BUSS975 Causal Inference in Financial Research

Ji-Woong Chung
chung_jiwoong@korea.ac.kr
Korea University Business School

Chapter 5

Hypothesis Testing

5.1 Introduction and Recap

In the previous chapter, we discussed the problem of *estimation*: constructing estimators for unknown parameters and characterizing their properties in finite samples and asymptotically. We emphasized that an estimator $\hat{\theta}_n$ is a random variable that may differ from the true (fixed but unknown) parameter θ. In this chapter, we turn to the formal analysis of questions concerning whether a parameter equals or differs from some particular value (or falls in a certain range). This leads us to the framework of **hypothesis testing**.

For example, consider the causal parameter

$$\tau_{ATT} \;=\; E[Y_i(1) - Y_i(0) \mid D_i = 1]\,,$$

the average treatment effect on the treated (such as the expected return to a college education for those who attended college). We might be specifically interested in whether $\tau_{ATT} > 0$, i.e. whether the expected return to education is positive. Hypothesis testing provides a systematic way to answer such questions with statistical rigor.

5.2 Formulating Statistical Hypotheses

A **statistical hypothesis** is a claim or assertion about a population parameter (or the distribution of a random variable). To conduct a test, we begin by formalizing the question of interest as two competing hypotheses:

$$H_0 : \theta \in \Theta_0 \qquad \text{versus} \qquad H_1 : \theta \in \Theta_1,$$

where θ is the parameter of interest, Θ is the set of all possible values for θ, and Θ_0 and Θ_1 are disjoint subsets of Θ that partition it (so $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \emptyset$).

- H_0 is called the **null hypothesis**. It represents the baseline claim or status quo that we seek evidence against.
- H_1 is called the **alternative hypothesis**. It represents the competing claim we will accept if the null is rejected.
- If $\Theta_0 = \{\theta_0\}$ is a single point (i.e. we are testing that θ equals a specific value), then H_0 is a **simple hypothesis**.
- If Θ_0 contains multiple values (a range or composite condition), then H_0 is a **composite hypothesis**.

Example 5.1. (Simple vs. Composite Hypotheses) Let Y be hourly wages and D indicate being a college graduate. Suppose we want to test whether college graduates earn at least \$600 per week on average. Define $\mu_{Y|1} = E[Y \mid D = 1]$ as the mean weekly wage of college graduates. We can formulate:

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

Here the null allows any value $\mu_{Y|1}$ greater or equal to 600, so H_0 is composite. The alternative asserts the mean is less than 600.

If instead we ask, “Do college graduates earn \$600 per week on average?” the hypotheses would be:

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

testing equality against any difference. In this case H_0 is simple ($\Theta_0 = \{600\}$).

In both cases, we have translated an economic or substantive question into a hypothesis about a statistical parameter. The next step is to design a procedure for deciding between H_0 and H_1 using sample data.

5.3 Test Statistics and Decision Rules

To test a hypothesis, we need a rule that maps the observed sample to a decision: either "do not reject H_0" (i.e. continue to regard H_0 as plausible) or "reject H_0" (and conclude H_1 instead). We construct a **test statistic** $T_n = T_n(X_1, \ldots, X_n)$, which is a known function of the sample. Being a function of random data, T_n is itself a random variable with some probability distribution. The test statistic is chosen so that extreme values of T_n provide evidence against H_0.

Once we have a test statistic, we select a **rejection region** $\mathcal{R}$, a subset of possible values of T_n. The decision rule is:

$$\begin{cases} \text{Reject } H_0, & \text{if } T_n \in \mathcal{R}, \\ \text{Do not reject } H_0, & \text{if } T_n \notin \mathcal{R}. \end{cases}$$

Typically, the rejection region is chosen such that T_n falling in $\mathcal{R}$ corresponds to T_n being "large" in magnitude, meaning far from the typical values expected under H_0. In this chapter, for simplicity, we will mostly consider rejection regions of the form

$$\mathcal{R}(c) = \{t : t > c\},$$

for some threshold (critical value) c. In other words, we define T_n so that larger values indicate more evidence against H_0, and we reject for sufficiently large T_n. (For two-sided tests, "large" will effectively mean large in absolute value; we will address this later.)

