LECTURE 4. RANDOMISED EXPERIMENTS: INFERENCE AND EXTERNAL VALIDITY

Dr. Tom O'Grady

1. Inference with Randomisation

So far in your QStep career, we have learned only one way of carrying out *statistical inference*, by which I mean making inferences about something unknown (a population quantity) from something known (a sample). We have relied on standard statistical inference, which assumes that we have a large sample from a population. There is uncertainty about the population quantity of interest, e.g., an average treatment effect, because rather than observing the whole population, we observe only a sample from it. We know, however, what the sampling distribution of our quantity of interest looks like (either deriving it mathematically or by simulation). From there we can derive a test statistic – usually a standardised regression coefficient or ATE – as well as a null distribution, and use them to test null hypotheses about our quantity of interest.

This type of testing relies on a number of assumptions. For instance, we must invoke the Central Limit Theorem to justify the use of standard errors when working with small samples whose own sample variances may be very far from normally distributed. Particularly in small samples with non-normal sample variances, the usual methods of inference may perform very poorly. And it is by no means uncommon to have experiments that are carried out on very small samples - e.g., on 16 units as in question 1 of the third tutorial.

It turns out that in experiments, we can rely on an entirely different method of statistical inference, known as **randomisation inference**, first invented in a setting involving cups of tea.

1.1. The Lady Tasting Tea: Fisher's Exact Test

It was invented by the statistician Ronald Fisher in a famous example known as **The Lady Tasting Tea** which made its way into his foundational textbook on experiments in the 1930s. The story goes like this: Fisher was made a bet by a colleague, Muriel Bristol, who claimed that she could tell, in a blind taste test, whether or not a cup of tea was poured with the milk first followed by the tea, or the tea first followed by the milk.

Fisher and his colleagues set up an experiment. Bristol was presented with 8 cups of tea, four poured with the milk first and four poured with the tea first, in random order. She then had to say which cup was which. She guessed all 8 correctly - or so the story goes. Does this mean that she can really tell the difference, or was she just lucky? Fisher's method involves

asking: suppose that she actually had no idea what she was doing, and guessed purely at random. What is the probability that she could have got the answer correct by chance alone?

To answer that, we think of her labelling as being fixed: for instance, she labels them 1-1-1-1-0-0-0-0. Our test statistic is the number of successes, and its distribution consists of all of the possible ways in which she could have got 0, 1, 2, 3, or 4 successes. Under the null hypothesis that she cannot distinguish the cups of tea, there is only one way for her to get all four correct: if all 8 cups are arranged exactly as she guessed. But there are a lot of ways for her to get 3 of them correct: in fact, there are 4 ways in which the first 4 cups could be arranged (xxx0, 0xxx, x0xx, xx0x) and in turn another 4 ways for the second 4 cups to be arranged. This gives 16 possible ways to get three correct guesses. Continuing that logic, there are 36 ways to get 2 correct answers, 16 ways to get 1 correct answer, and 1 way to get 0 correct answers, giving a total of 70 possible permutations (ways to distribute 4 cups among 8), which we can also calculate using the formula for combinations (with repetition):

$$\frac{8!}{4!(8-4)!} = 70$$

Under the null hypothesis that she in fact has no idea what she's doing, there is only a probability of $(1/70) = 0.014$ that she gets all four answers exactly correct. This is known as the **exact p-value**. Clearly, we have very strong evidence that she can indeed guess correctly. But suppose she got 3 correct and one wrong. In such a case, the exact p-value would be $\frac{17}{70} = 0.24$, because there are 17 ways for her to get 3 or more successes under the null hypotheses (there is 1 way for her to get 4 successes and 16 ways for her to get 3 successes). With three correct guesses, we can't reject the null hypothesis that she is unable to test tea as she claims.

1.2. Randomisation Inference in Experiments

This test does not rely on any notion of the eight cups being a random sample from a larger population of cups of tea. It is not an *approximate* test based on an approximation to the true sampling distribution, but rather an *exact* test because we know the entire distribution of possible cups of tea, given the randomisation scheme. We can carry out exact tests in any experiment where treatment is randomly assigned. We can think of uncertainty as arising in experiments not because of the process of sampling, but because of the process of randomisation. Just as it is possible that Fisher happened, by chance alone, to pour the four cups with milk first in exactly the four cups that Bristol picked, it is also possible for large Average Treatment Effects to arise by chance alone because we happen to allocate the units with highest potential outcomes to treatment, and the units with the lowest potential outcomes to control. In other words, randomisation might 'accidentally' produce a large treatment effect even though in reality the true treatment effect is 0.

