36-705: Intermediate Statistics **Fall 2017**

Lecture 25: November 1

Lecturer: Siva Balakrishnan

We will wrap up our discussion of confidence intervals and then begin discussing the basics of causal inference.

25.1 Pivots

This is something that is known as pivotal inference in statistics. This idea is really motivated by our first example where we constructed two different confidence intervals for the uniform parameter θ. One of them was well-behaved (i.e. we could easily get it to have the right coverage probability) and the other one failed (i.e. we needed to set the length of the interval to depend on the unknown parameter θ).

A function $Q(X_1, \ldots, X_n, \theta)$ is a *pivot* if the distribution of Q does not depend on θ. For example, if $X_1, \ldots, X_n \sim N(\theta, 1)$ then

$$\overline{X}_n - \theta \sim N(0, 1/n)$$

so $Q = \overline{X}_n - \theta$ is a pivot.

Let a and b be such that

$$P_\theta(a \leq Q(X, \theta) \leq b) \geq 1 - \alpha$$

for all θ. We can find such an a and b because Q is a pivot. It follows immediately that

$$C(X) = \{\theta : \; a \leq Q(X, \theta) \leq b\}$$

has coverage $1 - \alpha$.

Example 25.1 *Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. ($\sigma$ known.) Then*

$$Z = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \sim N(0, 1).$$

We know that

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$

and so

$$P\left(-z_{\alpha/2} \leq \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \leq z_{\alpha/2}\right) = 1 - \alpha.$$

Thus

$$C = \overline{X} \pm \frac{\sigma}{\sqrt{n}} z_{\alpha/2}.$$

If σ is unknown, then this becomes

$$C = \overline{X} \pm \frac{S}{\sqrt{n}} t_{n-1,\alpha/2}$$

because

$$T = \frac{\sqrt{n}(\overline{X} - \mu)}{S} \sim t_{n-1}.$$

We can similarly re-interpret the uniform interval we created in terms of a pivot.

Example 25.2 *Let $X_1, \ldots, X_n \sim \text{Uniform}(0, \theta)$. Let $Q = X_{(n)}/\theta$. Then*

$$\mathbb{P}(Q \leq t) = \prod_i \mathbb{P}(X_i \leq t\theta) = t^n$$

so Q is a pivot. Let $c = \alpha^{1/n}$. Then

$$\mathbb{P}(Q \leq c) = \alpha.$$

Also, $\mathbb{P}(Q \leq 1) = 1$. Therefore,

$$\begin{aligned} 1 - \alpha &= \mathbb{P}(c \leq Q \leq 1) = \mathbb{P}\left(c \leq \frac{X_{(n)}}{\theta} \leq 1\right) \\ &= \mathbb{P}\left(\frac{1}{c} \geq \frac{\theta}{X_{(n)}} \geq 1\right) \\ &= \mathbb{P}\left(X_{(n)} \leq \theta \leq \frac{X_{(n)}}{c}\right) \end{aligned}$$

so a $1 - \alpha$ confidence interval is

$$\left(X_{(n)},\ \frac{X_{(n)}}{\alpha^{1/n}}\right).$$

25.2 Large Sample Intervals

The Wald Interval. We know that, under regularity conditions,

$$\frac{\widehat{\theta}_n - \theta}{\text{se}} \rightsquigarrow N(0,1)$$

where $\widehat{\theta}_n$ is the mle and $\text{se} = 1/\sqrt{I_n(\widehat{\theta})}$. So this is an asymptotic pivot and an approximate confidence interval is

$$\widehat{\theta}_n \pm z_{\alpha/2}\text{se}.$$

By the delta method, a confidence interval for $\tau(\theta)$ is

$$\tau(\widehat{\theta}_n) \pm z_{\alpha/2}\text{se}(\widehat{\theta})|\tau'(\widehat{\theta}_n)|.$$

The Likelihood-Based Confidence Set. We saw an example before of trying to exactly the invert the LRT. This is often difficult to do. Let's consider inverting the asymptotic LRT. We test