The choice of the critical value c is crucial. It will be determined based on controlling the probability of making an error, as we discuss next.

5.3.1 Type I and Type II Errors

Because our decision is based on random data, there is a chance we draw the wrong conclusion. There are two types of errors in hypothesis testing:

- A **Type I error** occurs when we reject H_0 even though H_0 is actually true (a "false positive").
- A **Type II error** occurs when we fail to reject H_0 even though H_0 is false (a "false negative").

Any hypothesis test can result in one of four possible outcomes, as summarized in Table 5.1.

	Do not reject H_0	Reject H_0
H_0 true	Correct decision	Type I error
H_0 false	Type II error	Correct decision

Table 5.1: Outcomes of a hypothesis test and associated error types.

There is an inherent trade-off between Type I and Type II errors: if we make the rejection region $\mathcal{R}$ very "conservative" (small) to rarely reject H_0 (thus minimizing Type I errors), we increase the chance of missing real effects (more Type II errors). Conversely, if we make $\mathcal{R}$ very permissive (reject H_0 for even slight evidence), we reduce Type II errors but incur more Type I errors.

In practice, hypothesis testing procedures are usually designed to control the probability of a Type I error at some pre-specified low level (denoted α). This α is called the **significance level** of the test. It represents the tolerable probability of wrongly rejecting a true null hypothesis.

A classical analogy (attributed to Wasserman, 2003) is that hypothesis testing is like a criminal trial: the accused is presumed innocent (H_0 true) until proven guilty. The court requires "strong evidence" to convict (reject H_0) because convicting an innocent person (Type I error) is deemed worse than letting a guilty person go free (Type II error). Thus, we bias the procedure toward not rejecting H_0 unless the data provide compelling evidence against it.

> *"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."*
>
> — *Wasserman (2003)*

5.3.2 Size, Power, and p-Values

We now define measures to quantify the error probabilities of a test:

Definition 5.2 (Size (Significance Level) and Power)**.** The **size** of a test is the probability of committing a Type I error, evaluated at the worst case under H_0. In other words, it is the maximal rejection probability when the null hypothesis is true:

$$\alpha \;=\; \sup_{\theta\in\Theta_0} P_\theta(T_n \in \mathcal{R} \mid H_0 \text{ true}) \;=\; P_{\theta_0}(T_n \in \mathcal{R} \mid H_0 \text{ true}),$$

if H_0 is simple with $\theta = \theta_0$. Often α is chosen in advance (common values are 0.10, 0.05, or 0.01, with $\alpha = 0.05$ being a popular convention).

The **power** of a test at a particular alternative $\theta \in \Theta_1$ is the probability of correctly rejecting the null when that alternative is true:

$$\text{Power}(\theta) \;=\; P_\theta(T_n \in \mathcal{R} \mid H_0 \text{ false}).$$

The *power function* of the test is the function $\theta \mapsto P_\theta(\text{reject } H_0)$ for θ in the parameter space. A test's **power (at a given alternative)** is $1-\beta$, where β is the probability of Type II error (i.e. $\beta = P(\text{fail to reject } H_0 \mid \theta \in \Theta_1)$).

When we say a test has level (size) α, we mean its Type I error probability is controlled to be α (often exactly α in the worst case), and we often seek tests that maximize power among those with a given size α.

In practice, we choose the critical value c (and thus the rejection region $\mathcal{R}$) to achieve a desired size α. For example, we might choose c such that $P(T_n > c \mid H_0) = \alpha$. This ensures $P(\text{Type I error}) = \alpha$. We then hope that for plausible alternatives, $P(T_n > c \mid H_1)$ is as large as possible (high power), but there is usually a trade-off.

