LECTURE 3. RANDOMISED EXPERIMENTS: INTERNAL VALIDITY

Dr. Tom O'Grady

Note: Some of this lecture is based very closely on Gerber and Green's textbook, but with simplifications. In particular I use their terminology and equations. It is a good idea to read this alongside their book.

1. Things that can go Wrong: The Concept of Internal Validity

Concerns about the robustness of experimental results are all part of what we call *internal validity*. An experiment has high internal validity if we are very sure that the findings are genuinely due to the experimental treatments we have administered. Another way of putting this is that in an internally valid experiment, we have very high certainty that we have uncovered a genuine causal effect of the treatment. The discussion of the four different experiments at the end of Lecture 2 suggests several ways in which experiments can fail to uncover a true causal effect, or might uncover a causal effect that is different from the one that we intended to study:

- *Randomisation Failures*: Randomisation may not be carried out properly
- *Non-Compliance*: People may not comply with their assigned treatment or control
- *Attrition*: People may leave the experiment before it has finished
- *Interference*: The treatment and control groups may have effects on one another

Remember that the key strength of a properly-run experiment is that it allows us to compare two groups (treatment and control) who in expectation have exactly the same potential outcomes under treatment and control. The first three issues may all alter this useful property. They could result in us unwittingly comparing groups whose potential outcomes may be very different, rather than the same. We will be unable to recover the true average treatment effect The fourth problem does not result in treatment and control groups with different potential outcomes at the outset of the experiment. However, it still prevents us from obtaining the true average treatment effect because the two groups can affect each others' outcomes once the experiment is under way, as in the initial de-worming experiments within schools that we talked about at the end of Lecture 2, which under-estimated its impact.

When designing or analysing an experiment, or reading an experimental paper, it's important to consider carefully whether or not these problems are likely to exist. Tests for randomisation, and statistics like compliance and attrition rates, should always be reported. Note that there is a key difference to regression analysis, where our concerns tend to be about modelling: whether we have the correct functional form, whether we should be using logit or probit estimators, and so on. Here, the concerns are about the appropriate *design* and *operation* of our experiment. Did we set it up properly? And having done so, did it run as we intended?

Now, we'll consider each concern in turn, asking how they can be dealt with. We'll also delve briefly into the ethics of experimentation.

2. Randomisation, Balance Tests and Blocking

Randomisation is nice in theory, but can be hard to implement in practice. It is therefore important to verify that units really were randomly assigned to treatment and control. Randomisation may not occur as planned. Sometimes, for instance, the administrators of an experiment might cut corners and decide to assign treatment differently than was planned. Or more insidiously, some control units might bribe or coerce the administrators into changing their status to the treatment group.

2.1. Balance Tests and Covariate Imbalance

Therefore it is important, after an experiment, to check for such failures of randomisation. If randomisation succeeds, the distributions of both observed and unobserved characteristics should be near-identical in the treatment and control units. We say that the two groups are *balanced*. We check for balance using *balance tests* on characteristics that are unaffected by the treatment (like age, sex, etc. for individuals).

> **Balance Test**: A statistical procedure designed to test whether the treated and control units are on average the same as each other.

Characteristics that are unaffected by treatment are often called "baseline covariates."

> **Baseline Covariates**: Characteristics of the treated and control units measured before the experiment and unaffected by the treatment. Typically fixed characteristics of the units like age, gender, location, etc.

The idea is that if randomisation was carried out properly, there should not be large differences in the composition of the two groups by age, sex and so on. Thus even in an experiment, it is important to collect covariates for the purpose of balance tests. Balance tests can take several forms:

1. T-tests for equality of means in the baseline covariates between the treatment and control groups

2. Regress the treatment variable on the baseline covariates and test their individual and joint significance in explaining treatment assignment

3. Graphical comparisons of the densities of baseline covariates between treated and untreated units

In the first two cases, we are hoping to find results that are statistically indistinguishable from zero. In the third, we hope to find that both densities are near-identical. Then, there is strong evidence that randomisation was carried out correctly, but not definitive proof. There could still be imbalance on unobserved variables.