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0.$$

Let k be the dimension of θ. We don't reject if

$$-2\log\left(\frac{L(\theta_0)}{L(\widehat{\theta})}\right) \leq \chi^2_{k,\alpha}$$

that is, if

$$\frac{L(\theta_0)}{L(\widehat{\theta})} > e^{-\chi^2_{k,\alpha}/2}.$$

So, the set of non-rejected nulls is

$$C_n = \left\{\theta : \frac{L(\theta)}{L(\widehat{\theta})} > e^{-\frac{\chi^2_{k,\alpha}}{2}}\right\}.$$

This confidence set has a somewhat pleasing interpretation that it is just a collection of parameters that have high likelihood. Then

$$P_\theta(\theta \in C) \to 1 - \alpha$$

for each θ.

Example 25.3 *Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$, let $S = \sum X_i$. Using the Wald statistic*

$$\frac{\widehat{p} - p}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}} \rightsquigarrow N(0,1)$$

so an approximate confidence interval is

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

Using the LRT we get

$$C = \left\{ p : \ -2\log\left(\frac{p^S(1-p)^{n-S}}{\widehat{p}^S(1-\widehat{p})^{n-S}} \right) \leq \chi^2_{1,\alpha} \right\}.$$

These intervals are different but, for large n, they are nearly the same.

25.3 Multiple Confidence Intervals

When we have multiple parameters we might want to construct intervals $C_1, \ldots, C_d$ such that we control:

$$\mathbb{P}(\exists\ j \text{ such that } \theta_j \notin C_j) \leq \alpha.$$

In this case, we could just Bonferroni correct our confidence intervals, i.e. we could construct confidence intervals which have coverage probability α/d, and these would then have the desired property. This is analogous to controlling the FWER.

It is worth pondering what the analog of FDR control might be for confidence intervals (maybe look up False Coverage-statement Rate).

25.4 Tests Versus Confidence Intervals

Confidence intervals are more informative than tests. Intuitively, p-values are more informative than an accept/reject decision because it summarizes all the significance levels for which we would reject the null hypothesis. Similarly, a confidence interval is more informative that a test because it summarizes all the parameters for which we would (fail to) reject the null hypothesis. More practically, a confidence interval tells us something about the "effect size" as well as something about the uncertainty in our estimate of the "effect size".

Look at Figure 25.1. Suppose we are testing $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. We see 5 different confidence intervals. The first two cases (top two) correspond to not rejecting H_0. The other three correspond to rejecting H_0. Reporting the confidence intervals is much more informative than simply reporting "reject" or "don't reject."

25.5 Causal Inference

A lot of statistics focusses on questions of association. Are X and Y correlated? Is X predictive of Y, and so on.

