UNDERSTANDING STATISTICAL SIGNIFICANCE: A SHORT GUIDE

Farooq Sabri and Tracey Gyateng

September 2015

Using a control or comparison group is a powerful way to measure the impact of an intervention, but doing this in a robust way requires statistical expertise. NPC's Data Labs project, funded by the Oak Foundation, aims to help charities by opening up government data that will allow organisations to compare the longer-term outcomes of their service to non-users. This guide explains the terminology used in comparison group analysis and how to interpret the results.

Introduction

With budgets tightening across the charity sector, it is helpful to test whether services are actually helping beneficiaries by measuring their impact. Robust evaluation means assessing whether programmes have made a difference *over and above* what would have happened without them[1]. This is known as the 'counterfactual'. Obviously we can only estimate this difference. The best way to do so is to find a control or comparison group[2] of people who have similar characteristics to the service users, the only difference being that they did not receive the intervention in question.

Assessment is made by comparing the outcomes for service users with the comparison group to see if there is any statistically significant difference. Creating comparison groups and testing for statistical significance can involve complex calculations, and interpreting the results can be difficult, especially when the result is not clear cut. That's why NPC launched the Data Labs project to respond to this need for robust quantitative evaluations. This paper is designed as an introduction to the field, explaining the key terminology to non-specialists. The guide covers comparison groups and how they are used to benchmark against your outcomes. It explains what statistical significance testing means, and how to interpret results—especially when your outcome is inconclusive.

[1] The standards of evidence used by innovation charity Nesta look to provide a more holistic hierarchy of evidence, combining both methodological confidence and other aspects of the maturity of the service. A single report of outcomes compared to a comparison group would provide evidence for Level 3 of the standards.

[2] Generally a control group is used to describe situations where people have been randomly assigned to a treatment or control (non-treatment) group, whilst comparison is generally used to describe when the non-treatment group has not been randomly assigned.

The counterfactual—estimating what would have happened without your intervention

To really understand whether a programme has achieved an impact, many funders and other stakeholders want to see charities comparing the outcomes of their service users (the treatment group) with a different group who have not received this particular service (the comparison group). The introduction of a comparison group eliminates a whole host of issues that normally complicate the evaluation process. For example, if you introduce a new programme to support young people into work, how will you know whether those receiving the extra support might not have found employment anyway? In statistical terms, this hypothetical estimation of what would have happened in the absence of the intervention is called the 'counterfactual', as mentioned above. By comparing this counterfactual data to what actually happened, we can estimate the impact of the intervention. With a well-matched comparison group, we can be reasonably sure that it was the intervention that achieved the impact and not other factors (such as, for example, improved economic conditions).

A good comparison group is as similar as possible to the group of people who are using the service or receiving an intervention. This means we can be reasonably confident that any difference in outcomes between the groups is likely to be caused by the fact that one group received the intervention and the other did not. Demographic and socio-economic characteristics, such as age, gender and ethnicity, are often used to check the similarity of a comparison group.

The most robust way to ensure a fair comparison group is to randomly assign people to the groups that will and will not receive the intervention[i]. The aim is to ensure that all factors that could affect the results are evenly distributed across the two groups.

Matched groups

If randomly assigning people is impractical, an alternative method is to use existing data to establish a group of individuals who seem to 'match' the characteristics of your service users. However, charities often struggle to access the data needed (such as the socio-economic details of people who have not used their services) and may lack the skills and capacity to create a comparison group.

That is why NPC first made the case for—and then supported the development of—the Justice Data Lab[iii]. This is a Ministry of Justice service that enables organisations to compare the re-offending rates (and more nuanced outcomes such as the frequency and time to re-offending) of cohorts of service users with those of a matched comparison group, using a technique called 'Propensity Score Matching' (see Box 1 for more details). From this first Data Lab, we have already seen some charities use results to complement other qualitative research on impact, helping them to better understand how their interventions work and to make the case about why their service should be funded. Over time, we expect that the use of Data Labs will support qualitative research and help organisations to improve their services.

Box 1: Propensity Score Matching

Propensity Score Matching (PSM) is the statistical method the Justice Data Lab uses to identify a comparison group. Using PSM, offenders in the treatment group are matched to non-treated offenders with similar propensity scores. Propensity scores are derived for each offender using a statistical technique that factors in a range of offender and offence characteristics associated either with being chosen for the intervention programme or with re-offending. Factors could include gender, age at offence and/or criminal history.[ii]