On the other hand, finding significant differences between the two groups is not definitive proof of failure, either. Random assignment ensures that the treatment and control groups have the same potential outcomes *in expectation*. If we kept randomising over and over again, there should be no systematic differences between the two groups. But in any one instance of randomisation, it is perfectly possible for some differences to emerge between the two groups due to chance alone. This is particularly true in small samples. Thus balance tests are best thought of as the beginning, not the end, of an investigation. If large covariate imbalances are detected in an experiment, it would be a good idea to investigate how the randomisation actually took place.

In cases where some minor covariate imbalances are detected, but there does not appear to be a violation of randomisation, it is standard practice to analyse the experiment with a regression that includes both the treatment variable and the baseline covariates as independent variables. The coefficient on the treatment variable is the average ATE, 'purged' of any influence from the imbalanced background covariate. The standard error on the treatment variable will also usually be lower: we investigate this below.

2.2. Block Randomisation and Standard Errors

An obvious question is whether there is some way in which we can avoid creating serious imbalances by accident. Block randomisation, also known as stratified randomisation, is a principled way to do this. The idea is to pre-stratify the sample for an experiment, and then randomise separately within each stratum (block). Intuitively, this is like running a separate experiment within each stratum. Each unit still has an equal chance of ending up in either treatment or control, but we can avoid certain randomisations that would lead to serious imbalances in a particular instance.

To see this, suppose we have 50 people, and we want to randomly allocate half of them them to treatment and half to control. Suppose we collect data on their incomes before the experiment. Without blocking, there is one possible random allocation where all 25 of the richest people get the treatment, and all 25 of the poorest are the control group. Now, suppose that instead, we pre-stratify the sample into five income quintiles, and carry out a block randomisation scheme where within each quintile, people are randomly allocated to treatment or control. Then, we are guaranteed to have an outcome where 5 people from every quintile are in the treatment group, and 5 are in the control group. It is no longer possble to have a 'crazy' outcome where all 25 of the treatment group are the richest people. Nonetheless, this is still a randomised experiment, because every unit has an equal probability of being in the treatment or control group. In expectation, the treatment group will not have systematically different potential outcomes to the control group.

Such block randomisation is likely to reduce the standard error of the estimated ATE, so long as income is correlated with the outcome. It does this by removing 'crazy' randomizations from

the sampling distribution of the ATE. In the example above, the randomization outcome where the 25 richest people end up in the treatment would produce an outlandishly high difference-in-means estimate, if income is correlated with the outcome. In that case, a large part of this outlandish ATE estimate would be due to the influence of income, not the treatment. The true ATE could still be recovered by controlling for income in a regression of the outcome variable on the treatment variable. Block randomisation, on the other hand, prevents us from needing to carry out any regression in the first place.

In fact, in small samples of less than 100 units or so, block randomisation will usually lead to a substantially lower standard error for the ATE. This is typically the main motivation for blocking. To see why, we'll shift from thinking in terms of the sampling distribution of the ATE to thinking about the outcomes of any one experiment - which we always use as an unbiased estimate of the variance of the sampling distribution in any case. Suppose we have an experiment with a treatment variable D, an outcome Y and a background covariate X that is stratified into K blocks. Without accounting for blocking, we might run a simple regression to find the ATE:

$$y = \beta_0 + \beta_1 D + \epsilon_r$$

In this case without blocking, $var(\beta)$ is

$$\frac{\sigma_r^2}{\sum_i (D_i - \bar{D}_i)^2}$$

where σ_r^2 comes from the residual without blocking, i.e. $\frac{1}{N-1}\sum(\epsilon_r^2)$. Here, the standard error will be relatively high, because some of the variation in Y is due to unmodeled influences – including X – that are included in ϵ_r^2.