This suggests a test, known as Fisher's Exact Test. Suppose the true treatment effect is 0 for every unit. In such a case, we are assuming a **Sharp Null Hypothesis:**

$$H_0 : Y_i(1) = Y_i(0) \text{ , for all } i$$

Here, we can create a null distribution that is the exact sampling distribution under the sharp null hypothesis. Why? Because even in our one realization of the experiment, we're observing

all of the potential outcomes under both treatment and control if the sharp null is correct. That means that we can use the observed outcomes to tell us what would happen under any other randomization. Usually, we couldn't do this because each unit reveals a different potential outcome depending on whether they are in the treatment or control condition. The sharp null hypothesis is a special case where they reveal the same potential outcome always - because the treatment has zero effect.

We can then assess how unlikely our actual treatment effect is under the sharp null by calculating the ATE under every alternative randomization scheme and observing which percentiles our ATE falls into. This is just like the Lady Tasting Tea example. Intuitively, we're asking how often a treatment effect at least as big as the one that we observed would occur merely due to the quirks of randomisation alone, if the true treatment effect is zero. Remember that in most cases in the social sciences we carry out *two-tailed* tests, so"at least as big" needs to be interepreted in terms of absolute values.[1]

Fisher's Exact Test is a common method for analysing experiments carried out on small samples, where inference techniques designed for large samples may perform poorly. It performs particularly well in small samples with non-normal (e.g. bimodal) distributions. There are a couple of caveats/extensions to it that are worth mentioning:

1. In large samples, the test can still be carried out but is unlikely to perform much better than standard statistical inference. In addition, it would be impossible to calculate every possible randomisation. As Gerber and Green point out, even an experiment with 50 units, half of which are assigned to treatment, has over 126 trillion possible randomisation schemes. In these cases, one instead needs to take a sample from the randomisation distribution. We in fact saw an example of this in the second tutorial exercise where we had 100 experimental units and took a sample of 10,000 possible randomisations. Even in 'medium-sized' datasets, the code can also be very slow to run because it takes a long time for R to permute long treatment vectors.

2. If randomisation was carried out with blocking, then the randomisation distribution needs to be calculated calculated in the same way, with each possible randomisation generated by the same blocking scheme that was used in the original experiment.

1.3. Fisher's Exact Test: Example

Now we'll look at an example to see how Fisher's Exact Test works in practice to analyse an experiment.

[1]You might have noticed that Fisher's test in the Lady Tasting Tea example is one-tailed. If you were really eagle-eyed, you might also have noticed that Kalla and Broockman, in this week's example paper, used a one-tailed test without explicitly justifying why. You have always been taught to use two-tailed tests, but it's not necessarily crazy to use a one-tailed test instead. In fact it's fair to say that statisticians don't always agree, these days, on the use of one-tailed versus two-tailed tests. Two-tailed tests are more conservative – we are less likely to reject the null hypothesis for a given significance level – but one-tailed tests can arguably be justified in a situation where it makes no sense for our alternative hypothesis to fall on both sides of the null hypothesis. In a medical trial, for instance, it may be thought of as unlikely that a new drug could ever make a disease's symptom's *worse*, so some argue for alternative hypotheses that are one-tailed and strictly positive. Note, then, a downside of one-tailed tests: they rule out the possibility of discovering a significant effect in the opposite direction to the one we expected.

2. Experiments and External Validity

In the second part of today's lecture we move from analysing experiments to considering how much we can really learn from them. In particular, we're asking how well experiments can actually serve the needs of policymakers and social scientists. Dani Rodrik's article is fairly balanced and gives a good introduction to some of the debates. He emphasizes a clear shift in emphasis amongst international organisations, governments and academics towards a more practical, experimental approach to public policy. This can mean less reliance on theoretical, sometimes dogmatic, arguments in development and economic policy. It can mean an emphasis on only advocating policies based on 'what works'; on clear empirical evidence, rather than ideology. And it can mean taking an experimental approach to policy-making more broadly, focusing on piloting, testing and improving new policies before rolling them out in full.

