Review D: Hypothesis Testing

Professor Ji-Woong Chung
Korea University

This lecture note is based on Thomas Wiemann's.

Recap

In Part C of the statistics review, we discussed estimation:

- Developed estimators via the sample analogue principle;
- Characterized estimators with finite and large sample properties.

Our analysis highlighted that an estimator $\hat{\theta}_n$ is a random variable and may thus differ from the true (fixed) parameter θ.

In Part D, we consider the question of whether the true parameter is equal to a particular value or within a particular set.

- For example, when interested in the expected returns to education:

$$\tau_{ATT} = E\left[Y_i(1) - Y_i(0) \mid D = 1\right],$$

  we may be particularly curious about whether $\tau_{ATT} > 0$.

The formal analysis of such questions is known as hypothesis testing.

Outline

Outline

Outline

Hypothesis Testing

Our analysis begins with defining a hypothesis to be tested.

Let θ denote the parameter of interest and Θ its possible values.

Consider a partition of Θ into two disjoint subsets Θ_0 and Θ_1 and that we wish to test

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1.$$

Some terminology:

- H_0 is referred to as the null hypothesis;
- H_1 is referred to as the alternative hypothesis;
- When $\Theta_0 = \{\theta_0\}$ is a single element, H_0 is a simple hypothesis;
- When Θ_0 is a non-singleton set, H_0 is a composite hypothesis.

Hypothesis Testing (Contd.)

Example Let Y denote hourly wages and D denote being a college graduate. Do college graduates earn upwards of $600 a week?

To formulate a corresponding hypothesis, let $\mu_{Y|1} \equiv E[Y \mid D = 1]$. Then

$$H_0 : \mu_{Y|1} \geq 600 \quad \text{versus} \quad H_1 : \mu_{Y|1} < 600.$$

Here H_0 is a composite hypothesis.

If we had instead asked, "Do college graduates earn $600 a week?", the corresponding hypothesis would be

$$H_0 : \mu_{Y|1} = 600 \quad \text{versus} \quad H_1 : \mu_{Y|1} \neq 600.$$

Here H_0 is a simple hypothesis.

Hypothesis Testing (Contd.)

Hypotheses pose economic questions in terms of statistical parameters.

- Now we need a procedure to answer these questions.

For this purpose, define a test statistic T_n, which denotes a known function of the sample $X_1, \ldots, X_n$.

- $T_n(X_1, \ldots, X_n)$ is a function of random variables and hence random.

Hypothesis testing finds an appropriate region $\mathcal{R} \subset \text{supp } T_n$ such that

$$T_n \in \mathcal{R} \implies \text{reject } H_0, \quad T_n \notin \mathcal{R} \implies \text{don't reject } H_0.$$

$\mathcal{R}$ is known as the rejection region. We exclusively consider $\mathcal{R}$ of the form

$$\mathcal{R}(c) = \{t \in \mathbb{R} \mid t > c\},$$

for a critical value $c \in \mathbb{R}$. Note: "large" T_n is evidence against H_0.

Type I and Type II Errors

Because T_n is random, we are bound to make errors at some point.

Table: Outcomes of Hypothesis Testing

	Don't Reject H_0	Reject H_0
H_0 true	correct	type I error
H_0 false	type II error	correct

We will need to trade off type I and type II errors in our analysis.

- The less likely we make type I errors, the more likely are type II errors (and vice versa).
- We often focus on controlling the probability of a type I error.

Why? Wasserman (2003) has a nice analogy: *"Hypothesis testing is like a legal trial. We assume someone is innocent unless the evidence strongly suggests that they are guilty. Similarly, we don't reject H_0 unless there is strong evidence against H_0."*

Type I and Type II Errors (Contd.)

A test is characterized by its type I and type II error probabilities.

Definition (Size and Power)

The size of a test is the (maximum) probability of committing a Type I error, $\alpha \in (0, 1)$ such that

$$\begin{aligned}\alpha &= P(T_n \in \mathcal{R}(c_\alpha) \mid H_0 \text{ is true}) = P(T_n > c_\alpha \mid H_0 \text{ is true}) \\ &= P(\text{reject } H_0 \mid H_0 \text{ is true}) = P(\text{type I error}).\end{aligned}$$

The power of a test is the probability of rejecting the null hypothesis when the null hypothesis is false, $1 - \beta$, where

$$\begin{aligned}\beta &= P(T_n \notin \mathcal{R}(c_\alpha) \mid H_0 \text{ is false}) = P(T_n \leq c_\alpha \mid H_0 \text{ is false}) \\ &= P(\text{don't reject } H_0 \mid H_0 \text{ is false}) = P(\text{type II error})\end{aligned}$$

In practice, we choose a critical value c_α such that our test has the desired size. This controls the probability of a type I error.

Type I and Type II Errors (Contd.)

In practice, economists often consider a size of $\alpha = 0.05$ appropriate.