Now, suppose that X contains a set of dummy variables for each block, and the blocks are predictive of the outcome. Suppose further that we don't carry out block randomization, but instead simply include X in the regression after our experiment. We run the regression:

$$y = \beta_0 + \beta_1 D + \gamma X + \epsilon_b$$

Now, $var(\beta_D)$ is:

$$\frac{\sigma_b^2}{(1 - R_D^2)\sum_i (D_i - \bar{D}_i)^2}$$

where σ_b^2 is $\frac{1}{N-K-1}\sum(\hat{\epsilon}_b^2)$, and R_D^2 is the R^2 that comes from regressing the treatment indicator on all the other covariates. This shows us that when the block indicators X are introduced, two things happen to the variance of the coefficient on treatment. If the blocks are successful in explaining changes in the outcome variable, then the residuals will be much smaller, and thus $\sigma_b^2 < \sigma_r^2$, making the overall variance of the coefficient much smaller. At the same time, there is the risk that including other covariates can *increase* the variance because you have the $(1 - R_D^2)$ in the denominator. This is what Gerber and Green mean when they refer to the "collinearity penalty."[1] High collinearity in the independent variables makes their coefficients

[1]This is exactly the same reason why collinearity in multiple regression often leads to falls in the precision of estimates.

more uncertain. When is collinearity higher? When R^2_D is higher, which occurs when there is high covariate imbalance. High imbalance means that the block indicator variables are highly correlated with treatment assignment.

It turns out that as sample sizes increase to over 100, the benefits from reducing σ^2 far outweigh the countervailing influence from R^2_D. The reason is that high imbalance between the treatment and control groups is far more unlikely, the larger the sample size. See footnote 17 on page 114 of Gerber and Green for an example. Thus in large samples, it is more typical to simply control for baseline covariates in a regression when estimating the ATE, without worrying about blocking.

However, in small samples, the R^2_D penalty is high. Why does blocking help in these cases? Because it makes the R^2_D equal to zero! Suppose that in our example, we actually carried out block randomization, randomizing treatment within each of the K strata. For instance, suppose X indicates one of four age groups. Then, each of the four age groups has an equal chance of receiving treatment: there is no relationship at all between D and X, making R^2_D equal to zero. As long as the block variables are strongly correlated with the outcome, we get a big decrease in σ^2 without any downside. Our ATE is guaranteed to be more precisely estimated than in a case with just simple randomisation that doesn't account for the blocks. That's why block randomisation is such a good idea in small samples. For that reason, field experiments carried out on small numbers of units almost always use block randomisation whenever possible. By block-randomizing and then analysing the experiment with a regression that controls for the block indicator variables, in small samples we can ensure that the ATE is estimated as precisely as possible.

3. Interference Between Units: The Stable Unit Treatment Value (SUTVA) Assumption

The standard analysis of experiments relies on an assumption known as SUTVA: the Stable Unit Value Treatment Assumption.

> **SUTVA**: The potential outcomes of each unit i are not affected by the treatment assignment of other units in the experiment. They depend only on the treatment assignment of i.

Concretely, this is an assumption that units do not *interfere* with each other. Such interference could take many forms, but the most obvious one – and the form that most obviously leads to bias – is that the treated units affect the control units. This might include:

- *Contagion*: As in the Worms example, some of the benefits of the treatment may naturally spill over to the controls
- *Communication*: For instance, an experiment that randomly provides electoral campaign leaflets to some houses in an area, but not others. The treated households may tell the untreated households about the leaflets
- *Deterrence*: For instance, randomly carrying out corruption audits on some public officials may encourage untreated officials to be less corrupt

All three of these examples lead to the control units effectively receiving some of the treatment, and are likely to lead to provide under-estimates of the true treatment effect. This is because the control group no longer provides a valid counterfactual for what would have happened to the treatment group in the absence of treatment.

Correcting for SUTVA violations is a particularly technical area of experimental design, and we are not going to cover those techniques in this course: see Chapter 8 of Gerber and Green if you are interested. It is worth noting, though, that dealing with SUTVA is more often than not best achieved by design choices rather than complex analysis after the experiment. The Worms paper provides one example of this, by randomizing across rather than within villages. In settings where the analyst explicitly wants to avoid interference, it makes sense to try, as far as possible, to prevent the treatment and control groups from interacting. For example, making sure that treatment and control groups are kept geographically separate is often a good idea.

