LECTURE 2: CAUSATION AND RANDOMISED EXPERIMENTS

Dr. Tom O'Grady

1. Introduction: Causation and Experiments

We'll start by introducing some of the basic ideas of the course, using the example of the impact of the media on people's beliefs. This is just an introduction. We'll be returning to all of these ideas throughout the course in more depth.

1.1. Multiple Regression, Correlation and Causation

Consider the regression estimates in Figure 1, from a table contained in a recent paper by Niederdeppe *et al.*[1] They investigated whether watching a lot of local news causes people to form exaggerated, fatalistic beliefs about the dangers of cancer. They measured both TV viewing and people's beliefs about cancer from a survey. They ran a regression. Their dependent variable was a scale, where higher values indicated that the respondent held more false beliefs about cancer. Their key independent variable was the amount of local news that the respondent watched. They controlled for a very large number of other characteristics that may be correlated with both beliefs about cancer and TV viewing.[2] They argue that their results (in the first two columns) show a causal effect of TV viewing on beliefs. In the language of this course, we'll say that newspaper viewing is their *treatment*, and the impact of newspaper viewing on beliefs is their *quantity of interest*:

> **Treatment Variable:** The independent variable whose impact we want to measure
>
> **Quantity of Interest:** The effect that we want to measure. Typically, the impact of a treatment on an outcome (treatment effect)

We'll also say that they want to *identify* the impact of newspaper viewing on beliefs, using a particular *identification strategy*

> **Identification Strategy:** The approach that we use to infer our quantity of interest from the data

[1] Jeff Niederdeppe *et al* (2010). "Does Local Television News Coverage Cultivate Fatalistic Beliefs about Cancer Prevention?" *Journal of Communication* 60 (2): 230-253

[2] Actually, I've spared you. Figure 1 shows only the first of two pages of control variables. Writing "continued overleaf" on a regression table should make you think twice about whether your research is really sensible.

Figure 1: **Regression Estimates of the Impact of Local News Viewing on Beliefs about Cancer from Niederdeppe *et al* (2013)**

Table 3 Ordinary Least Squares Regression Models Testing Variables Associated With an Index of Fatalistic Beliefs About Cancer Prevention

Independent Variables	Bivariate Models	Multivariate Model	Model With Interactions
Variables related to Hypothesis 4			
Number of days watching local TV news in past week	.07*(.026)	.10*(.026)	.11(.080)
Variables related to Hypothesis 5			
Education less than high-school diploma (vs. college degree)	.09**(.009)	.06(.137)	.11(.101)
Education completed high school (vs. college degree)	.12**(.001)	.09*(.018)	.07(.309)
Education some college (vs. college degree)	.07 (.072)	.05(.206)	.06(.427)
Education less than high-school diploma × local TV news viewing	—	—	−.07(.342)
Education completed high school × local TV news viewing	—	—	.02(.807)
Education some college × local TV news viewing	—	—	−.02(.801)
Control variables			
Age	−.05(.092)	−.03(.452)	−.03(.475)
Female (vs. male)	.02(.491)	.01(.670)	.01(.658)
Non-Hispanic (NH) African American (vs. NH White)	−.02(.445)	−.06(.056)	− .06(.059)
Hispanic (vs. NH White)	−.04(.202)	−.05(.086)	−.05(.075)
NH other (vs. NH White)	−.06(.106)	−.03(.404)	−.03(.415)
Working full- or part-time	.03(.383)	.08*(.041)	.08*(.042)
Married	−.03(.333)	−.01(.779)	−.01(.779)
Household size	.03(.356)	.02(.616)	.02(.624)
Self-rated poor or very poor health (vs. very good or excellent)	.09**(.002)	.08*(.016)	.08*(.016)
Self-rated fair health (vs. very good or excellent)	.08*(.018)	.06(.091)	.07(.082)
Self-rated good health (vs. very good or excellent)	.07*(.035)	.06(.069)	.06(.081)
Body mass index (BMI)	.04(.210)	.02(.626)	.02(.590)

(*continued overleaf*)

In this case, their identification strategy hinges on a key assumption: that the control variables in their regression account for all possible *confounders*. This type of research is sometimes called *observational*, because it involves observing the world and then trying to make inferences about it. This is the opposite of experimental research, where the researcher manipulates something in the world in order to study a causal effect.