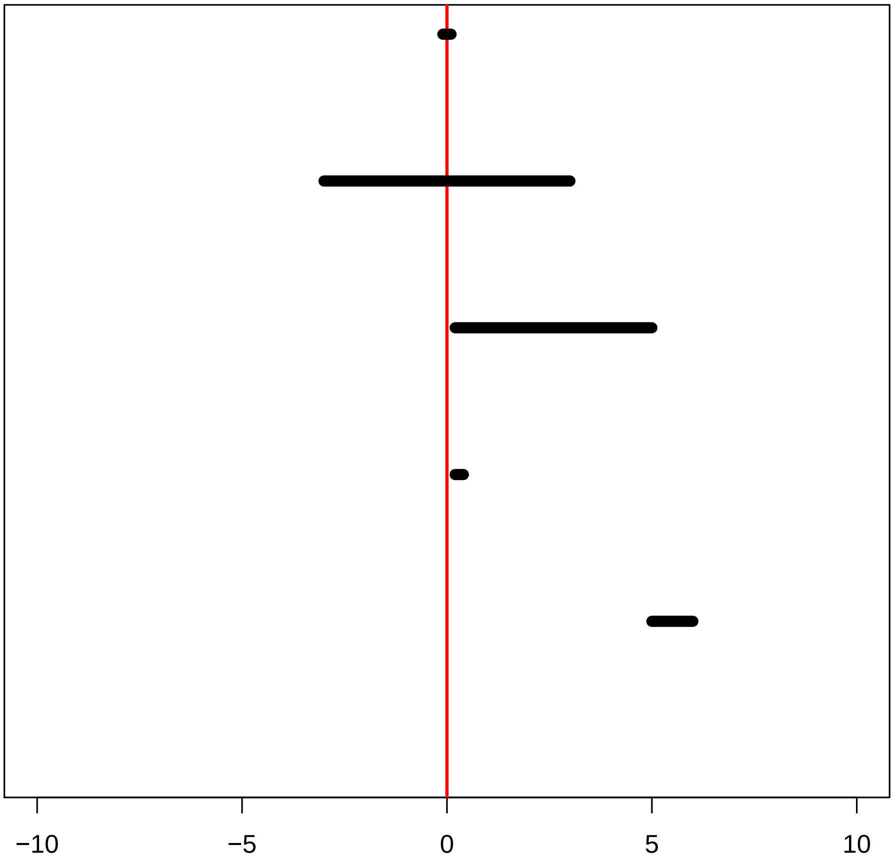

Figure 25.1: Five examples: 1. Not significant, precise. 2. Not significant, imprecise. 3. Barely significant, imprecise. 4. Barely significant, precise. 5. Significant and precise.

In many applications however, our questions are inherently causal: in medicine we wish to know if a new drug is effective against a disease. This is not a question of association, because if I went out in the world and measured all the people taking aspirin, most likely many of them would have headaches so I could (correctly?) conclude aspirin and headaches are associated. It is almost certainly not the case that aspirin causes headaches and this is what we usually mean by the phrase: "correlation does not imply causation."

It will take a bit of work to get to the questions of interest but broadly you should think of the two statistical questions in causal inference as analogous to ones we have considered so far: we want to estimate the causal effect (point estimation) and construct confidence intervals for the causal effect (inference/hypothesis testing).

25.6 The Potential Outcomes Framework

The basic language of causal inference that we will adopt comes from the work of Neyman (and later Rubin). Causality is tied to something known as a manipulation/intervention applied to a *unit* (think person).

We will think of the case when there are two possible actions (or treatments). Think of taking an aspirin and not taking an aspirin as the two treatments. Often we refer to one of the treatments as the active treatment (or just treatment) and the other as the control treatment (or just control).

We associate every unit and the two treatments with two *potential outcomes*: the potential outcome if the unit received the treatment and the potential outcome if the unit received control. A priori both potential outcomes are possible. However, every unit only receives one of the two treatments (i.e. either treatment or control) and so we only observe one of the two potential outcomes. This is known as the fundamental problem of causal inference. We only observe one of the potential outcomes for each unit.

While all of this might seem rather obvious, thinking formally about treatment and control, and the potential outcomes is extremely important to causal inference. A point of particular emphasis is that if you are asking a causal question, ideally you need to be able to meaningfully say what the "treatment" is and what the potential outcomes are.

Here are a few examples of statements:

1. "Aspirin cures headaches." In order to cast this is the potential outcomes framework we could imagine that for a person with a headache (a unit) we could either give the person aspirin (treatment) or a placebo (control), and observe the corresponding potential outcome.

2. "She has long hair because she is a girl." This sounds like a causal statement so we

should be able to describe the experiment. Is a unit a girl/boy? What exactly is a treatment? Can we meaningfully say what the potential outcomes are?

For some causal questions we can naturally define an associated "experiment". Murky causal questions are ubiquitous, and are in some sense interesting and challenging. For instance, I might like to know the effect of race on life expectancy. If you go through the exercise above again you will have a lot of trouble. Research in social science, political science, epidemiology and economics (to name a few fields) are centered around how to make sense of these difficult questions. I particularly like the terminology of Angrist and Pischke (Mostly Harmless Econometrics) who term questions for which one cannot design an experiment as a "Fundamentally Unidentified Question (FUQ)" – the book goes on to describe some casual sounding questions which are FUQed.