4. Non-Compliance

Non-compliance refers to a situation where the units in our experiment do not comply with their assignment to treatment or control. Non-compliance can be *one-sided* or *two-sided*.

> **One-Sided Non-Compliance**: occurs when some members of the treatment group take the control instead.
>
> **Two-Sided Non-Compliance**: occurs when some members of the treatment group take the control, and some members of the control group take the treatment.

For instance, in the Krueger paper on class sizes, there was two-sided non-compliance. Some children assigned to small classes ended up in regular-sized classes and vice versa. In that case, non-compliance by the control units was more prevalent, because the parents of these families had obvious incentives to try to get their children into smaller classes. In many other applications where the control group simply receives a placebo or no treatment at all, one-sided non-compliance is much more common. This is particularly true of medical trials, where we cannot ethically compel people to take the treatment (more on this below), and the treatment is usually completely unavailable to the control group. This was the case in the JTPA study, where only around 2% of the control group managed to obtain the treatment, but one third of the treatment group never actually enrolled in the training program.

Non-compliance is extremely common in experiments and should be considered part and parcel of experimenting on humans. In fact, it is relatively easy to fix – at some cost, as we'll find out – with the technique of instrumental variables, which we'll come onto in a few weeks. Intuitively, instrumental variables uses the fact that while treatment *receipt* is non-random under non-compliance (the people who choose to take treatment have different potential outcomes to those who don't), treatment *assignment* is random. In turn, this allows us to estimate average ATEs for the sub-group of people who comply with the experiment, known as "compliers." This is what Krueger does in the right-hand columns of Table V.

5. The Ethics of Experimentation

Having mentioned non-compliance, now is a good time to talk about research ethics and experimentation on humans. This is a big topic that we can't hope to cover in detail, but it is important to note that social science experiments raise very clear ethical concerns. When developing an experiment, researchers are typically required to obtain pre-approval from a university committee that investigates whether the experiment raises ethical issues. Some guidelines for experimentation include the following:

1. *Full and Informed Consent*: Human subjects can never be compelled to participate in a trial or to remain in it any point in time. They must also give consent freely to take part, and consent must be informed by a clear expanation of any risks involved. That is one reason why "non-compliance" is so common in experiments and should be anticipated in many cases. No randomised trial should ever be allowed to go ahead unless those involved can remove themselves at any point.

2. *Confidentiality*: Experiments that involve the collection of sensitive information or that may elicit sensitive behaviours must ensure that participants' data remain anonymous after the experiment. Published articles must never identify participants, and universities always require researchers to formulate a clear plan for how data will be stored and disseminated in a way that maintains this anonymity.

3. *Minimising Harms*: This is the most difficult ethical concern to deal with. Some experiments, particularly medical trials, carry risks for participants that can never be wholly mitigated. That is one reason why informed consent is so important, and why in many cases, risky new treatments for patients with conditions like cancer are typically only offered when all other treatment options have failed. A particularly thorny issue is what happens to control groups. Medical trials by their very nature might require that a control group receives a placebo, even though we might be able to afford to give them what could be a very valuable, even life-saving treatment. One minimal ethical requirement is that the control group should never be left worse off than they would have been outside of the experiment. Trials may encourage a treatment group to undertake some beneficial action (stopping smoking, for instance), but no trial would ever be allowed to encourage a control group to undertake clearly harmful behaviour, like smoking more. One way to deal with this concern is that trials will often compare a treatment group to a control group that receives the currently best-available treatment rather than a placebo, particularly when there exists reasonable doubt about whether the treatment is likely to work. This principle is sometimes called "equipoise" in the medical trials literature. Another common experimental design is to offer the treatment to everyone in the trial, but administer it in different phases to a treatment and control group. In that way, the progress of the treatment group can be compared in the first phase of the experiment to a control group who have not yet received the treatment, but will in the next phase.