> **Confounder:** A variable that influences both uptake of the treatment and the outcome.

> **Observational Research:** The opposite of experimental research: a study that involves observing the world and making inferences about it from our observations.

A failure to adequately control for confounders may lead to drawing the wrong conclusions about a quantity of interest. For example, the authors control for education, since low education may be positively correlated both with TV viewing and holding false beliefs about cancer. Without controlling for it, the estimated treatment effect would be too high. The general problem is that certain types of people are more likely to *select into treatment*. People who already hold certain views are likely to choose to watch (select) media that puts forward those views. We need to account for every possible confounder in regressions that relate media consumption to opinions. Otherwise, we cannot be certain that results such as Niederdeppe *et al* are due to the influence of the media, or merely due to certain types of people selecting certain news in the first place. If we fail to account for all confounders, the result is *selection bias*, which we'll define properly below. It is, however, the same thing as *endogeneity* or *omitted variable bias*.

Equivalently, we wish to know whether consumption of particular news *causes* certain opinions, or is merely *correlated* with them. In the case of a positive relationship:

> **Correlation:** Higher values of X are associated with higher values of Y

> **Causation:** An increase in X leads to an increase in Y

Correlation is a necessary but not sufficient for causation. Many variables in the social sciences are associated with each other, without being causally related. Disentangling which relationships are causal, and which are not, is a key task for social scientists. Until the past 20 years or so, multiple regression was the main tool that social scientists used to assess causality. This raises an obvious question: how do we know whether we have controlled for everything important? Can we ever hope to exhaust the possible confounders through multiple regression? Most modern social scientists would say 'probably not.' Such scepticism about regression as a tool for causal inference is the motivation for this module - and in particular, for the use of experiments.

1.2. Randomised Experiments

It is possible to study media effects with an experiment, instead. One of the first experiments in poltical science was carried out by Shanto Iyengar, Mark Peters and Donald Kinder in

1982.[3] They wanted to know whether watching particular news stories – for instance, about environmental pollution – cause citizens to care more about those issues. They recruited some study participants, and divided them at random into *treatment* and *control* groups.

> **Treatment Group:** The group of experimental subjects that receives the treatment

> **Control Group:** The group of experimental subjects that does not receive the treatment.

They were asked to come, for several evenings in a row, to one of two specially designed rooms at Yale University. The participants were told that they were watching the evening news. But unbeknownst to them, Iyengar and colleagues edited the evening news first. The treatment group watched programs with heavy coverage of environmental pollution, while the control group saw news about other issues. The result of the experiment was that the treatment group, after the four nights of news viewing, reported much greater concern for pollution than the control group.

Does this solve the problem of selection bias? Yes! The reason why is that the study participants couldn't assign themselves to treatment. People who cared more about the environment couldn't select into news channels that talked about pollution a lot. Because they were assigned randomly to treatment and control, environmentalists were not systematically more likely to take the treatment. As we'll cover in more detail below, the treatment and control groups should be exactly the same as each other, in expectation. This means that the control group serve as a counterfactual. They tell us what would have happened to the treatment group, had they not received the treatment. Therefore the treatment effect is simply the average difference between the treatment and control groups in the outcome variable. The key point here is that the authors themselves exercised control over treatment assignment by carefully designing the study. This control is what distinguishes experiments from observational studies.

A major advantage of this strategy of experiments is their relative simplicity. Iyengar, Peters and Kinder's claim that "X causes Y" rests on far fewer assumptions than in regression analysis. Their identification strategy mainly relies on randomization being carried out properly. This is much easier to achieve than accounting for all possible confounders in a regression. If it worked, then any differences in outcome between the treatment and control group should be attributable only to the treatment – and not to pre-existing differences between the two groups. Thus we can be pretty certain that watching the news programs genuinely changed people's self-reported concern for pollution. We say that this experiment has much higher *internal validity* than the multiple regression setup.