- This is rather arbitrary: Is $1/20$ rare enough?
- Practitioners may disagree on the size they would like to consider.

The next definition allows for side-stepping the issue of pre-specified sizes.

Definition (p-Value)

The p-value of a test is defined as

$$\inf\{\alpha \in (0,1) \mid T_n \in \mathcal{R}(c_\alpha)\},$$

that is, the smallest size of the test such that H_0 would be rejected.

Small p-values are interpreted as evidence against H_0:

- The smaller the p-value, the stronger the evidence against H_0.

Importantly: Large p-values are not evidence in favor of H_0!

- Large p-values may also occur because our test has low power.

Outline

Two-Sided Hypothesis Testing

Let's make things more concrete: Consider a sample $X_1, \dots, X_n \overset{iid}{\sim} X$.

Suppose we are interested in a parameter $\theta \in \mathbb{R}$ (e.g., $\theta = E[X]$), and that we developed an estimator $\hat{\theta}_n$ such that

$$\frac{\hat{\theta}_n - \theta}{\text{se}(\hat{\theta}_n)} \xrightarrow{d} N(0,1).$$

Is θ equal to a particular value, say, θ_0?

For this purpose, we consider testing

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

We are now in need of an appropriate test statistic T_n and a corresponding critical value c_α such that the size of our test is $\alpha \in (0,1)$.

Two-Sided Hypothesis Testing (Contd.)

Given the standard normal limit of the previous slide, a natural choice of test statistic is

$$T_n = \left| \frac{\hat{\theta}_n - \theta_0}{\mathrm{se}(\hat{\theta}_n)} \right|.$$

- Recall that we reject H_0 if T_n is "large".
- Here, T_n increases in deviations of $\hat{\theta}_n$ from θ_0: Seems sensible!

The following theorem shows that T_n is indeed a useful test statistic:

Theorem

Let $\hat{\theta}_n$ be an estimator for θ such that the previous slide's limit holds. Then for T_n defined above, it holds that

$$P(T_n > z_{1-\alpha/2} \mid H_0 \text{ is true}) \to \alpha,$$

where $z_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$ is the $1-\alpha/2$ quantile of a standard normal.

Two-Sided Hypothesis Testing (Contd.)

Proof.

$$
\begin{aligned}
P(T_n > c \mid H_0) &= P\left(\left|\frac{\hat{\theta}_n - \theta_0}{se(\hat{\theta}_n)}\right| > c \mid H_0\right) \\
&= P\left(\frac{\hat{\theta}_n - \theta_0}{se(\hat{\theta}_n)} > c \mid H_0\right) + P\left(\frac{\hat{\theta}_n - \theta_0}{se(\hat{\theta}_n)} < -c \mid H_0\right) \\
&= 1 - P\left(\frac{\hat{\theta}_n - \theta_0}{se(\hat{\theta}_n)} \leq c \mid H_0\right) + P\left(\frac{\hat{\theta}_n - \theta_0}{se(\hat{\theta}_n)} < -c \mid H_0\right) \\
&\to 1 - \Phi(c) + \Phi(-c) \quad \because \frac{\hat{\theta}_n - \theta_0}{se(\hat{\theta}_n)} \xrightarrow{d} N(0,1) \\
&= 1 - \Phi(c) + (1 - \Phi(c)) = 2(1 - \Phi(c))
\end{aligned}
$$

When $c = z_{1-\alpha/2}$, then
$2(1 - \Phi(c)) = 2(1 - \Phi(z_{1-\alpha/2})) = 2(1 - (1 - \alpha/2)) = \alpha$
Note: It's worth memorizing that when $\alpha = 0.05$, we have $z_{1-\alpha/2} \approx 1.96$.

Two-Sided Hypothesis Testing (Contd.)

Example Consider the test statistic T_n defined in the previous slide. By the theorem, we reject $H_0 : \theta = \theta_0$ at significance level α when

$$T_n > z_{1-\alpha/2}.$$

Hence, the p-value is given by

$$\begin{aligned}
&\Rightarrow \Phi(T_n) > \Phi(z_{1-\alpha/2}) = 1 - \frac{\alpha}{2} \\
&\Rightarrow \alpha > 2(1 - \Phi(T_n)) \\
&\Rightarrow 2(1 - \Phi(T_n)) = \text{p-value}
\end{aligned}$$

Outline

One-Sided Hypothesis Testing

Instead of the simple hypothesis considered before, suppose we test

$$H_0 : \theta \leq \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0,$$

or

$$H_0 : \theta \geq \theta_0 \quad \text{versus} \quad H_1 : \theta < \theta_0.$$

Recall that we want large T_n to be evidence against H_0.

- For $H_0 : \theta \leq \theta_0$, choose

$$T_n = \frac{\hat{\theta}_n - \theta_0}{\text{se}(\hat{\theta}_n)}.$$

- For $H_0 : \theta \geq \theta_0$, choose

$$T_n = -\frac{\hat{\theta}_n - \theta_0}{\text{se}(\hat{\theta}_n)}.$$

One-Sided Hypothesis Testing (Contd.)