We can see some of these ethical issues at work in the experiments we are looking at today. One reason why the JTPA study ran into so much trouble is that very few training providers wanted to sign up for an experiment in which they had to deny treatment to some participants. Nonetheless, the control group were not left any worse off than they would have been outside the

trial. In the same way, children in the Tennessee Star experiment's control group were enrolled in classes that were the same size as existing classes, and there did exist genuine debate about whether smaller class sizes were really beneficial or not. Finally, the Worms study did in fact use a phased treatment, where the treatment was given to three sets of randomly-selected villages in successive years. The researchers would probably never have been given ethical approval to deny treatment to the control group, because deworming is extremely cheap and is known to have many benefits for those who receive it.

These ethical dilemmas are not necessarily easy to resolve, and there will always be some who remain deeply uneasy about experimentation on humans. In return, one might argue that we should weigh any potential harms from an experiment against the potential costs of ignorance about which public policies are effective. The costs of spending vast amounts of public money on policies that do not work – or could be improved upon – could be extremely large and are borne by all taxpayers. Thus it is probably reasonable to weigh up the potential harms and gains from any experiment.

6. Attrition

Many experiments unfold over time, and some units may drop out of the experiment before it ends. Equivalently, we may need to collect outcome data some time after the experiment is over, and some units may be unavailable to provide that data. For example, in the JTPA experiment, earnings data was collected sometime after the training actually took place, and some people were difficult to trace or did not provide information. Similarly, in the Tennessee Star experiment, inevitably some families moved away from the area or moved to different schools over time, so that many students who began the study were not still enrolled in it after four years.

Attrition is ultimately an issue of *missing data*. We end up lacking information on some fraction of those who began the trial. Non-compliance, on the other hand, refers to a situation where some people end up in the opposite group (treatment or control) to their initial random assignment. When the control condition is simply the absence of any treatment, attrition and non-compliance may amount to the same thing, but this is a special case, and it is useful to consider the two as separate problems. In the Tennessee Star experiment, for example, non-compliance and attrition are separate issues. Non-compliance occurred for children who were assigned to small classes but actually ended up in a regular-sized class, and vice versa. Whereas attrition occurred for children who began the study (in either group), but subsequently left the schools being studied altogether, for whatever reason. Thus attrition involves initially complying with the experiment, but later leaving it, or complying all the way until the end, and then failing to provide outcome data.

A key difference with non-compliance is that attrition often leaves us unable to estimate anything interesting from our experiment. With non-compliance, the ATE for compliers is at least a theoretically interesting quantity. It tells us the effect of the treatment for the population of people who, in the real world, would actually take the treatment when offered it by policymakers. But there is no guarantee that we should be interested in the ATE for the group who remain in an experiment until the end, as we'll see.

6.1. Missingness, Potential Outcomes and Bias

As in the last lecture, we'll use the following notation:

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

Observed outcome: Y_i

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

We'll also introduce a new random variable, $R_i(d)$, unit i's *Reporting Status*:

$$R_i(d) = \begin{cases} R_i(1) & \text{Reporting status for unit } i \text{ with treatment} \\ R_i(0) & \text{Reporting status for unit } i \text{ without treatment} \end{cases}$$

R_i is binary, and clearly we can only observe the potential outcomes of those for whom $R_i = 1$. Therefore if we simply calculate a ATE for those that report an outcome, we obtain:

$$E[Y_i(1)|R_i(1) = 1] - E[Y_i(0)|R_i(0) = 1)]$$

When is this the same as the true ATE of $E[Y_i(1) - Y_i(0)]$? The answer is that it's true only when missingness occurs independently of potential outcomes. This means that some units are simply *Missing at Random*. In this case, a random subset of units are missing, meaning that:

$$E[Y_i(1)|R_i(1) = 1] = E[Y_i(1)|R_i(1) = 0]$$

and the same is true of the control group. Under true randomisation, here the control group is still a valid counterfactual for the treatment group, and vice versa. There is no selection bias. People with higher potential outcomes are not more likely to be in the treatment group, for example. This is difficult to test empirically. Following the logic of a balance test, we could regress units' actual reporting status on their baseline covariates and hope to find that they do not predict missingness. But this does not address whether missingness is related to unobserved characteristics.