We will focus on simpler case, where there are well-defined interventions and potential outcomes.

In this case, for the i^{th} unit we will denote the potential outcome if the unit receives control as $Y_i(0)$ and the potential outcome if the unit receives treatment as $Y_i(1)$. A natural definition of the *causal effect* of treatment on the i^{th} unit is $Y_i(1) - Y_i(0)$ (you could consider any other meaningful function of the potential outcomes and we will discuss this soon). Again, the fundamental problem of causal inference is that we only observe $Y_i(1)$ or $Y_i(0)$ and not both.

25.7 Multiple Units

Defining the causal effect does not require multiple units, however, estimating causal effects does. The idea is simple, suppose I observe the potential outcomes under treatment for some units and the potential outcomes under control for some units, then maybe in some cases I can put these together to get a sense of the average causal effect.

This actually requires another assumption. This is called the Stable-Unit-Treatment-Value-Assumption (SUTVA). The assumption has two parts:

1. Giving treatment/control to one unit does not affect the potential outcomes of other units,

2. For units receiving treatment (or control) there is only one level of treatment (it cannot be that some units take one aspirin, some take two and a few take 10000).

25.8 The assignment mechanism

The next part of the story, is what is called the assignment mechanism. Suppose we have n units $\{1, \ldots, n\}$. The assignment mechanism, is what determines which potential outcome we observe for each unit.

We will denote the assignment vector: $W \in \{0,1\}^n$ where $W_i = 0$ means the unit i is assigned to control and $W_i = 1$ means that unit i is assigned to treatment.

The treatment mechanism that we will focus on today is what is known as a completely randomized trial. This means that we select a number m of units to receive treatment, and then select m out of the n units uniformly at random. In mathematical notation, this means that:

$$\mathbb{P}(W = w) = \frac{1}{\binom{n}{m}},$$

for any binary vector w, with $\sum_i w_i = m$.

An alternative that is popular is something called a Bernoulli trial, where each individual has some fixed probability p (think 0.5) of receiving treatment. Alternatively, you could imagine *stratified* randomized assignments where the units are grouped and then randomly assigned treatment/control within the group. With this notation we can now write the observed and missing potential outcomes:

$$\begin{aligned} Y_i^{\text{obs}} &= Y_i(W_i) \\ Y_i^{\text{mis}} &= Y_i(1 - W_i). \end{aligned}$$

More generally, you could imagine situations where a doctor measures some covariates of the patient (her blood pressure, age, and height say) and then decides whether to recommend the treatment or not. In this case, the assignment is not random, and we will discuss situations like this in a future lecture.

25.9 Causal Estimands

Finally, let us be a bit more precise about what we'd like to estimate. There are many things we might care about estimating:

1. Unit level causal effects: things like $Y_i(1) - Y_i(0)$ or $Y_i(1)/Y_i(0)$.
2. The average treatment effect:

$$\tau = \frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0)).$$

This is what we will focus on in this class.

3. Average treatment effect over sub-populations:

$$\tau_S = \frac{1}{|S|}\sum_{i=1}^{n}(Y_i(1) - Y_i(0))\mathbb{I}(i \in S).$$

For instance the set S could be all men in the population (i.e. I am interested in whether aspiring relieves headaches in men).

25.10 Finite sample versus population causal inference

You might notice that we defined our setup in terms of having n units, and our estimands are things that depend on the n units.

It is perhaps more common to suppose that units are sampled i.i.d from a (super)population, in which case the average treatment effect for instance would be defined as:

$$\tau = \mathbb{E}(Y(1) - Y(0)).$$

25.11 The basic issue in causal inference

A natural quantity that we can estimate is:

$$\alpha = \mathbb{E}[Y(1)|W = 1] - \mathbb{E}[Y(0)|W = 0].$$

Suppose that m individuals receive treatment. A natural way to try to estimate α is to use:

$$\widehat{\alpha} = \frac{1}{m}\sum_{i \in T} Y_i(1) - \frac{1}{n-m}\sum_{i \notin T} Y_i(0).$$

The most important point is that in general:

$$\tau \neq \alpha.$$

There are broadly two ways to try to fix this issue: (1) to randomly assign treatment, i.e. conduct a randomized trial (2) to adjust for confounders. We will spend today on (1) and the next lecture on (2).

