Section 29

Lecture 12

Section 30

IV inequalities

Motivation: IV inequalities

Theorem (IV inequalities)

Suppose $Z \perp\!\!\!\perp Y^a$, positivity and consistency hold. Then,

$$\begin{aligned}
P[Y=0,A=0 \mid Z=0] + P[Y=1,A=0 \mid Z=1] &\leq 1;\\
P[Y=0,A=1 \mid Z=0] + P[Y=1,A=1 \mid Z=1] &\leq 1;\\
P[Y=1,A=0 \mid Z=0] + P[Y=0,A=0 \mid Z=1] &\leq 1;\\
P[Y=1,A=1 \mid Z=0] + P[Y=0,A=1 \mid Z=1] &\leq 1.
\end{aligned}$$

The idea is that the instrumental variable assumptions put constraints on the joint law $p(y, a, z)$. This is interesting, because , in principle, we can use these logical bounds to use evaluate the IV assumptions: we can derive a test of whether the IV assumption $Z \perp\!\!\!\perp Y^a$ holds. If any of the above inequalities fail, then the core conditions must be violated; however, it is possible that the core IV conditions are violated without failing the inequalities.

Cont

Proof.

For $i, j, k \in \{0, 1\}$,

$$
\begin{aligned}
& P[Y^{a=i} = j] \\
& = P[Y^{a=i} = j \mid Z = k] \quad \text{bc. } (Z \perp\!\!\!\perp Y^{a}) \\
& = P[Y^{a=i} = j, A = i \mid Z = k] + P[Y^{a=i} = j, A = 1 - i \mid Z = k] \quad \text{laws of prob.} \\
& = P[Y = j, A = i \mid Z = k] + P[Y^{a=i} = j, A = 1 - i \mid Z = k] \quad \text{const.} \\
& \leq P[Y = j, A = i \mid Z = k] + P[A = 1 - i \mid Z = k] \\
& = 1 - P[Y = 1 - j, A = i \mid Z = k];
\end{aligned}
$$

Thus

$$
\begin{aligned}
\max_{k} P[Y = 1, A = i \mid Z = k] \leq P[Y^{a=i} = 1] \\
\leq \min_{k^*} 1 - P[Y = 0, X = i \mid Z = k^*],
\end{aligned}
$$

where the lower bounds follows by taking $j = 0$ in the exp. for $P[Y^{a=i} = j]$ □

According to Pearl

"The instrumental inequality can be used in the detection of undesirable side- effects. Violations of this inequality can be attributed to one of two possibilities: either there is a direct causal effect of the assignment (Z) on the response (Y), unmediated by the treatment (A), or there is a common causal factor influencing both variables. If the assignment is carefully randomized, then the latter possibility is ruled out and any violation of the instrumental inequality (even un- der conditions of imperfect compliance) can safely be attributed to some direct influence of the assignment process on subjects' response (e.g., psychological aversion to being treated). Alternatively, if one can rule out any direct effects of Z on Y , say through effective use of a placebo, then any observed violation of the instrumental inequality can safely be attributed to spurious dependence between Z and Y, namely, to selection bias.

Section 31

Motivation for bounds

Bounds

- Motivation: Can we derive *partial* identification results (i.e. bounds) under weaker assumptions (than those imposed so far)?
- Anyway are bounds useful? I think the answer is yes.
  The following text is from Robins and Greenland:
  - "Some argue against reporting bounds for nonidentifiable parameters, because bounds are often so wide as to be useless for making public health decisions.
  - But we view the latter problem as a reason for reporting bounds in conjunction with other analyses: Wide bounds make clear that the degree to which public health decisions are dependent on merging the data with strong prior beliefs.
  - Even when the ITT[43] null hypothesis of equality of treatment arm-specific means is rejected, the bounds may appropriately include zero. If treatment benefits some subjects and harms others, the ATE parameter may be zero even though both the sharp and ITT null hypotheses are false.

[43] say, the effect of Z in our considerations.

According to Pearl

When conditions for identification are not met, the best one can do is derive bounds for the quantities of interest—namely, a range of possible values that represents our ignorance about the data-generating process and that cannot be improved with increasing sample size.