In reality, missingness-at-random is probably very rare. Attrition is often related to potential outcomes. For instance, imagine in the JTPA experiment that "lazier" people were less likely to fill out the post-experiment questionnaire. Or in the Tennessee Star experiment, children from "higher-achieving" families were more likely to be economically mobile, and move to another city before the end of the trial. In both cases, the same traits that explain attrition are surely related to potential outcomes. Lazy people are less likely than the non-lazy to have high earnings, and children from high-achieving families are more likely than those from non-high-achieving families to have high test scores.

6.2. The ATE for Always-Reporters

What if missingness is related to potential outcomes, but occurs in the same way amongst both the treatment and control groups? Then, even though people are missing not-at-random, the treatment and control groups still remain valid counterfactuals for each other. Suppose, for instance, that in the JTPA experiment there is some subset of the population who are "lazy" and would never fill in the JTPA post questionnaire, regardless of whether they end up in treatment or control. Or, suppose in the Tennessee experiment, some subset of the population are "mobile" and would move to another area or school, regardless of what size class their child is assigned to. What we end up with in such cases is an ATE estimated for a particular subset of the population, called *Always-Reporters*.

> **Always Reporter**: A unit that always reports its outcome, regardless of assignment to treatment or control

Formally, we require that *all* units are either never missing (Always-Reporters) or always missing (Never-Reporters). The observed treatment and control groups will then contain only Always-Reporters. Thus we observe $E[Y_i(1)|R_i(1) = 1)]$ for the treatment group and $E[Y_i(0)|R_i(0) = 1)]$ for the control group, and estimate the ATE for Always-Reporters as:

$$E[Y_i(1) - Y_i(0)|R_i(0) = R_i(1) = 1]$$

Is this actually an interesting or a useful quantity to measure? It could still be a highly biased guide to the ATE in the population as a whole, and is probably of limited usefulness. Nonetheless, we might argue that if there is (say) a positive ATE for the population as a whole, there should also be one for the sub-group that always reports outcomes.

More importantly, there are probably relatively few circumstances where attrition behaviour would be the same for all units regardless of assignment to treatment or control. For example, high-achieving families were probably more likely to leave the Tennessee Star experiment if their child was assigned to a larger class than if they were assigned to a smaller class. If that is true, we can no longer estimate an ATE for Always-Reporters, because the treatment and control groups will no longer contain units with the same potential outcomes in expectation. Children from high-achieving families would be more likely to end up in the control group at the end of the experiment. Whenever missingness is related to potential outcomes and it occurs differently in the treatment and control groups, then the treatment and control groups are no longer valid counterfactuals for each other: we no longer have a randomised experiment, even for a subgroup.

6.3. Missingness Conditional on an Observed Covariate

A solution is possible in cases where we can observe the variable that directly drives missingness. Suppose in the Tennessee Star experiment that we could measure families' "high-achievingness", and that children from high-achieving families are more likely to leave the experiment. It may be that within children from similar families, missingness occurs at random. In such a case, we can use *inverse probability weighting* to recover the correct ATE:

$$\tau_{ATE} = \frac{1}{N}\sum_{i=1}^{N}\frac{Y_i(1)r_i(1)}{\pi(d=1,x_i)} - \frac{Y_i(0)r_i(0)}{\pi(d=0,x_i)}$$

where:

- $r_i(d)$ refers to the actual reporting status that occurs in condition d, and can be either =1 (unit i reports) or =0 (unit i does not report)
- $\pi(d=1,x_i)$ refers to the share (π) of non-missing units amongst those in condition d with characteristic x_i.

For example, suppose that a child in the Tennessee Star experiment can come from either a high-achieving ($x_i = 1$) or low-achieving ($x_i = 2$) family. Then, when i is in the treatment condition, i may or may not report an outcome ($r_i(1) = 1$ or $= 0$) and likewise in the control condition ($r_i(0) = 1$ or $= 0$). In the same way, a particular share (π) of all children from the same type of family as i will report data when in the treatment ($d = 1$) or control ($d = 0$) conditions.