Experiments fit easily into this new paradigm in policy-making. So much so that there are some academics, including those associated with the MIT 'Poverty Action Lab' that Rodrik discusses, who say that no new policies should be adopted by governments unless they have first been subject to randomised control trials that test their effectiveness. There are some empirical researchers who now simply refuse to even consider evidence from observational studies, arguing that only studies with very high internal validity – from experiments or natural experiments – should be considered as worthwhile evidence. These views are hard to dismiss easily. Those who hold them could rightly point to cases like HRT therapy in women, where thousands of women may have died unnecessarily based on faulty observational evidence, as proof that observational studies are fatally flawed. They also point to large-scale policy interventions like the IMF's 'Washington Consensus' programs in developing and former Soviet countries that were based on limited evidence and had very mixed results.

Critics like Rodrik, though, argue that even though this may be true, it does not necessarily follow that experimental evidence is a 'gold standard' that clearly trumps any observational study. The arguments of these critics generally focus on two areas. First, many policy areas that are deeply important, particularly macro policies like economic policies, cannot even in principle be subjected to randomised trials. As Rodrik says, many of the biggest gains in human welfare have come not from small, easily randomisable interventions like mosquito nets, but rather from whole government programs, like industrial and growth policies, that no-one could first test in a trial. What should we make of this suggestion? Undoubtedly, many important policies can't be randomly tested. But this surely isn't a good argument against using experiments where they are feasible. It simply suggests that greater humility in implementing large-scale policy changes would be a good thing, and that smaller elements of a larger program could be subjected to randomised trials wherever possible.

The second reason that experiments are attacked is harder to rebut quickly. Rodrik is uneasy about the shift toward experimentation because he feels that it is difficult to use experiments to accumulate systematic evidence that can be applied in many contexts. While 'true believers' will often say that the evidence from *any* one experiment is inherently more valuable than dozens or hundreds of observational studies, detractors will say that experiments have limited 'generalisability.' By that, they mean that there may be limits on the extent to which the findings of experiments can be applied to groups or settings outside the original experiment.

This is the problem of external validity. It matters because usually the whole point of running an experiment is to help social scientists and policy-makers draw conclusions about

how the treatment would affect a population in the real world. The problem is that the pursuit of external validity can directly conflict with achieving high internal validity. It is precisely the ability to hold all aspects of the world constant except the treatment, exerting very close control over randomisation, that makes experiments so attractive to begin with. But this often requires us to carry out experiments on small groups or in artificial, controlled settings.

It is worth pointing out that issues of external validity are not completely unique to experiments. Observational studies may also be carried out on unusual groups of people or in specific settings. Nonetheless, it is more often the case that observational research can study the units of interest in their 'natural setting', and can more often include a broader sample of those units. Thus it is reasonable to consider experiments as more prone to these issues. There are three main problems of external validity in experiments, which we'll look at briefly in turn: unrepresentative samples, unrepresentative settings, and the impact of being studied. Note that these problems are not wholly distinct from each other.

2.1. Unrepresentative Samples

While observational studies often focus on the whole population of interest, experiments are often carried out on small and unusual samples, like university students. It is unclear whether the results from these small subgroups would generalise to larger populations, on whom policies are typically enacted or about whom social science theories are said to apply. Likewise, it is expensive and logistically challenging to run field experiments in many locations at once. Like the mosquito nets example in Rodrik's paper, field experiments can usually only be carried out in one country/town, etc., whereas observational studies can often include many countries at once in a single regression. The single sub-populations on which experiments are carried out may be atypical of the wider population that observational studies are able to cover.

A well-known example concerns the use of university students in experiments. Particularly in fields like psychology or experimental microeconomics, students have often been used because they are cheap and are easy for academics to access. Indeed, they are sometimes called "convenience samples." How typical are students of the general population? In some ways, they are unusual. In a famous article from 1986, David Sears argued that some of the conclusions of social pyschology may be merely an artefact of using students to test theories.[2] Pyschologists have often argued, based on experimental results, that people's behaviour is very strongly influenced by the beliefs of others around them, and that people's attitudes in life are quite changeable. However, it is also well known that these are traits that are associated with young people, and with university students in particular, who often update their opinions in early adulthood as they are exposed to new ideas. The results might well differ with samples that have more variable ages.