If we randomly assign treatment then we have that,

$$W \perp\!\!\!\perp (Y(1), Y(0)),$$

and in this case,

$$\alpha = \tau.$$

25.12 Estimating the average treatment effect in a trial

Under the assumption of a completely randomized trial (and SUTVA), it is easy to construct an estimator of the average treatment effect. Let us denote the set of treated units as T and the number of treated units as m (remember that m is fixed). Our estimate is:

$$
\begin{aligned}
\widehat{\tau} &= \frac{1}{m}\sum_{i\in T} Y_i(1) - \frac{1}{n-m}\sum_{i\notin T} Y_i(0) \\
&= \sum_{i=1}^{n}\left(\frac{W_i}{m}Y_i(1) - \frac{(1-W_i)}{n-m}Y_i(0)\right).
\end{aligned}
$$

Now, we can see that this is an unbiased estimator of the average treatment effect. It is worth noting that the only thing that is surely random here is our treatment assignment, the potential outcomes can be fixed or random (it does not matter which).

$$
\mathbb{E}[\widehat{\tau}] = \sum_{i=1}^{n} \frac{\mathbb{E}\left(W_i\right)}{m}Y_i(1) - \frac{\mathbb{E}\left((1-W_i)\right)}{n-m}Y_i(0).
$$

The mean of W_i is given by:

$$
\mathbb{E}(W_i) = \frac{\binom{n-1}{m-1}}{\binom{n}{m}} = \frac{m}{n},
$$

and

$$
\mathbb{E}(1-W_i) = \frac{n-m}{n}.
$$

This gives us that:

$$
\mathbb{E}[\widehat{\tau}] = \tau.
$$

The next important question is how to we construct valid confidence intervals for the causal effect. It is a somewhat deep question because natural strategies (compute the variance of the estimator) will fail. We won't go into details here but instead we will discuss approaches that work.

25.13 Hypothesis testing: Fisher's Exact p-values

Fisher was one of the first statisticians to understand the power of a randomized trial. In agricultural experiments, he advocated randomized experiments in order to draw rigorous causal conclusions.

A natural subsequent problem is: given an estimate of the causal effect, assess its significance (or construct confidence intervals for it).

Fisher gave a way to construct valid p-values under what is called the *sharp null*, i.e. the null hypothesis that for every unit i the potential outcomes are the same under the treatment and control, i.e. the treatment has no effect. The method is reminiscent of the permutation method we used for two-sample testing.

Suppose for simplicity that we are using the estimator described in the previous section and we reject the null hypothesis if $|\widehat{\tau}|$ is large. Under the null hypothesis, we can determine both potential outcomes $Y_i(0)$ and $Y_i(1)$ for all the units.

We can now use the permutation method, suppose a different set T' of m units were to receive treatment: then our estimate would be:

$$\widehat{\tau}_{T'} = \frac{1}{m}\sum_{i \in T'} Y_i(1) - \frac{1}{n-m}\sum_{i \notin T'} Y_i(0),$$

where we can use the sharp null hypothesis to "fill in" the potential outcomes we do not observe. We can repeat this many times (say B) and compute the p-value:

$$\text{p-value} = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(|\widehat{\tau}_{T_b}| \geq |\widehat{\tau}|).$$

It is easy to verify that this is a valid p-value.

The intuition is identical to the permutation test, if there was in fact a different in outcomes under treatment and controls (say treatment potential outcomes were much higher than control potential outcomes) then we would expect the p-value to be small, since the difference in means will get smaller when we randomly swap some of the treatment and control outcomes.

Of course, the sharp null is a very strong null hypothesis, and we often have much weaker null hypotheses. Like perhaps our null hypothesis is just that $\tau = 0$. We will discuss these in a future lecture.