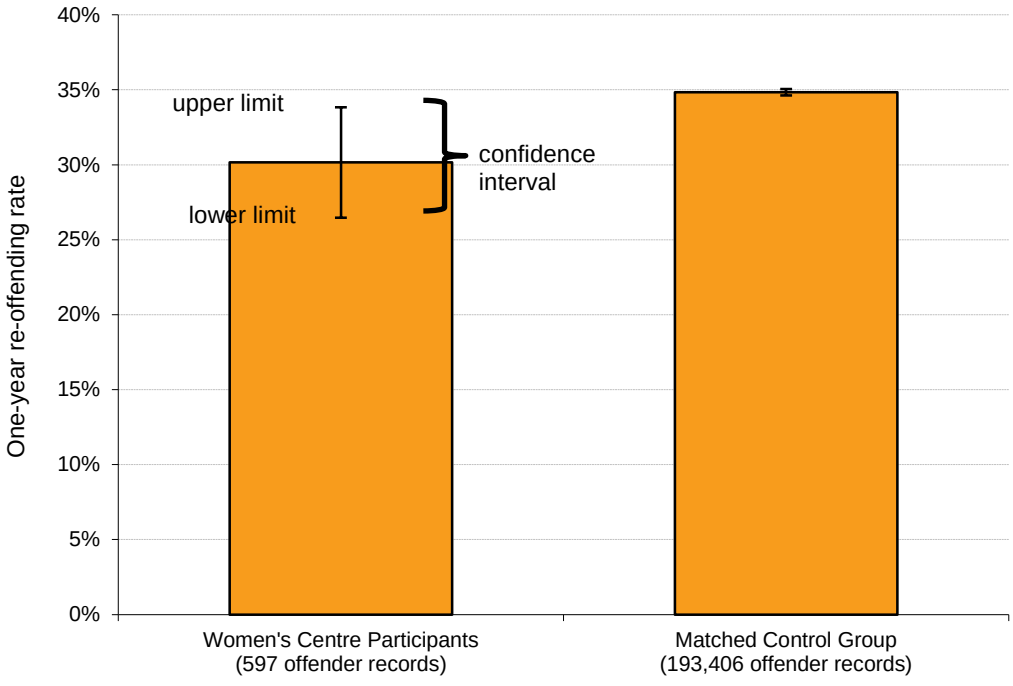
Statistical significance—comparing intervention and comparison group outcomes

Once our comparison group has been identified, the next stage is to test whether there's a statistically significant difference between the outcomes of the treatment and comparison groups. For example, a percentage point difference between the employment rate of the comparison and treatment groups of an employment programme does not necessarily mean that there was a genuine difference between the two groups. After all, the matched comparison group is an estimate of what would have happened to the treatment group in the absence of the intervention, and the treatment group is only a sample of people you have worked with from the population of (for example, unemployed) people. For this reason, there is always a level of uncertainty and potential error when estimating the impact of an intervention.

By convention, the acceptable level of statistical significance necessary to demonstrate impact is 0.05 (that is, a chance of error of 5%). This means that you can be 95% confident that any difference between your comparison and treatment groups (whether positive or negative) was not down to chance. For example, if your result indicates a difference between your comparison and treatment groups but the 'p-value' (a way of working out the probability of a programme having no impact given the data available) is greater than 0.05, there is insufficient evidence to conclude with confidence that the intervention had an impact on the outcome of interest. The lower the p-value, the more confidence we can have in concluding that there is evidence of impact.

To illustrate this, figure 1 is an example from the Justice Data Lab, which shows best estimates for the one-year proven re-offending rate for offenders—comparing those who received support provided by Women's Centres throughout England and a matched control group.

Figure 1



Source: Ministry of Justice (2015) *Justice Data Lab Re-offending Analysis: Women's Centres throughout England*[iv]

The Women's Centre Participants had an estimated re-offending rate of 30%, while the matched comparison group had a rate of 35%. Through statistical testing, the p-value was calculated as 0.01, which is less than the conventional 0.05. This indicates that the difference between the re-offending rates for the Women's Centre and

the comparison group were indeed statistically significant. However, p-values are not the only way to compare the differences between two groups. The confidence intervals (the lines drawn near the top of the bar chart) illustrate the boundaries within which an estimate of re-offending can be found. Figure 1 illustrates that the true difference in re-offending between the two groups is between one percentage point[3] and nine percentage points[4].

Among professional statisticians there are debates about the over-use of the p-value in statistical significance tests[5], as it appears to provide a definitive answer when really it is an estimation. Re-running the analysis can actually lead to changes in the p-value. Therefore it is important that p-value estimating isn't the only method used to categorise whether a programme is effective, and why it is equally important for evaluations to be repeated to test the robustness of initial results.

How to improve the chances of a statistically significant result

The ability to detect a difference between comparison and treatment groups with a reasonable degree of confidence is called 'statistical power'. The higher the statistical power, the more confident we can be that any differences are due to the intervention rather than random variation.