In a similar vein, the economist Ariel Rubinstein has carried out experiments on different populations of undergraduates, designed to assess the extent to which students behave like "profit-maximisers."[3] When faced with a hypothetical company that needs to sack workers in order to increase profits, what should the owner do? Economics students were far more likely than those from other disciplines to want to sack workers. Rubinstein attributes this

[2]David Sears (1986). "College Sophomores in the Laboratory: Influences of a Narrow Data Base on Social Psychology's View of Human Nature." *Journal of Personality and Social Psychology* 51 (3), pp. 515-530.

[3]Ariel Rubinstein (2006). "A Skeptic's Comment on the Study of Economics." *The Economic Journal* 116: C1-9

to economics students being "indoctrinated" by their studies. His findings call into question whether the results of economics experiments, which examine theories on altruism and profit-maximising behaviours, should be carried out on economics students, who may exhibit more of a tendency to behave in line with neo-classical economic models than other people.

2.2. Unrepresentative Experimental Settings

While observational studies examine units in their 'natural' environment, experiments often take place in artificial settings like labs or internet surveys. Behavioural patterns observed in these artificial settings may have little relationship to behaviour in the real world. In particular, subjects may pay more attention to tasks like reading news articles than they would in the real world, and cannot choose not to consume them (unless they leave the experiment). Thus survey experiments in particular may 'exaggerate the power' of the treatments being studied.

2.3. The Impact of Being Studied

Beyond an unrepresentative sample or unrealistic setting, subjects may change their behaviour from that which would occur in the 'real world' simply because they are aware that they are being studied. Datapoints in a regression, on the other hand, don't know that they are being watched. The Tennessee Star experiment from last week contained two possible examples of these types of effects, where teachers might have tried extra hard simply because they were in an experiment. They may have thought that if smaller class sizes proved a success, policy-makers would be more likely to adopt the policy permanently. This may have given the treatment group an added incentive to exert effort, exaggerating the impact of smaller classes. Or conversely, teachers in the control group may have felt a need to prove themselves, showing that they were such good teachers that they could overcome the potential hurdle of larger classes. In both cases, behaviour in the experiment differs from the behaviour that we would observe if the policy were to be rolled out in the real world, outside of an experiment.

The phenomenon of experimental subjects working or trying harder due to being under direct observation is sometimes referred to as a 'Hawthorne Effect' after an experiment that was carried out at the Hawthorne factory in Chicago, to see whether greater natural light on the factory floor caused workers to work harder. They found that it did. Nonetheless, others later pointed out that the workers were aware of being studied and may have simply enjoyed the novelty of being in an experiment and having someone care about their work. In a similar vein, subjects may want to try to please experimenters or may figure out what the experiment is about, and try to put forward the behaviour that experimenters are hoping for. This is a particular concern with groups who are commonly employed in experiments, like psychology undergraduates.

3. Differences Across Types of Experiments

Just as concern for external validity tends to differ across fields of study, so it also tends to vary across *types* of experiments. Although the differences aren't always clear-cut, we can generally delineate four types of experiments:

Survey Experiment: an experiment that is embedded within an online, phone or face-to-face opinion survey.

Lab Experiment: an experiment that takes place in person at a specially-designed facility, like a computer lab

Field Experiment: an experiment that takes place in a real-life setting that is unaltered by the analyst, except for the assignment of treatment and control

Natural Experiment: an experiment that takes place in a real-life setting that is unaltered by the analyst, and where treatment occurs without the analyst's intervention

In some respects these have a natural hierarchy, where the first two might have the highest internal validity – the experimenter has the most control over the situation – and the last two, particularly natural experiments, have the highest external validity. As we saw in the last lecture, field experiments are prone to problems like non-compliance, attrition and randomisation failures. Nonetheless, it is worth pointing out that this hierarchy only really holds for two of the three aspects of external validity, unrepresentative experimental settings and behaviour being altered by being studied. There is not necessarily any ranking when it comes to the representativeness of study populations, and one might even argue that the ranking is inverted. Survey and lab experimenters, at least in theory, have a lot of choice about who they invite to do their experiments. It is frequently possible to conduct survey experiments on a nationally representative sample. But field experiments can rarely be rolled out across large populations. Natural experiments often take place in one particular location or at one particular time only. Here, analysts are stuck with whatever happened to occur in the real world, which may or may not be related to a sub-population that we would ideally choose to study.