The next result shows that these are indeed useful test statistics:

Theorem

Let $\hat{\theta}_n$ be an estimator for θ such that the previous slide's limit holds. Then for T_n defined above, it holds that

$$P(T_n > z_{1-\alpha} \mid H_0 \text{ is true}) \to \alpha,$$

where $z_{1-\alpha} = \Phi^{-1}(1-\alpha)$ is the $1-\alpha$ quantile of a standard normal.

An analogous result holds for T_n defined for the opposite hypothesis.

Proof.

$$P\left(\frac{\hat{\theta}_n - \theta_0}{\mathrm{se}(\hat{\theta}_n)} > c \mid H_0\right) = 1 - P\left(\frac{\hat{\theta}_n - \theta_0}{\mathrm{se}(\hat{\theta}_n)} \leq c \mid H_0\right) \to 1 - \Phi(c)$$

Taking $c = z_{1-\alpha}$ implies $1 - \Phi(z_{1-\alpha}) = 1 - (1-\alpha) = \alpha$

Note: It's worth memorizing that when $\alpha = 0.05$, we have $z_{1-\alpha} \approx 1.64$.

One-Sided Hypothesis Testing (Contd.)

Example Consider the test statistic $T_n = \frac{\hat{\theta}_n - \theta_0}{\mathrm{se}(\hat{\theta}_n)}$. By the previous theorem, we reject $H_0 : \theta = \theta_0$ at significance level α when

$$T_n > z_{1-\alpha}.$$

Hence, the p-value is given by

$$\begin{aligned} &\Rightarrow \Phi(T_n) > \Phi(z_{1-\alpha}) = 1 - \alpha \\ &\Rightarrow \alpha > 1 - \Phi(T_n) \\ &\Rightarrow 1 - \Phi(T_n) = \text{p-value} \end{aligned}$$

Outline

Hypothesis Testing and Confidence Intervals

Consider the following thought experiment: Suppose you test

$$H_0 : \theta = \tilde{\theta}_0 \quad \text{versus} \quad H_1 : \theta \neq \tilde{\theta}_0,$$

for all possible values $\tilde{\theta}_0 \in \Theta$ using a test of size α.

- Whenever H_0 is not rejected, you write down the value of $\tilde{\theta}_0$.
- This gives the set (say, C_n) of $\tilde{\theta}_0$ for which H_0 would not be rejected.
- C_n summarizes the collection of hypotheses we would not reject.

It turns out that this newly constructed set C_n is the confidence interval discussed in Part C of the review!

- This is known as the duality between hypothesis testing and confidence intervals.

This implies that we can use a $1 - \alpha$ confidence interval to test hypotheses at a significance level α.

- Step 1: Construct the $1 - \alpha$ confidence interval C_n;
- Step 2: Check whether $\theta_0 \in C_n$. If not, reject $H_0 : \theta = \theta_0$.

Hypothesis Testing and Confidence Intervals (Contd.)

To see this dual relationship, recall that we would include $\tilde{\theta}_0$ in the set C_n if our test of size α does not reject $H_0 : \theta = \tilde{\theta}_0$. That is, whenever

$$T_n \leq c_\alpha.$$

Take $T_n = \left|\frac{\hat{\theta}_n - \theta_0}{\text{se}(\hat{\theta}_n)}\right|$ so that $c_\alpha = z_{1-\alpha/2}$. Then

$$\begin{aligned}\left|\frac{\hat{\theta}_n - \theta_0}{\text{se}(\hat{\theta}_n)}\right| \leq z_{1-\alpha/2} &\Rightarrow -z_{1-\alpha/2} \leq \frac{\hat{\theta}_n - \theta_0}{\text{se}(\hat{\theta}_n)} \leq z_{1-\alpha/2} \\ &\Rightarrow \hat{\theta}_n - z_{1-\alpha/2} \cdot \text{se}(\hat{\theta}_n) \leq \theta_0 \leq \hat{\theta}_n + z_{1-\alpha/2} \cdot \text{se}(\hat{\theta}_n)\end{aligned}$$

Hence, the set of $\tilde{\theta}_0$ for which we don't reject H_0 at significance level α is

$$C_n = \left[\hat{\theta}_n - z_{1-\alpha/2} \cdot \text{se}(\hat{\theta}_n), \hat{\theta}_n + z_{1-\alpha/2} \cdot \text{se}(\hat{\theta}_n)\right].$$

which is identical to our definition of the symmetric confidence interval.

Summary

This conclude our statistics review:

- Discussed the construction of estimators;
- Introduced tools to study the properties of estimators;
- Developed procedures for testing hypotheses about parameters.

Now we're fully equipped to delve into the analysis of causal questions!

- Can leverage our probability expertise for defining and identifying target parameter.
- Can leverage our statistics expertise for estimating the estimand.