Traditionally, researchers often use $\alpha = 0.05$ as a benchmark for "statistical significance." This choice is somewhat arbitrary (why not 0.01 or 0.10?), and in some contexts a different significance level may be more appropriate. Instead of focusing solely on a fixed α, it is often useful to consider the *p-value* of a test outcome:

Definition 5.3 (p-Value)**.** Given the observed value t_{obs} of the test statistic T_n, the **p-value** is the smallest significance level α at which the null hypothesis would be rejected. Formally, it is

$$\text{p-value} \;=\; \inf\{\alpha \in (0,1) : t_{\text{obs}} \in \mathcal{R}(c_\alpha)\},$$

where c_α is the critical value that yields a test of size α. Equivalently, the p-value is the probability (under H_0) of obtaining a test statistic as extreme as or more extreme than the observed t_{obs}.

A small p-value indicates that the observed data are very unlikely under H_0, hence provides strong evidence against H_0. We reject H_0 if the p-value is less than our chosen α. For example, if we observe a p-value of 0.003, this is much smaller than $\alpha = 0.05$, so we would reject H_0 (and typically report the result as "significant at the 5% level" or even the 1% level, since 0.003 $<$ 0.01).

It is crucial to understand that a large p-value does *not* constitute evidence that H_0 is true; it merely indicates a lack of evidence against H_0. A high p-value could occur either because H_0 is true or because H_0 is false but our test had low power or the particular sample did not exhibit a strong effect.

Having defined the general framework of hypothesis testing, we next discuss specific common tests: two-sided tests for an equality hypothesis, and one-sided tests for an inequality hypothesis.

5.4 Two-Sided Hypothesis Tests

Suppose we have a sample $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F_X$ (i.e. i.i.d. observations from some distribution), and we are interested in a real-valued parameter $\theta = \theta(F_X)$ (for example, θ could be $E[X]$, the population mean). We have an estimator $\hat{\theta}_n$ for θ. In many cases, we know (or can derive) the approximate sampling distribution of $\hat{\theta}_n$. A very common situation is that $\hat{\theta}_n$ is asymptotically normal:

$$\frac{\hat{\theta}_n - \theta}{\text{s.e.}(\hat{\theta}_n)} \xrightarrow{d} N(0,1) \qquad \text{as } n \to \infty,$$

where s.e.$(\hat{\theta}_n)$ denotes the standard error of $\hat{\theta}_n$ (the standard deviation of its sampling distribution, or an estimate of that). This was the case for many estimators discussed in the previous chapter (by the Central Limit Theorem or other large-sample results).

Now we consider testing whether θ equals some specific value θ_0. The hypotheses are:

$$H_0 : \theta = \theta_0, \qquad H_1 : \theta \neq \theta_0,$$

a two-sided hypothesis (the alternative allows θ to be either less or greater than θ_0). Intuitively, if the estimator $\hat{\theta}_n$ is much different from θ_0, that would be evidence against H_0.

A natural choice of test statistic in this scenario is the standardized difference between $\hat{\theta}_n$ and the null value θ_0:

$$T_n = \left| \frac{\hat{\theta}_n - \theta_0}{\text{s.e.}(\hat{\theta}_n)} \right| .$$

We take the absolute value because departures on either side (too high or too low relative to θ_0) are both evidence against H_0. By taking T_n to be nonnegative (absolute value), we can use a single rejection region of the form $T_n > c$. In words, if $\hat{\theta}_n$ is sufficiently far from θ_0 in either direction, we reject H_0.

How do we choose the critical value c? Under H_0 (which specifies $\theta = \theta_0$), for large n the statistic T_n should approximately follow a half-normal or (more conveniently) we can say $\frac{\hat{\theta}_n - \theta_0}{\text{s.e.}(\hat{\theta}_n)} \approx N(0,1)$. Thus $T_n = |Z|$ for $Z \sim N(0,1)$ under H_0 in the large-sample limit. We want

$$P_{H_0}(T_n > c) = \alpha,$$

to achieve size α. If $Z \sim N(0,1)$,

$$P(|Z| > c) = 2\,[1 - \Phi(c)],$$

since $P(|Z| > c)$ is the probability of falling in either tail beyond c. Setting this equal to α, we get