> **Internal Validity:** How confident are we that our causal findings are correct?

However, this high internal validity comes at the expense of lower *external validity*:

> **External Validity:** How confident are we that our causal findings can be generalised to other populations, times, settings and treatment variables?

[3] Iyengar, Peters and Kinder (1982). "Experimental Demonstrations of the 'Not-so-Minimal' Consequences of Television News Programs." *American Political Science Review* 76 (4): 848-858

Most obviously, the experiment took place in an artificial setting. In the real world, people watch TV in their own living rooms. They may be distracted, or not watching properly, whereas the experimental participants may have felt the need to watch more carefully. Thus the experimental effect may be larger than would occur in a more 'real-world' setting. The experiment was also carried out in only one town: can we be sure that the findings would apply to people from other towns? More broadly, exposure to news in the real world takes place over many years, but the experiment only lasted four nights. The experimental treatment effect, therefore, may be different to the one that we really care about as researchers.

This illustrates a crucial point, that we'll come back to again and again: causal analysis involves *trade-offs*. Procedures with high internal validity, like experiments, often lack external validity.

1.3. Natural Experiments

A third possibility is to look for randomness in the real world. Della Vigna and Kaplan (2007) wanted to find out whether viewing Fox News causes Americans to become more supportive of the Republican Party.[4] Again, causation is difficult to disentangle here. Does Fox News make people conservative, or is it just that conservative people choose to watch Fox News? They found a neat trick to get around that problem. In the year 2000 – when George Bush Jr. was elected for the first time – Fox News was being gradually rolled out across the country. Some Cable providers had started to offer it, and some hadn't. They make a strong case that, within a given geographical area, this occurred essentially at random. Some cable providers had re-negotiated contracts to offer Fox News, and others had not. Such a situation, where a treatment varies at random in the real world, is known as a *natural experiment*

> **Natural Experiment:** Occurs when a treatment variable varies randomly in the real world, without the intervention of a researcher

To uncover the impact of Fox News, the authors essentially compare the vote for George Bush in places that had received Fox News by the time of the election and places that hadn't. They find a sizeable effect: an increase of about 0.5 percentage points – enough to have given Bush an advantage in close states.

Natural experiments are now enormously popular in the social sciences. One major reason is that they seem to combine many of the advantages of both multiple regression and randomised experiments. Unlike many experiments, they occur in real, natural settings that researchers care about. The viewers in this study were watching Fox News at home, not in a lab. And if cable provision really did vary randomly, we can have a high degree of confidence that this is a true causal effect. Thus natural experiments might be a 'best of both worlds', with relatively high internal and external validity. The key problem though, is whether or not the treatment really varied at random. Can we be sure that Fox News didn't try to expand into Republican areas first, where it was likely to be more popular? Making sure that this (and other) assumptions are really fulfilled takes up a lot of time and effort when analysing natural experiments, as we will discover in this course.

[4] Dellla Vigna and Kaplan (2007). "The Fox News Effect: Media Bias and Voting." *Quarterly Journal of Economics* 122 (3): 1187-1234

1.4. Why Studying Causal Inference Matters

These are not just academic debates. Thinking clearly about causation in the social sciences has major implications for the way we conduct public policy and make public health recommendations, to name but two possibilities. This week's readings include a sobering story from the New York Times on the use of estrogen hormone replacement therapy (HRT). It was once thought to prevent heart attacks in women, on the basis that in surveys, women who took estrogen were much less likely to have heart attacks. It was widely prescribed to women of all ages. Later, this finding was over-turned by a randomised control trial that administered estrogen to a treatment group and a placebo to a control group. It turned out that in fact, estrogen was very harmful to women, and considerably raises the risk of heart disease among most groups.[5] A conservative estimate may be that tens of thousands of women have died as result of faulty evidence that relied on observational research.