Bounds on the ATE

$\mathbb{E}[Y^1 - Y^0] = \mathbb{E}[Y^1] - \mathbb{E}[Y^0]$ can be decomposed as

$$\sum_{a=0}^{1} E\left[Y^1 \mid A = a\right] P[A = a] - \sum_{a=0}^{1} E\left[Y^0 \mid A = a\right] P[A = a]. \qquad (10)$$

- $\mathbb{E}[Y^a \mid A = a] = \mathbb{E}[Y \mid A = a]$ by consistency.
- $\mathbb{E}[Y^a \mid A = a]$ and $P[A = a]$ are identifiable and can be consistently estimated by their empirical counterparts.
- the observed data provide no information about $\mathbb{E}[Y^a \mid A = 1 - a]$, such that (10) is only partially identifiable without additional assumptions (such as exchangeability).

Bounds on the ATE

- $\mathbb{E}[Y^1 - Y^0]$ is bounded by smallest and largest possible values for $\mathbb{E}[Y^a \mid A = 1 - a]$.
- If Y^1 and Y^0 are not bounded then bounds on $\mathbb{E}[Y^1 - Y^0]$ will be ranging from $-\infty$ to ∞.
- Informative bounds are only possible if Y^0 and Y^1 are bounded.
- Because any bounded variable can be rescaled to take values in the unit interval, without loss of generality assume $Y^a \in [0, 1]$ for $a = 0, 1$. Then $0 \leq \mathbb{E}[Y^a \mid A = 1 - a] \leq 1$ and from (10) it follows that $\mathbb{E}[Y^1 - Y^0]$ is bounded below by setting $\mathbb{E}[Y^1 \mid A = 0] = 0$ and $\mathbb{E}[Y^0 \mid A = 1] = 1$, which yields the lower bound

$$E\left[Y^1 \mid A = 1\right] P[A = 1] - E\left[Y^0 \mid A = 0\right] P[A = 0] - P[A = 1].$$

Similarly, $\mathbb{E}[Y^1 - Y^0]$ is bounded above by setting $\mathbb{E}[Y^1 \mid A = 0] = 1$ and $\mathbb{E}[Y^0 \mid A = 1] = 0$, which yields the upper bound

$$E\left[Y^1 \mid A = 1\right] P[A = 1] - E\left[Y^0 \mid A = 0\right] P[A = 0] + P[A = 0].$$

Width of bounds

Determining treatment effect bounds can be viewed as a constrained optimization problem. The assumptions we make, for example exchangeabilities, determine the constraints.

- The bounds from the previous slide have width 1 and are contained in $[-1, 1]$, and are called the Manski-Robins bounds.

Motivating example 2: bounds

- We will consider a setting where Z, A, Y are all binary. This could for example be plausible in a randomized controlled trial, where
  - Z is treatment assignment
  - A is the treatment taken
  - Y is the outcome
- In our motivation, we will assume no defiers.
- What do we know about the average treatment effect?
  - We will explore this (and build some intuition) in the next slides.

Motivating example 2 (cont.): always-takers

- Suppose we consider an IV setting with monotonicitiy (no defiers).
  - Then we can simply identify always-takers by $A^{z=0} = 1$.
  - The fraction of always-takers is $P(A = 1 \mid Z = 0)$
  - $\mathbb{E}(Y^{a=1} \mid A = 1, Z = 0) = \mathbb{E}(Y \mid A = 1, Z = 0)$ $= \mathbb{E}(Y \mid A^{z=0} = 1, A^{z=1} = 1)$.
  - $\mathbb{E}(Y^{a=1} - Y^{a=0} \mid A = 1, Z = 0) \leq \mathbb{E}(Y \mid A = 1, Z = 0)$ with equality when all always-takers have $Y^{a=0} = 0$.

Motivating example 2 (cont.): never-takers

- Suppose monotonicitiy (no defiers).
  - Then we can simply identify never-takers by $A^{z=1} = 0$.
  - The fraction of never-takers is $P(A = 0 \mid Z = 1)$
  - $\mathbb{E}(Y^{a=0} \mid A = 0, Z = 1) = \mathbb{E}(Y \mid A = 0, Z = 1) = \mathbb{E}(Y \mid A^{z=0} = 0, A^{z=1} = 0)$.
  - $\mathbb{E}(Y^{a=1} - Y^{a=0} \mid A = 0, Z = 1) \leq 1 - \mathbb{E}(Y^{a=0} \mid A = 0, Z = 1)$ with equality when all never-takers have $Y^{a=1} = 1$.