$$2\,[1 - \Phi(c)] = \alpha,$$

which implies $1 - \Phi(c) = \alpha/2$, so $\Phi(c) = 1 - \frac{\alpha}{2}$. Thus c should be the $1 - \frac{\alpha}{2}$ quantile of the standard normal distribution. We denote this critical value by

$$z_{1-\alpha/2} = \Phi^{-1}(1 - \tfrac{\alpha}{2}),$$

the $(1 - \frac{\alpha}{2})$-quantile of $N(0,1)$. For example, if $\alpha = 0.05$, then $1 - \alpha/2 = 0.975$, so $z_{0.975} \approx 1.96$. This is the familiar 1.96 appearing in 95% confidence intervals and two-sided 5% tests.

The following theorem formalizes the justification of this test in large samples:

Theorem 5.4 (Asymptotic size of two-sided Z-test)**.** *Let $\hat{\theta}_n$ be an estimator for θ such that*

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

Consider the test statistic $T_n = \left|\frac{\hat{\theta}_n - \theta_0}{\text{s.e.}(\hat{\theta}_n)}\right|$. Then under $H_0 : \theta = \theta_0$, we have

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

as $n \to \infty$. In other words, the test that rejects H_0 if $T_n > z_{1-\alpha/2}$ has (asymptotic) size α.

Proof. Under H_0, $\theta = \theta_0$. By assumption,

$$\frac{\hat{\theta}_n - \theta_0}{\text{s.e.}(\hat{\theta}_n)} \xrightarrow{d} N(0,1) \quad \text{as } n \to \infty.$$

Let $Z_n = \frac{\hat{\theta}_n - \theta_0}{\text{s.e.}(\hat{\theta}_n)}$ and note $T_n = |Z_n|$. We want the probability of rejection:

$$\begin{aligned}
P(T_n > c \mid H_0) &= P\Big(|Z_n| > c \mid H_0\Big) \\
&= P(Z_n > c \mid H_0) \;+\; P(Z_n < -c \mid H_0) \\
&= 1 - P(Z_n \le c \mid H_0) \;+\; P(Z_n < -c \mid H_0) \\
&= 1 - P(Z_n \le c \mid H_0) \;+\; P(Z_n \le -c \mid H_0).
\end{aligned}$$

As $n \to \infty$, Z_n converges in distribution to $N(0,1)$, so by the Continuous Mapping Theorem the above probability converges to

$$1 - P(Z \leq c) + P(Z \leq -c),$$

where $Z \sim N(0,1)$. But $P(Z \leq -c) = \Phi(-c) = 1 - \Phi(c)$ (since Φ is the CDF of $N(0,1)$). Therefore the limit is

$$1 - \Phi(c) + (1 - \Phi(c)) \; = \; 2\,[1 - \Phi(c)].$$

Setting $c = z_{1-\alpha/2}$, by definition $1 - \Phi(c) = \alpha/2$. Thus $2[1-\Phi(c)] = 2(\alpha/2) = \alpha$. This proves that

$$\lim_{n\to\infty} P(T_n > z_{1-\alpha/2} \mid H_0) = \alpha,$$

as required. □

This result shows that for large n, our test rejects H_0 with probability about α when H_0 is true (thus controlling the Type I error rate at α).

In practice, then, we reject $H_0 : \theta = \theta_0$ at significance level α if

$$T_n = \left| \frac{\hat{\theta}_n - \theta_0}{\text{s.e.}(\hat{\theta}_n)} \right| > z_{1-\alpha/2}.$$

Equivalently, we can phrase the decision in terms of the p-value. The p-value in this two-sided test is

$$\text{p-value} = 2\left[1 - \Phi\Big(|z_{\text{obs}}|\Big)\right],$$

where $z_{\text{obs}} = \frac{\hat{\theta}_n - \theta_0}{\text{s.e.}(\hat{\theta}_n)}$ is the observed standardized value. This formula $2(1 - \Phi(|z|))$ gives the two-tail area under the standard normal beyond the observed $|z|$. We reject H_0 at level α if and only if p-value $< \alpha$.