This is a dramatic example, but it's not uncommon. Many causal claims that we see reported in the media are based on deeply questionable evidence and research strategies, especially when it comes to diets, exercise and health. Vitamin C, for example, was once thought to prevent heart disease, amongst many other supposed benefits. But randomised control trials have found no impact at all. Very occasionally, observational research in public health is so startling that a causal effect seems unquestionable. Smoking and lung cancer is the most famous example, which (for obvious ethical reasons) has never been subject to a randomised clinical trial. Lung cancer was virtually non-existent before smoking became widespread, and smokers have a risk of lung cancer that is around 2-3,000 times that of non-smokers. But examples like this are rare. For the most part, epidemiologists study effects that may be small and hard to detect, and rely on observational evidence. The problem is that it's virtually impossible to control for every confounder that might matter, and epidemiological studies may therefore be subject to what is called 'healthy-user bias' – another name for selection bias. A quote from the New York Times article on this week's reading list sums it up nicely:

> "At its simplest, the problem is that people who faithfully engage in activities that are good for them – taking a drug as prescribed, for instance, or eating what they believe is a healthy diet – are fundamentally different from those who don't. One thing epidemiologists have established with certainty, for example, is that women who take H.R.T. differ from those who don't in many ways, virtually all of which associate with lower heart-disease risk: they're thinner; they have fewer risk factors for heart disease to begin with; they tend to be more educated and wealthier; to exercise more; and to be generally more health conscious. Considering all these factors, is it possible to isolate one factor – hormone-replacement therapy – as the legitimate cause of the small association observed or even part of it?"

Nor is healthy-user bias the end of the story, as the NYT story makes clear. There is also a "prescriber effect": doctors don't tend to give medicines to the most hopeless cases, and an "eager patient effect": patients who are already likely to be healthy are more likely to demand medicines from their doctors. The bottom line is that we should be extremely cautious about making bold causal claims from observational data – and that flashy news stories about the latest diet fads are more often than not simply wrong.

[5] The exception may be women going through the menopause, for whom HRT is very valuable.

Randomised control trials are having a major impact outside of public health, too. Across the social sciences, multiple regression was pretty much the *only* technique of causal inference in use even thirty years ago. This has changed. In public policy and development, in particular, there is now huge skepticism about policy prescriptions that are based only on a combination of theory and observational evidence. Instead, *program evaluation* has become a new paradigm. By testing different policies out experimentally, we can find out what works, and what doesn't. Many governments and international agencies now devote a lot of resources to program evaluation, and it is increasingly difficult to publish research in journals that is based on observational evidence alone. Economists call this the 'credibility revolution', by which they mean an attempt to provide a much firmer evidence base for social science theories. We'll spend the rest of the course finding out how different social scientists are going about this – and how dificult it is to do in practice.

2. Causation and Experiments

Here, we'll develop a model of causality, and apply it to experiments

2.1. Counterfactuals and Potential Outcomes

Social scientists – and to a large degree, philosophers – now largely agree on what it means for a treatment D to cause Y. Causality is about counterfactuals: if D causes Y, then in the absence of D, Y would not have happened. D is a binary variable, here. For example, D indicates taking a drug (=1) or not taking a drug (=0). Then,

- Treatment: $D_i = \begin{cases} 1 & \text{if person } i \text{ received the treatment} \\ 0 & \text{otherwise} \end{cases}$
- Observed outcome: Y_i

The subscript "i" indicates person, or more generally, the unit of analysis. With N people in the data, the dataset looks like this:

$$\{D_1, D_2, ..., D_N\} \quad \text{and} \quad \{Y_1, Y_2, ..., Y_N\}.$$

We'll focus for now on any one individual i in the dataset. What is the causal effect of D for this person? To answer this, we need to think counterfactually. For that person, we only observe one outcome, $Y_i(d)$: either they got the treatment, or they didn't. But if they *hadn't* been treated – if they'd been in the control group instead – their outcome might have been different. Thus we say that each individual has two *potential outcomes*:

$$Y_i(d) = \begin{cases} Y_i(1) & \text{Potential outcome for unit } i \text{ with treatment} \\ Y_i(0) & \text{Potential outcome for unit } i \text{ without treatment} \end{cases}$$