Suppose no effect in compliers

Combine the simple results from the two previous slides to gain some insight:

- Suppose monotonicitiy (no defiers).
  - Suppose no effect in compliers $\implies \mathbb{E}(Y^{z=1} = Y^{z=0}) = 0$, in other words no intention to treat effect (ITT). Think about it, if it isn't clear!
  - Then the maximum possilbe effect of actually taking treatment is

$$\begin{aligned}&\mathbb{E}(Y^{a=1} - Y^{a=0})\\ \leq&\mathbb{E}(Y^{a=1} \mid A = 1, Z = 0)P(A = 1 \mid Z = 0)\\ &+ [1 - \mathbb{E}(Y^{a=0} \mid A = 0, Z = 1)]P(A = 0 \mid Z = 1),\end{aligned}$$

    even if the intention to treat (ITT) effect $\mathbb{E}(Y^{z=1} = Y^{z=0}) = 0$.
  - Thus, even if the ITT effect is zero, there could be considerable causal effects of taking treatment. In other words, even if the ITT is null, the ATE can be nonzero, which seriously complicate the interpretation of hypothesis tests of the ITT in settings with (a substantial amount of) noncompliance.

Section 32

Finite sample inference

Finite sample inference: Where does randomness come from?

- We have considered superpopulation inference, where the randomness comes from the fact that we have a random draw from a superpopulation.
- However, in a randomised trial, we do not necessarily need to consider a superpopulation at all.
- In the (simple) setting of an experiment, we can often do *finite* sample, or randomization-based, inference.
- Yet, we shall see that to generalize the results outside of the study – which is really what researcher would like to do in most settings – it is necessary to consider large sample extensions (which fundamentally ends up being superpopulations).

Superpopulation inference and finite sample inference

- We have suppose that our study population is sampled at random from an (essentially) infinite superpopulation, sometimes referred to as the target population.
- Broadly speaking, we aimed to generalize our results to this superpopulation.
- It is possible to take a different point of view in randomised trials, often called "design-based inference", which we will study now. This does not require the consideration of a superpopulation at all.[44]

Definition (Design-based inference)

Inference is drawn from a *finite* population, where the potential outcomes of the experimental units are fixed and the randomness comes solely from the treatment assignment.

[44] However, to generalize results from finite samples to settings outside of the experiment – even if we start in the design based setting – it is difficult to proceed without consider a target (super)population. Thus, if we are interested in using the results from the trials for decisions (or rigorous reasoning more broadly) outside of the experiment, it seems that we need to rely on superpopulation inference anyway.

Notation

- We have a sample of n individuals
- As before, for each individual i, let A_i, L_i, Y_i be treatment, baseline covariates and outcomes, respectively.
- We use bold symbols to denote n-vectors:
  - $\boldsymbol{A} = (A_1, \dots, A_n)$,
  - $\boldsymbol{L} = (L_1, \dots, L_n)$,
  - $\boldsymbol{Y} = (Y_1, \dots, Y_n)$.
- We will consider settings where A is randomly assigned. Thus, for a binary A there are 2^n possible values of $\boldsymbol{A}$. Define $\mathbb{A} = \{0, 1\}^n$.
- Let $\mathbb{A}^+$ denote the set of vectors $\boldsymbol{a}$ (i.e. realizations of $\boldsymbol{A}$), with positive probability under a particular experimental design.