In practical terms, making sure we have enough statistical power is particularly important if the people in our sample are likely to have very different outcomes (a high sample variance) or if the expected difference between our comparison and treatment group (the treatment effect) is small. There is often very little that can be done to reduce sample variance and so we can only really increase our chance of detecting a genuine difference by making sure that we pick a large enough sample size[6] (see Box 2 for how to boost sample size). Almost any estimate may be found to be statistically significant if a large enough sample size is used.

In reality, especially for small organisations, large sample sizes can prove difficult to achieve. One solution is to consider pooling samples across cohorts/years or, for federated organisations, across member organisations that deliver similar programmes. However, it is important that the pooled samples share a similar intervention model and demographic characteristics of service users, because significant variation within the pooled sample will reduce the likelihood of a meaningful or statistically significant result.

Box 2: Boosting sample size when using government data labs

To match your service users to government datasets, it is necessary that accurate personal data such as names, date of birth and gender is collected. Without this information, it will be difficult to identify your service users in the government datasets. Any individuals with missing information may have to be excluded from the sample, thereby reducing the sample size and the ability to detect a significant result. Additional information, such as any unique IDs used by government (for example, Police National Computer IDs for justice, National Insurance number for employment), will improve the likelihood of a service user being found—helping to keep your sample as large as possible.

[3] The highest estimate of Women's Centre re-offending rate: 34%
Minus the lowest estimate of the comparison group: 35%
=1%

[4] The lowest estimate of Women's Centre re-offending rate: 26%
Minus the highest estimate of the comparison group: 35%
= 9%

[5] See for example Ioannidis, JPA. (2005) Why Most Published Research Findings Are False *PLoS Med* 2(8): e124. DOI:10.1371/journal.pmed.0020124; or Flanagan, O. (2015) Journal's ban on null hypothesis significance testing: reactions from the statistical arena, *StatsLife* (Royal Statistical Society).

[6] The MoJ provided an example of sample sizes needed to detect different effect sizes in the first report of the pilot year of the Justice Data Lab. A Google search of 'power calculators, sample size' can provide links to websites where sample sizes needed to detect effects can be produced. However, a basic understanding of statistics is needed to use these calculators.

Interpreting and responding to statistical significance

A statistically inconclusive result does not necessarily indicate that your project isn't effective. Quantitative evaluations only provide part of the story. How accurate that story is depends on careful interpretation of statistical significance tests. Here are some important considerations for interpreting the results of comparison group analysis:

1. Statistical insignificance can mean more research is needed

Statistical significance does not conclusively show that an intervention causes a particular outcome. It is better seen as a test of how precisely the relationship between an intervention and an outcome can be measured with the data collected. So, if statistical insignificance tells us there is insufficient evidence to draw conclusions, we do not know whether:

- the intervention has an impact, but we don't have sufficient data to confirm that it has an impact; or
- the intervention has no impact over and above the comparison group.

The first of these, where we falsely accept that the intervention has no impact, is called a Type II error[7]. As described above, the risk of this can be reduced by increasing sample size and statistical power.

In addition, an inconclusive result could mask whether the intervention worked particularly well, or poorly, for different types of service users in different circumstances. Your first response to an inconclusive results should be to think about for whom your project worked and did not work and why—conducting more research could help support this[8].

2. Economic significance is also important

The size of the difference measured has important implications. For example a 20% difference in employment rates of long-term unemployed clients (compared to the comparison group) is useful even if it is not statistically significant because it indicates that the intervention improves an economically important outcome, which could have significant implications for the welfare of beneficiaries and costs to the state. In this situation it will be important to repeat the study with a bigger sample and examine the qualitative research evidence to better understand how it has been effective and with which beneficiaries.

On the other hand, we could have a result that is statistically significant, but economically insignificant. For example, a statistically significant 1% improvement in an outcome that comes from an expensive intervention would need to be reviewed, taking account of whether the benefits of scaling up the service outweigh the costs.

3. Is the right outcome being targeted?

Improvements in long-term and ambitious policy outcomes like reductions in youth offending or improvements in employment rates can take many years to measure. A youth employment programme's impact on long-term unemployment, for example, may be statistically insignificant if measured several months after the programme has finished, but it could be statistically significant if measured years later. This makes measuring intermediate outcomes particularly important. These are factors that are known to contribute to the long-term outcome we want and are observable and measurable more quickly, which makes them a useful way to test impact over the short term as well as supporting claims about long-term impact once that data is available. Charities operating in the criminal justice sector considering their own intermediate outcomes should review the National Offender Management Service Commissioning Intentions and Evidence Tables[v].

[7] As described above, the chance of this type of error occurring is higher for samples that have a large variation in outcomes and where the expected impact of the intervention is small.