That is, $Y_i(d)$ = the value of the outcome that would be realised if unit i received the treatment D, where $D = 0$ or $= 1$. The causal effect τ_i of D for i is therefore the difference in potential outcomes under treatment and control. Equivalently if i is in the treatment group,

the causal effect for i is the difference between what actually happened to i given $D = 1$ and what would have happened to i if $D = 0$:

$$\tau_i = Y_i(1) - Y_i(0)$$

The key problem that motivates the entire field of causal analysis is that we can never directly observe this causal effect. This problem is known as

> **The Fundamental Problem of Causal Inference**: the same unit can never be in both the treatment and control conditions at the same time. We observe only one potential outcome (the observed, realised outcome), but we need to know both.

This an issue of missing data. For each individual, we only observe half of the data that we need. We know the actual outcome under 1 or 0, but not the **counterfactual** outcome under the opposite assignment.

2.2. The Average Treatment Effect, Selection Bias and Random Experiments

To begin filling in this missing data, instead of focusing on causal inference at the individual level, we look at everyone in our dataset, calculating the *Average Treatment Effect*. In expectation, for any unit i this is equal to:

$$\tau_{ATE} = E[Y_i(1) - Y_i(0)]$$

An unbiased estimator of this quantity is:

$$\tau_{ATE} = \frac{1}{N}\sum_{i=1}^{N}[Y_i(1) - Y_i(0)]$$

That is, we focus on the average difference in potential outcomes across all units. Just like in a linear regression, the average relationship across all units provides an unbiased estimate of the outcome for any single unit.

But we haven't fully solved the fundamental problem yet. This ATE still includes the same units under both treatment and control. One way to think about everything we will study for the rest of this module is that it represents various attempts to come as close as possible to estimating this theoretical, unobservable quantity. To understand ATEs and causation more precisely, we'll use an example. Suppose our data consist of 6 units, and we observe the following:

i	D_i	Y_i
1	1	3
2	1	1
3	1	2
4	0	0
5	0	2
6	0	1
$E[Y_i \mid D_i = 1]$		2
$E[Y_i \mid D_i = 0]$		1
$E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0]$		1

Here, our estimated ATE is equal to 1, if we simply take the observed difference in means between the treated and control units. But this is not necessarily equal to the true ATE, which is defined in terms of *potential* outcomes, not observed outcomes. Now, let's add the potential outcomes that we actually observe to the table:

i	D_i	Y_i	$Y_i(1)$	$Y_i(0)$	τ_i
1	1	3	3	?	?
2	1	1	1	?	?
3	1	2	2	?	?
4	0	0	?	0	?
5	0	2	?	2	?
6	0	1	?	1	?

This lays bare the fundamental problem of causal inference. Only one half of the necessary potential outcomes are revealed to us in any study. We cannot know what the outcome of the treated units would have been if those same units had been in the control condition. Now, hypothetically, let's add the unobserved potential outcomes to the table:

i	D_i	Y_i	$Y_i(1)$	$Y_i(0)$	τ_i
1	1	3	3	3	0
2	1	1	1	1	0
3	1	2	2	1	1
4	0	0	0	0	0
5	0	2	2	2	0
6	0	1	1	1	0
$E[Y_i(1)]$			1.5		
$E[Y_i(0)]$				1.33	
$E[Y_i(1) - Y_i(0)]$					0.17

The true ATE of 0.17 is much smaller than the ATE of 1 calculated from observed data alone. The problem is that the average potential outcomes are not the same for the treatment and control groups in this particular study. In fact, if you look carefully, you can see that the treatment group have higher potential outcomes than the control group *under both treatment*

and control. They would have had a relatively high outcome even in the absence of the treatment. This is an example of selection bias! It corresponds to a situation where healthier people (people with higher potential outcomes) have selected into the treatment. As in the example of HRT therapy and heart attacks, the resulting ATE from a naive comparison of the observed groups is much too large. Mathematically, the observed comparison yields:

$$\begin{aligned}\tilde{\tau} &= E[Y_i|D_i=1] - E[Y_i|D_i=0] \\ &= E[Y_i(1)|D_i=1] - E[Y_i(0)|D_i=0] \\ &= \underbrace{E[Y_i(1)-Y_i(0)|D_i=1]}_{\text{Causal Effect for the treated}} + \underbrace{E[Y_i(0)|D_i=1] - E[Y_i(0)|D_i=0]}_{\text{Selection Bias}}\end{aligned}$$