Let's derive that explicitly: the observed test statistic is $T_n^{obs} = |z_{\text{obs}}|$. The p-value is the probability (under H_0) of seeing a result as extreme as what we saw. "As extreme as" means $|Z| \geq |z_{\text{obs}}|$ if $Z \sim N(0,1)$. So

$$\text{p-value} = P_{H_0}\big(|Z| \geq |z_{\text{obs}}|\big) = 2\,[1 - \Phi(|z_{\text{obs}}|)].$$

Setting this equal to α and solving for $|z_{\text{obs}}|$ yields $|z_{\text{obs}}| = z_{1-\alpha/2}$. Thus rejection $|z_{\text{obs}}| > z_{1-\alpha/2}$ is equivalent to p-value $< \alpha$, as it should be.

Remark. In finite samples, if the distribution of the estimator is known, one could use the exact critical value from that distribution. For example, if X_i are i.i.d. normal and $\hat{\theta}_n = \bar{X}$, and we estimate the variance from data, then $(\bar{X} - \theta_0)/(S/\sqrt{n})$ follows a Student t distribution with $n-1$ degrees of freedom under H_0. In that case, the exact finite-sample test would reject if $|\bar{X} - \theta_0| > t_{n-1,\,1-\alpha/2}\, S/\sqrt{n}$, where $t_{n-1,\,1-\alpha/2}$ is the $1-\alpha/2$ quantile of t_{n-1}. In large n, $t_{n-1,\,0.975} \approx 1.96$ and the distinction blurs. In econometrics and many large-sample settings, one often simply uses the normal approximation as we have done, especially when n is moderate or large.

5.5 One-Sided Hypothesis Tests

We now consider testing a hypothesis where the alternative is one-sided. There are two forms, depending on the direction of the inequality:

- **Right-tailed test:** $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$. Here the alternative claims θ is greater than some threshold θ_0, and the null says θ is at most θ_0.
- **Left-tailed test:** $H_0 : \theta \geq \theta_0$ versus $H_1 : \theta < \theta_0$. Here the alternative claims θ is less than θ_0.

Many practical questions lead to one-sided alternatives. For instance, in our earlier example of returns to education, we might specifically hypothesize H_0: “the return is *non-positive* (≤ 0)” against H_1: “the return is *positive* (> 0)”. This is a right-tailed test because the alternative is that θ is greater than 0.

The testing framework is similar, but now “extreme” evidence against H_0 occurs only in one direction. We want to reject H_0 only if $\hat{\theta}_n$ is sufficiently larger than θ_0 (in a right-tailed test) or sufficiently smaller (in a left-tailed test). We can adapt our test statistic accordingly so that, again, large values of T_n favor the alternative:

- For $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$, a natural choice is

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

If the estimator $\hat{\theta}_n$ is much bigger than θ_0, then T_n will be large (positive), signaling evidence for H_1. If $\hat{\theta}_n$ is below or near θ_0, T_n will not exceed the threshold.

- For $H_0 : \theta \geq \theta_0$ versus $H_1 : \theta < \theta_0$, we can simply take the negative of the above:

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

This way, if $\hat{\theta}_n$ is far below θ_0, the test statistic T_n becomes large (positive), indicating evidence for the alternative. (Equivalently, one could keep $T_n = (\hat{\theta}_n - \theta_0)/\text{s.e.}$ but then define the rejection region for a left-tailed test as $T_n < -c$, which is less convenient for a unified treatment. By flipping the sign in T_n, we maintain the convention of rejecting for T_n exceeding a positive critical value c.)