4. Differences Across Fields of Study

Some areas of the social sciences show greater concern for external validity than others. The amount of concern shown tends to vary with the universality of the behaviour being considered. Microeconomists and psychologists often study (or at least think they study) behaviours that are quite intrinsic to all humans, and are therefore less likely to differ across different subgroups of people. This tends to make them more comfortable with using convenience samples in experiments. The aim of these experiments is often to test some underlying motivation or theory of human behaviour (like 'greed' or 'atruism') without wishing to immediately influence public policy for a whole population.

This tends to lessen their concern for unrealistic experimental settings, too. Perhaps the most famous, and most controversial, example of this sort of abstraction is provided by Stanley Milgram's 1961 experiment on obedience. He began by asking how prison guards can possibly carry out atrocities like those that occurred in Nazi prison camps. He argued that under the right conditions, even very ordinary people are capable of following orders to inflict pain when told to do so.

Subjects participating in his experiment arrived at his center believing they were being randomly assigned to be either a 'teacher' or 'learner.' In fact, the drawing of straws was faked

and all became teachers, and were told to teach a learner a series of word pairs via audio link. In reality, the audio was simply pre-recorded by an actor and was the same for all participants. Subjects were instructed to administer an electric shock to the 'learner' when they gave a wrong answer, to encourage them to learn. Gradually, the size of the electric shocks were increased and the learner became more and more distressed and apparently in pain. Nonetheless, the experimenter told the subjects very firmly that they must continue, even though the fake 'learner' appeared to be suffering and many of the experimental subjects became distressed themselves. The experimenter wore a lab coat, spoke authoritatively, and identified himself as being in charge. Many of the subjects went on to administer maximum-strength electric shocks, particularly after the experimenter instructed them "You have no choice. You must continue." Obviously this experiment is very far removed from the experience of a prison camp. But it does quite powerfully demonstrate the potential for humans to obey figures of authority issuing orders. It sheds light on the general phenomenon of obedience, despite taking place in a deeply abstracted setting. Milgram used the experiments to build a more general theory of obedience to authority.[4]

Political scientists, unlike psychologists and experimental economists, tend to be extremely concerned about external validity. This is perhaps because political behaviour can often be highly contextual and influenced by particular institutional settings. Voters clearly do not think and behave the same over time or across different political contexts, even though consumer behaviour may be quite similar. And most voters in the real world are far less engaged in the world of politics than they are when they become involved in an experiment. For these reasons, experiments remain quite contested and controversial in political science, although they are gaining in popularity. Epidemiologists probably fall somewhere in the middle of the pack; they have become increasingly worried that the type of people who show up for medical trials are atypical of patients in general, but almost always tend to conduct their experiments in real-world settings. They also face particularly acute ethical and practical difficulties in actually carrying out experiments, which has tended to keep observational methods quite popular.

5. Making Experiments more Externally Valid

It is possible to mitigate or minimise some of the problems of external validity with good research design. We saw last week that there are often ways to design experiments in ways that minimize potential problems of internal validity before they occur: good *research design* is very often the best solution to problems. The same is true of concerns about external validity. There are several ways in which we can improve the design of experiments with external validity in mind:

- *Use more representative samples:* Where possible, try to carry out experiments on individuals whose characteristics more closely match the population that you are ultimately interested in
- *Repeat the experiment on multiple samples:* If you are restricted to a small or unrepresentative sample, try to repeat the experiment again on other samples with different characteristics

[4]It's worth noting that this is a pretty controversial experiment. I can't imagine any university ethics review board allowing such an experiment to go ahead today.

- *Make the treatments and research setting as close to the real world as you can*
- *Use Manipulation checks:* Check to see whether the treatment produced the changes in the treatment group that you were expecting to see or would see in a real-world situation. E.g., if you ask them to read a newspaper article, ask them factual questions to see if they actually read and understood it.
- *Gather feedback after the experiment:* Try to find out whether or not people knew that they were in an experiment, or whether they guessed what the aim of the study was.
- *Use deception and/or subtle treatments:* To prevent hawthorne effects or similar issues, don't inform people that they were in an experiment, if this is feasible and ethical, or use subtle experimental manipulations.