When $r_i(d) = 1$, meaning that i reports data, observation i will be *re-weighted* to account for missingness amongst other units from the same family type. If, for example, half of all high-achieving families fail to report when $d = 1$, the remaining units will be weighted doubly (dividing by one half). Effectively, we fill in missing data from children who leave the study with children who remain in the study and share the same x. This avoids the ATE being biased by having too few high-achieving children in the final sample; we just replace the missing children by giving a higher weight to those who stay in the study.[2] Effectively, we make our data look the way it would look if all units were observed.

The key assumption here is that missingness occurs at random amongst units with the same x. Then, the potential outcomes of those who remain in the study are representative of the potential outcomes of those who leave, amongst those with the same x. If this is not the case, then bias will still occur, because the potential outcomes of those who remain in the study will be systematically different to the potential outcomes of those who leave. This assumption that one covariate (or set of covariates) can perfectly account for missingness is very strong, and is rarely likely to be satisfied. Suppose that we proxied families' "high-achievingness" with their incomes. Even amongst families with the same incomes, it is probably still the case that children from families who choose to leave the study early have different potential outcomes to children who choose to remain until the end. Thus this solution is rarely available to us in practice, unless we have a very strong understanding of what drives attrition.

6.4. Extreme Value Bounds

A final possibility – and the strategy that is actually used by Krueger to analyse the Tennessee experiment – is to accept that non-random attrition occurs, and try to place bounds on the range of possible ATEs under attrition. One way to think about this is that we ask "what is the

[2]Note that this is exactly the same sort of procedure that is used to weight survey data when survey samples are not representative of the general population on some observed characteristics.

worst that could happen?" and then re-calculate our ATE under these worst-case scenarios. Specifically, we create an *upper bound* and *lower bound* on the ATE in the following ways:[3]

1. To calculate the upper bound, replace all missing values in the control group with the *lowest observed outcome* and all missing values in the treatment group with the *highest observed outcome*

2. To calculate the lower bound, do the opposite: replace all missing values in the control group with the *highest observed outcome* and all missing values in the treatment group with the *lowest observed outcome*

Alternatively, instead of using the lowest and highest *observed* outcomes, we might use the lowest and highest *possible* outcomes, which is what Krueger does, replacing the score for missing children with either 0 or 100, the range of possible grades on the test administered to students. Either way, the idea is to see whether we would have a treatment effect that is statistically distringuishable from zero under these two extreme scenarios. If under both of these scenarios the treatment effect remains significant and in the same direction as the original ATE, this should raise our confidence that attrition doesn't endanger the results. After all, real attrition is unlikely to occur in as extreme a way as either of the modelled scenarios.

Therefore this procedure places bounds on the possible values of the ATE, and will typically contain the true ATE within them. It is an intuitive and easily implemented procedure, but tends to work well only when attrition occurs on quite a small scale. The problem is that with even quite moderate amounts of attrition, the extreme value bounds will often be so wide as to be virtually meaningless. For experiments with small amounts of attrition, though, it is sensible to calculate bounds.

7. Design Trumps Analysis

The major advantage of experiments, in theory, is that they allow us to uncover causal effects with minimal assumptions needed, very little reliance on modelling to make inferences, and very simple analysis. But as we saw today, many of those advantages quickly crumble away when we face problems like randomisation failures, interference between units, non-compliance and attrition. For example, solving the problem of attrition quickly got us back into a reliance on strong assumptions or modelling.

This has implications for the way we design experiments. Ultimately, it makes most sense to anticipate problems up front and design experiments in such as way that they are robust to these issues. This will tend to help us get closer to the simple, clean ideal world of experiments. The Worms paper provides one example of a design-based solution to interference, while blocking provides a design-based solution to problems of randomisation in small samples. In the same way, it probably makes most sense to avoid designing experiments that are likely to be subject to a large amount of non-compliance or attrition, or to at least take steps in advance to minimise them wherever possible. In this respect, "design trumps analysis": in the *design-based inference* perspective of this course, good research design in advance is almost always the best way to solve problems.

[3]This assumes an ATE that is expected to be greater than zero; the calculations would be reversed for a negative ATE