[8] See Kay, J. (2014) *The Significance of No Significance.* Education Endowment Foundation.

4. Does our comparison group benefit from other interventions?

As we've already explained, a comparison group needs to be as similar as possible to our treatment group, with the exception of the intervention in question. But what if the comparison group receives other interventions? Within justice it is very unlikely that an offender will receive no interventions whatsoever (for example, under Transforming Rehabilitation, all prisoners now receive some supervision after their release). The Justice Data Lab is unable to control for all the possible interventions offenders could have received. Therefore a comparison between a treatment group and a matched control is really a comparison between a specific intervention and a group that has received a mixture of other interventions, not a comparison between our intervention and a group that has received no intervention. This is highlighted in Box 3.

Box 3: Treatment groups versus comparison groups

Treatment Group A - Comparison Group B (received no treatment at all) = Impact of intervention compared to no intervention.

However, in a sector where all offenders are likely to receive some type of intervention/supervision:

Treatment Group A - Treated Comparison Group B = Impact of intervention compared to general treatment of offenders.

Thus, if charities work in an area where beneficiaries are likely to be subject to multiple interventions at any one time, statistical significance may become a measure of the relative impact of a programme compared to other programmes. In the same way, statistically insignificant results could suggest that your programme is just no worse than other programmes that are out there.

5. How good a 'match' is our comparison group?

A limitation of the matched control group ('quasi experimental') approach to identifying a comparison group is that it can only match individuals on characteristics where data is available. The Justice Data Lab, for example, matches offenders using the offender characteristic data that is actually recorded by the Ministry of Justice. However, there may be other factors, such as an offender's motivation to receive treatment, that are not reflected and are likely to have an influence on re-offending rates.

It is always important to consider this point when looking at the results from quasi-experimental evaluations. Very often, limitations in the matching process offer an explanation as to why a difference has or has not been found.

Conclusion

Interpreting the results of a matched control group analysis requires an understanding of how the statistics have been produced and how the assumptions have been made. This is not unique to matched control group analysis; all evaluation techniques have their strengths and limitations. The key issues to consider are:

- Matched control group analysis should not be interpreted as the final word on whether an intervention has made a difference. As a purely quantitative approach, it is limited in explaining how/why any detected impact came about.
- It is important to re-evaluate the intervention to check for stability of results where you have strong reasons (theoretical and qualitative research) for doing so.

Finally, an inconclusive result should not be considered as being worthless. In fact, it may signal that a bigger sample is needed, or indeed that your service is no better or worse than the other interventions an average person may receive in their day-to-day lives.

Acknowledgements and contact details

Thanks to Sarah French, Ministry of Justice; Jess Mullen, Clinks; and colleagues at NPC who reviewed this paper. Any errors remain the authors responsibility.

If you have any questions, please get in touch via info@thinknpc.org and visit www.NPCdatalabs.org for more on our Data Labs project, including our progress in new policy areas.

Useful links and references

[i] For a more detailed guide to comparison groups, see Clinks' and NPC's *Using comparison group approaches to understand impact* from the Improving Your Evidence project.

[ii] Ministry of Justice (2013) *Justice Data Lab Methodology Paper,* p. 13.

[iii] https://www.gov.uk/government/publications/justice-data-lab

[iv] https://www.gov.uk/government/statistics/justice-data-lab-statistics-may-2015

[v] The latest report (at the time of publication) can be found here: https://www.gov.uk/government/publications/guidelines-for-services-commissioned-by-noms

TRANSFORMING THE CHARITY SECTOR

NPC is a charity think tank and consultancy which occupies a unique position at the nexus between charities and funders, helping them achieve the greatest impact. We are driven by the values and mission of the charity sector, to which we bring the rigour, clarity and analysis needed to better achieve the outcomes we all seek. We also share the motivations and passion of funders, to which we bring our expertise, experience and track record of success.

Increasing the impact of charities: NPC exists to make charities and social enterprises more successful in achieving their missions. Through rigorous analysis, practical advice and innovative thinking, we make charities' money and energy go further, and help them to achieve the greatest impact.

Increasing the impact of funders: NPC's role is to make funders more successful too. We share the passion funders have for helping charities and changing people's lives. We understand their motivations and their objectives, and we know that giving is more rewarding if it achieves the greatest impact it can.

Strengthening the partnership between charities and funders: NPC's mission is also to bring the two sides of the funding equation together, improving understanding and enhancing their combined impact. We can help funders and those they fund to connect and transform the way they work together to achieve their vision.

New Philanthropy Capital
185 Park Street, London SE1 9BL
020 7620 4850
info@thinkNPC.org

Registered charity No 1091450
A company limited by guarantee
Registered in England and Wales No 4244715

www.thinkNPC.org