Stances on the use of deception, in particular, tend to differ across fields. In psychology, virtually every study uses deception to mask the experimenter's intent. One problem with this may be that many people doing a psychology experiment now expect to be deceived, and may spend much of the time trying to work out what type of deception is being employed, leading to distorted behaviours. Deception is also common in many survey experiments in political science. Here, though, respondents often have no idea that they are in an experiment in the first place. On the other hand, in field experiments and in particular in medical trials, informed consent rules rightly make it almost impossible to use deception. Only in settings like survey experiments, where the potential harms from taking part in an experiment are virtually zero, can it be ethical to use deception.

6. Assessment: Learning from Experiments

Some social scientists – although not many, these days – argue against all experiments in general on the basis of external validity concerns. How concerned should we really be about external validity? Rodrik's argument is that single experiments cannot settle any policy questions, and that this should be viewed as a somewhat damning critique of them. Certainly, he is right to say that the study he discusses on mosquito nets is not the last word on them, and that 'softer' contextual information is needed in order to interpret and use experimental results in practice. He then goes on to argue that, as a result, experiments have little inherent superiority to observational studies. They do not constitute 'better' evidence, in his view.

Many modern social scientists, me included, would fundamentally disagree with him. One of the main contributions of the 'credibility revolution' and renewed emphasis on design-based inference has been to bring internal validity more to the forefront. Most researchers today argue that internal validity is more important. They would rather draw causal conclusions about at least one sub-population, even in a highly abstracted setting, rather than drawing much weaker conclusions about larger groups in a natural setting, as in traditional multiple regressions. At least, they would say, Rodrik's mosquito study gives definitive guidance to policymakers in the one setting (country, time, population, etc.) where it was carried out. And if free mosquito nets work in one place, it is at least somewhat likely that they would work elsewhere.

One justification for experiments that is sometimes invoked – and one that I tend to endorse – is that some social scientists expect too much of any one study. No single study can provide

definitive proof of a theory. Sometimes experiments can serve as merely a 'sanity check' for a wider theory, particularly experiments that are carried out in highly artificial settings like a lab experiment. Take my own work on the influence of political discourse on public opinion toward welfare, introduced at the end of Lecture 1. No lab or survey experiment can possibly prove that large-scale historical movements in public opinion were due to politicians' rhetoric. But if the theory is true, it should at least be the case that in a survey experiment, reading actual political speeches that present 'extreme' rhetoric causes a shift in stated opinions. If it did not, there would be cause for doubt. Thus it may be wiser to view experiments carried out in artificial settings as a modest contribution to a wider *research program* that could encompass multiple experiments and multiple different types of evidence and research approaches. It is not problematic to carry out survey experiments in artificial settings provided we do not rely on them as our only source of evidence.

And when it comes to the fact that experiments may be carried out on 'unrepresentative' populations, we should ask: unrepresentative of what? A lot of social behaviour and institutions are inherently local and contextual, so that limited external validity is not so much a problem to be battled against as simply a part and parcel of the social world that we, as social scientists, seek to understand. As Rodrik points out, his argument could easily be re-framed to say that we simply need to run experiments on mosquito nets in more places. His call for an end to 'one size fits all' theories of development would be readily endorsed by most experimental researchers. It is perfectly consistent with the idea of running multiple experiments in many contexts, each of which on its own has high internal validity but limited external validity. This would allow policymakers to find out what works where, and how to tailor policies in different communities or countries. Thus in an ideal world, we would have multiple experiments all answering the same basic question in multiple different settings. The problem, as he also points out, is that it is difficult for academics to obtain funding to replicate existing studies, or to publish papers that tackle an identical question in a new setting. The incentives of academics are not, perhaps, well-aligned with the needs of policy-makers.

Overall, though, Rodrik is probably attacking a straw man when he suggests that experiments are flawed because they cannot definitively settle a research question. The point is that no study of *any* kind can do so. It is fanciful to argue that any one observational study could settle a research question either. Only a long-term research program encompassing multiple methods and approaches can establish firm facts about the social world. External validity is often invoked as a damning critique of experiments, but it is probably a less serious problem than critics imagine, once we think about it more carefully.