Fisher's Randomisation inference

- Key idea: do inference based solely on the *assignment* mechanism.
- The counterfactuals $Y_i^{a=1}, Y_i^{a=0}$ are considered to be *fixed* variables.
- All the *randomness* comes from the random assignment of A.
- Fisher's aim was to test the sharp null hypothesis using the so-called exact test.
- The idea is basically a stochastic proof by contradiction...
- Fisher's null hypothesis is $H_0 : Y_i^{a=1} \equiv Y_i^{a=0}$ for all $i \in \{1, 2, \dots, n\}$. In words, the treatment has no effect of the outcomes in no individual. Under the null hypothesis, $Y_i^{a=1} = Y_i^{a=0} = Y_i$, but of course this is not true under the alternative.
- This null hypothesis is called a **sharp** null hypothesis; it holds for every individual
    - it allows the researcher to fill in a hypothetical value for each unit's missing counterfactual outcome

Fisher's exact test: A test of individual effects

- Define the *sharp null hypothesis* $H_0 : Y_i^{a=1} = Y_i^{a=0}$ for all $i \in \{1, 2, \dots, n\}$.
- Define a test statistic[45], $S \equiv S(\boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{L})$, e.g. $S^{diff} = \frac{1}{n_1} \sum_{i:A_i=1} Y_i - \frac{1}{n_0} \sum_{i:A_i=0} Y_i$.
- Let s^* be an observed test statistic. Then $P(S \geq s^*)$ is a p-value, where the probability is under the law that describes the null hypothesis.
- Fisher suggested an exact test.
  - The idea is to ask the following question: How unusual or extreme is the observed statistic (say, absolute difference), assuming that the null hypothesis is true?
- Intuitively, we want to have *power* against alternative hypotheses, but this is somehow complicated because there are many alternative hypotheses. It seems reasonable to have good power against alternative hypotheses that are substantively most interesting.

[45] A statistic is a known, real-valued function of the data (here, $\boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{L}$)

Examples of statistics

- Averages (like above)
- Trimmed means
- Quantiles (medians)
- T-statistics
- Rank statistics (perhaps good when heavy-tailed distributions)

One example is the Kolmogorov-Smirnov Statistic. Define, the empirical distributions

$$\hat{F}_{a=1}(y) = \frac{1}{n_1} \sum_{i:A_i=1} I(Y_i \leq y) \quad \hat{F}_{a=0}(y) = \frac{1}{n_0} \sum_{i:A_i=1} I(Y_i \leq y).$$

The Kolmogorov-Smirnov Statistic is

$$S^{ks} = \sup_y |\hat{F}_{a=1}(y) - \hat{F}_{a=0}(y)| = \max_i |\hat{F}_{a=1}(Y_i) - \hat{F}_{a=0}(Y_i)|.$$

We can combine statistics

- Fisher's exact p-value inference is valid when there is one test statistic and one null hypothesis.
- However, we can combine test statistics.
  - Consider two statistics S^1 and S^2.
  - The combine $S^{comb} = g(S^1, S^2)$. (e.g. $S^{comb} = \max(S^1, S^2)$)
  - Then we can calculate a p-value

$$P(S^{comb} \leq s^{\star,comb})$$

Illustration of Fisher's exact test

Under the sharp H_0, we can impute missing values of the counterfactuals

i	$Y_i^{a=1}$	$Y_i^{a=0}$	A_i	Y_i
1	-5	-5	1	-5
2	6	6	0	6
3	8	8	1	8
4	0	0	0	0

Table 2: Fisher's idea

The idea is resampling without replacement

Consider the estimator $\frac{1}{n_1}\sum_{i:A_i=1} Y_i - \frac{1}{n_0}\sum_{i:A_i=0} Y_i$. Because we have a completely randomised experiment, the following $\binom{4}{2} = 6$ scenarios are equally possible under H_0,

$$\boldsymbol{A} = (1,1,0,0),\ \hat{\tau} = \frac{-5+6-8-0}{2} = -3.5$$

$$\boldsymbol{A} = (1,0,1,0),\ \hat{\tau} = \frac{-5-6+8-0}{2} = -1.5$$

$$\boldsymbol{A} = (1,0,0,1),\ \hat{\tau} = \frac{-5-6-8+0}{2} = -9.5$$

$$\boldsymbol{A} = (0,1,1,0),\ \hat{\tau} = \frac{5+6+8-0}{2} = 9.5$$

$$\boldsymbol{A} = (0,1,0,1),\ \hat{\tau} = \frac{5+6-8+0}{2} = 1.5$$

$$\boldsymbol{A} = (0,0,1,1),\ \hat{\tau} = \frac{5-6+8+0}{2} = 3.5$$

One way of explaining Fisher's exact test

1. Do the randomization.
2. Calculate a statistic $S = S(\boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{L})$, a function of the observed data.
3. Under the assumption of H_0, i.e. no individual level causal effect, fill in missing potential outcomes.
4. Under the assumption of H_0, generate many hypothetical replications of the randomization, and in each of which calculate a statistic $S_{rep} = S_{rep}(\boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{L})$
5. Compare S with the values S_{rep}

This is an example of a permutation test.