When is the selection bias zero? If and only if selection into treatment is not associated with potential outcomes: then, $E[Y_i(1) - Y_i(0)|D_i = 1]$ is the same as the ATE for untreated units, because the two groups have equal potential outcomes under treatment. Likewise, $E[Y_i(0)|D_i = 1] = E[Y_i(0)|D_i = 0]$. For instance, people who were healthier to begin with are not more likely to select into taking a drug. Or people who were more Republican to begin with are not more likely to select into watching Fox News. Note that zero selection bias is the same thing as zero omitted variable bias. Omitted variable bias occurs when, for example, the treatment group are healthier than the control group: "health" can be thought of as an omitted variable that is correlated both with the outcome and with the independent variable of interest (in this case, a variable that equals 1 for the treatment group and 0 for the control group).

In our hypothetical experiment, zero selection bias would correspond to the following type of situation:

i	D_i	Y_i
1	1	3
2	1	1
3	1	2
4	0	1
5	0	1
6	0	3
$E[Y_i \mid D_i = 1]$		2
$E[Y_i \mid D_i = 0]$		1.66
$E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0]$		0.33

i	D_i	Y_i	$Y_i(1)$	$Y_i(0)$	τ_i
1	1	3	3	3	0
2	1	1	1	1	0
3	1	2	2	1	1
4	0	1	1	3	-2
5	0	1	2	1	1
6	0	3	3	1	2
$E[Y_i(1)]$			2		
$E[Y_i(0)]$				1.66	
$E[Y_i(1) - Y_i(0)]$					0.33

Here, the treatment and control groups have the same average potential outcomes under both treatment and control. This means that a simple comparison of the mean *observed* outcomes between the two groups (the first table) yields the *actual* true Average Treatment Effect (the second table).

Now we can see why random assignment in an experimental setting solves selection bias. It automatically ensures that treatment assignment is not associated with potential outcomes. Healthier people are no more likely than unhealthy people to be assigned to take the drug. Another way to think about this is that random assignment creates two groups that are essentially interchangeable. The control group serves as a counterfactual for the treatment group, telling us what would have happened to the treatment group had they not received the treatment. In the same way, the treatment group serves as a counterfactual for the control group, telling us what would have happened to the control group if they had received the treatment.

It is important to note that this interchangeability only holds in expectation. In any one randomisation, we might accidentally end up with healthier people in the treatment group. But if we kept randomising over and over again, in expectation the treatment and control groups will have the same average characteristics.

3. Implementing and Analysing Experiments

Randomisation has a particular technical meaning in experiments: assignment to the treatment group must be unrelated to units' potential outcomes. Suppose we want to randomly assign treatment to a subset (say half) m units from a possible N. Some schemes that look like randomisation may be non-random. In particular, *alternation* schemes are risky. If you go through the N units in order, and assign unit 1 to treatment, unit 2 to control, and so on, this is not a random assignment scheme if there is some reason why the units appeared in an order to begin with. The even numbers may have different potential outcomes to the odd numbers. Even more obviously, the same might be true if you select the first half of a list into treatment and the second half into control. Instead, randomisation should proceed by:

1. Assign each unit a random number
2. Sort the units in order of these random numbers
3. Assign the first m units to treatment and the remaining $N - m$ units to control

Assuming that this randomisation is carried out properly, randomised experiments turn out to be a powerful tool for making causal inferences. They are also remarkably simple to analyse. For instance, we don't need any control variables at all (although we'll come back to this next week). We don't need to measure or control for any confounders, because randomisation ensures that the treatment and control groups are on average the same as each other in every respect except for assignment to treatment or control.