More formally

- Define $H_0 : Y_i^{a=1} = Y_i^{a=0}$.
- Now, consider the randomisation distribution of two statistics S
- Define $\mathcal{F} = (\boldsymbol{Y^0}, \boldsymbol{Y^1})$. In this case, the randomization distributions of $S = S(\boldsymbol{A}, \boldsymbol{Y}, \boldsymbol{L})$ is

$$F(s) = P(S \leq s \mid \mathcal{F})$$

- Then the one-sided p-value of observing the same value or more extreme of the observed statistics S is $F(S)$.
- In our example, the one-sided p-value is $1 - F(-1.5) = 1 - 0.5$.

Fisher’s randomization test formally

Theorem (Nominal coverage of the exact test)

Under consistency and H_0, $P(F(S) \leq \alpha \mid \mathcal{F}) \leq \alpha$ for all $\alpha \in (0,1)$.

Proof.

This follows from some basic properties of the distribution function: indeed, $F^{-1}(\alpha) = \sup\{s : F(s) \leq \alpha\}$. Also $F(s)$ is non-decreasing and right-continuous and therefore

$$P(F(S) \leq \alpha) = P(S < F^{-1}(\alpha)) = \lim_{s \rightarrow F^{-1}(\alpha)} P(S \leq s) \leq \alpha.$$

□

PS: you may have seen the probability integral transform before, i.e. if X is continuous, then $Z = F(X) \sim U(0,1)$

$$P(F(X) \leq \alpha) = P(X \leq F^{-1}(\alpha)) = F(F^{-1}(\alpha)) = \alpha.$$

Remark on the proof and randomness

We only use randomization to compute the p-values because, by definition of the statistic S, whose randomness is due to $\boldsymbol{A}$, we have that

$$F(s) = P(S \leq s \mid \mathcal{F}) = \sum_{\boldsymbol{a} \in \boldsymbol{\mathcal{A}}} P(\boldsymbol{A} = \boldsymbol{a} \mid \mathcal{F}) I(S \leq s)$$

and $P(\boldsymbol{A} = \boldsymbol{a} \mid \mathcal{F}) = P(\boldsymbol{A} = \boldsymbol{a})$ by design, i.e., by the randomisation scheme.

Conservative or good?

Conservative does not necessarily mean appropriate. Consider a confidence interval formed by stating that a random 95% of the time, the interval is any positive or negative number, and that 5% of the time, the interval is the number 0. Such an interval would cover the true value of any quantity of interest at least 95% of the time, and thus would also be a "conservative" interval. It would not, however, be of any use....
Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015

Checking for no causal effect (hypothesis testing)

- Suppose we want to check if there is no causal effect.
- A classical frequentist approach goes as follows
  - Assume no effect (the null hypothesis).
  - Calculate a statistic,[46] and see how surprising the statistics is, under the assumption of no effect.
  - If it is very surprising, we reject.
- This is contrapositive logic, applied to probabilities.

[46] A statistic is a known, real-valued function of the data

We should be careful with this (Example from Shpitser)

Suppose we do cancer screening.

- Consider a rare cancer, our outcome Y, such that $P(Y = 1) = 0.0001$
- Consider also a cancer lab test T. And suppose
  - Test false positive $P(T = 1 \mid Y = 0) = 0.01$.
  - Test false negative $P(T = 0 \mid Y = 1) = 0.001$.
- Suppose we had a positive test, $T = 1$. Should we worry?
- Just use Bayes theorem,

$$P(Y = 1 \mid T = 1) = \frac{P(T = 1 \mid Y = 1)P(Y = 1)}{P(T = 1)} \approx 0.01.$$

- What would the Frequentist do? Assume $Y = 0$, and check how surprised we would be, that is, calculate $P(T = 1 \mid Y = 0) = 0.01$, which is surprising....
- Lesson learned, if hypothesis probabilities are uneven, hypothesis testing might not be ideal..