After carrying out the experiment, all we need to do is calculate the average treatment effect from our data by subtracting the average outcome in the control group from the average outcome in the treatment group. Due to randomisation, this *difference in means* estimator is an unbiased estimator of the true average treatment effect. Suppose we have a randomised

experiment with N units, where m of them are assigned to treatment and $N - m$ to control. Then, the ATE can be calculated as:

$$\hat{\tau} \;=\; \frac{1}{m}\sum_{i=1}^{m}(Y_i \mid D_i = 1) - \frac{1}{N-m}\sum_{i=m+1}^{N}(Y_i \mid D_i = 0)$$

where the first half of the equation is the average outcome for treated units, and the second half is the average outcome for control units.

Of course, in general we only observe a sample of the whole population that we're interested in. We might want to know about the efficacy of a drug in the whole population, but we can usually only experiment on a small sample from that population. For that reason, the difference in means estimator has a *sampling distribution*, whose standard error we also need to estimate from the data. To do so, we use the estimator:

$$\hat{SE} \;=\; \sqrt{\frac{var(Y \mid D = 1)}{m} + \frac{var(Y \mid D = 0)}{N-m}}$$

This should look familiar. All we are in fact doing is carrying out a t-test for the equality of two means. Inference can then be carried out in the usual way: divide the estimated ATE by its estimated standard error, and compare the resulting statistic to the critical value at some pre-determined significance level. If our t-statistic exceeds that level in absolute terms, we conclude that the treatment has a statistically significant effect on the outcome.

4. Examples of Experiments

On the face of it, experiments should be very simple to analyse. That is one of their virtues. But they are often complex to administer. These practical issues can, in turn, increase the complexity of analysis. To illustrate this, we'll start by introducing some real-life experiments. In the next lecture, we'll discuss how to overcome some of these problems.

4.1. A Simple Survey Experiment

We'll start with an example from my own research. I am currently working on a book about reforms to welfare provision in the UK over the past thirty years - programs that provide benefits to the unemployed, disabled and those living in poverty. As part of that, I wanted to find out how politicians' discourse about welfare recipients affects public opinion about the benefits system. One way that I examine this is with a *survey experiment*: an experiment that is embedded within a public opinion survey. As part of an online survey, participants were (not to their knowledge) assigned to treatment and control groups, where the treatment group read a political speech that was very critical of welfare provision, and the control group read a more positive speech. Afterward, I asked them opinion questions about the welfare state, and found that the group who read the negative speech expressed much more negative opinions about benefits and those who rely on them for income.

I'm highlighting this not because the experiment is particularly exciting or ground breaking, but because it provides an example of an experiment that is extremely simple both to design and analyse. The main reason is that I had extremely close control over the experiment:

- It was very simple to randomly divide the group into treatment and control. The computer simply did it via standard software.
- There was no possibility of non-compliance: no-one could shift themselves from one experimental group to another.
- Because it was not very onerous, there was virtually no attrition from the experiment.
- It was not possible for the people in the experiment to interact, and therefore influence each other in ways that I did not intend.
- It was very easy to make sure that they actually complied with the experiment: all I had to do was ask some factual questions to make sure that they had read the speeches. Almost all of them showed evidence of having read and absorbed the material.

4.2. De-Worming in Africa

Intestinal worms such as ringworm affect hundreds of millions of people worldwide, many of them children in developing countries. They cause malnutrition and anaemia, and are widely thought to have a negative impact on education by making it more difficult for children to attend school or concentrate on their work. They are also extremely cheap and easy to treat with regular doses of deworming drugs in affected areas. Development economists and epidemiologists have had a robust debate about whether deworming programs should be widely rolled out by development organisations.

A set of experiments has tried to test this. Initial results were not positive. Several projects went to individual schools and randomly assigned some children to treatment (deworming drugs) and others to control. They found very limited evidence of benefits to the treated (in terms of increased school attendance, etc.). However, such studies may severely underestimate the true ATE because there are *spillover* effects, or "externalities" in the parlance of economics. Worms are contagious and are spread through contaminated water. Therefore providing deworming drugs to the treatment group is likely to provide benefits to the control group as well, because it reduces the control group's probability of infection. The difference in means after the intervention (in, say, school attendance) is likely to underestimate the true average ATE.

To get around this, a very famous paper by Miguel and Kremer looked at a project that randomly provided deworming by *school* instead of individual.[6] Schools were situated in different communities that were not close to each other. All pupils within treated schools received the treatment and all pupils within control schools did not. By comparing treated schools to untreated schools in different locations, they were able to capture the spillover effect of deworming, since pupils within a school could influence each other, but pupils across schools could not. They found very large effects on school attendance. Rates were substantially higher in treated versus untreated schools. Deworming is also a much cheaper way to boost attendance than other potential policies. The paper has had a major impact on development practice, with development agencies rolling out big increases in deworming for primary school children across Africa.

[6]Edward Miguel and Michael Kremer (2004). "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities" *Econometrica* 72 (1): 159-217. Note that the results of this paper have been subsequently disputed and remain controversial.

4.3. The US Job Training Partnership Act

In 1983, the US government began a major new active labour market policy, the Job Training Partnership Act (JTPA), which aimed to boost the earnings of low-income and unemployed people, mainly young people, through intensive training. At the same time, Congress mandated that an evaluation be carried out of the scheme's effectiveness: there was skepticism, especially amongst conservatives, on whether adult training programs really deliver value for money. This became perhaps the largest randomised control trial of a public policy ever carried out in a developed country. Around 20,000 adults who wanted to enrol in training were randomly assigned to either treatment (training) or control (no training), and were followed up for a number of years. The outcome of interest was the difference in subsequent earnings between the two groups.[7]

This was incredibly complex to administer in practice. For a start, the administrators found it extremely difficult to persuade any training sites to actually enrol in the trial. Nobody wanted to have to randomly deny training to half of the people who showed up. Eventually, sixteen sites did sign up. But then, the problem was that not everyone who was assigned to treatment actually received it. A large number of them dropped out before they actually received any training. Intuitively, a failure to account for this *non-compliance* is likely to overstate the impact of the training program. It seems likely that those who were assigned to the treatment group and dropped out were lower 'quality' (had lower potential outcomes) than those who remained in the program. Thus the program ended up being given to people who were more likely to benefit from it in the first place.

In fact once this non-compliance was adjusted for (we'll learn how to do that in week 6), the JTPA turned out to have rather disappointing results. It had only a small positive effect on the earnings of adults as a whole, and an effect that was indistinguishable from zero for young people, for whom the programme was mainly intended. Following the randomised control trial, Congress drastically cut back spending on the JTPA.

4.4. The Tennessee STAR Experiment

The economist Alan Krueger wrote a well-known evaluation of an experiment in education policy, which is on the reading list for next week.[8] This was another large-scale public policy experiment in the US. Education researchers have long debated whether smaller classes – which are expensive to bring about – actually lead to better educational outcomes. The Tennessee Star experiment set out to test this by randomly assigning schoolchildren from kindergarten into 'small' or 'regular' class sizes for several years. The children were tracked to see whether those in smaller class sizes got better grades.

This was, perhaps unsurprisingly, not only controversial but also very difficult to administer in practice. There were obvious incentives for parents in the control group to not comply with randomisation, attempting to push their children into the treatment group. Likewise, they had an incentive to take their children out of the school they were in and try to enrol them

[7]Several published evaluations exist. The most well-known is probably Howard Bloom *et al* (1997). "The Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Training Partnership Act Study." *Journal of Human Resources* 32 (3), pp. 549-576.

[8]Alan Krueger (1999). "Experimental estimates of education production functions." *Quarterly Journal of Economics* 114 (2): 497-532

in another school with smaller class sizes. Indeed, the results suggest that movement from the control group to the treatment was more common than the other way around, and that attrition was slightly more common in the control group. Once again, this matters because children with higher potential outcomes – with pushier parents – may have been more likely to leave the control group. This would bias upward the average ATE, because it results in a lower 'quality' control group.