Now T_n is (asymptotically) $N(0, 1)$ under the null hypothesis (since under H_0, $\theta = \theta_0$ and thus $\hat{\theta}_n - \theta_0$ is centered at 0). We want to choose c such that

$$P_{H_0}(T_n > c) = \alpha.$$

If $T_n \approx N(0, 1)$ under H_0, then $P(T_n > c) = 1 - \Phi(c)$. Setting this equal to α gives $1 - \Phi(c) = \alpha$, or $\Phi(c) = 1 - \alpha$. Therefore c should be the $1 - \alpha$ quantile of the standard normal:

$$c = z_{1-\alpha} = \Phi^{-1}(1 - \alpha).$$

For example, with $\alpha = 0.05$, $z_{0.95} \approx 1.645$. The next theorem mirrors Theorem 5.4 for the one-sided case:

Theorem 5.5 (Asymptotic size of one-sided test). *Suppose* $\frac{\hat{\theta}_n - \theta}{\text{s.e.}(\hat{\theta}_n)} \xrightarrow{d} N(0, 1)$ *as* $n \to \infty$. *For testing* $H_0 : \theta \leq \theta_0$ *vs.* $H_1 : \theta > \theta_0$, *consider the test statistic* $T_n = \frac{\hat{\theta}_n - \theta_0}{\text{s.e.}(\hat{\theta}_n)}$. *Then under* H_0,

$$P(T_n > z_{1-\alpha} \mid H_0) \ \to \ \alpha.$$

In other words, rejecting H_0 *when* $T_n > z_{1-\alpha}$ *yields an (asymptotic) level-*α *test. An analogous result holds for the left-tailed test using* $T_n = \frac{\theta_0 - \hat{\theta}_n}{\text{s.e.}(\hat{\theta}_n)}$.

Proof. Under $H_0 : \theta \leq \theta_0$, the "least favorable" case (that maximizes the Type I error) is $\theta = \theta_0$. So assume $\theta = \theta_0$. Then as n grows large, $T_n =$

$\frac{\hat{\theta}_n-\theta_0}{\text{s.e.}(\hat{\theta}_n)} \xrightarrow{d} N(0,1)$. We have

$$P(T_n > c \mid H_0) \;=\; 1 - P(T_n \leq c \mid H_0) \;\to\; 1 - \Phi(c),$$

as $n \to \infty$. If we choose $c = z_{1-\alpha}$, then $1 - \Phi(c) = 1 - (1-\alpha) = \alpha$. Thus $\lim_{n\to\infty} P(T_n > z_{1-\alpha} \mid H_0) = \alpha$, as required.

(The argument for the left-tailed test is similar: by defining $T_n = (\theta_0 - \hat{\theta}_n)/\text{s.e.}$, under $H_0 : \theta = \theta_0$ we again have $T_n \xrightarrow{d} N(0,1)$, and $P(T_n > z_{1-\alpha}|H_0) \to \alpha$. Rejecting for large T_n corresponds to $\hat{\theta}_n$ being sufficiently below θ_0, as desired.) □

Thus, for a one-sided test at significance α, we reject H_0 if

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

In a right-tailed test, this means

$$\frac{\hat{\theta}_n - \theta_0}{\text{s.e.}(\hat{\theta}_n)} > z_{1-\alpha},$$

and in a left-tailed test it means

$$\frac{\theta_0 - \hat{\theta}_n}{\text{s.e.}(\hat{\theta}_n)} > z_{1-\alpha}.$$

Equivalently, we can report a p-value. In a one-sided test, the p-value is the one-tail probability beyond the observed T_n under a $N(0,1)$ distribution. For example, if our test statistic is $T_n = \frac{\hat{\theta}_n-\theta_0}{\text{s.e.}(\hat{\theta}_n)}$ (right-tailed test) and we observe $T_n^{obs} = t$, then

$$\text{p-value} = P_{H_0}(Z \geq t) = 1 - \Phi(t).$$

If this p-value is below α, we reject H_0. For a left-tailed test (with $T_n = \frac{\theta_0-\hat{\theta}_n}{\text{s.e.}}$ as defined above), the p-value would likewise be $1 - \Phi(t_{\text{obs}})$, since we have defined T_n so that large values (in the right tail) indicate significance in either case.

Example 5.6. (College wage example revisited) In the hypothesis $H_0 : \mu_{Y|1} \geq 600$ vs $H_1 : \mu_{Y|1} < 600$ (do college grads earn less than 600 on average?), the alternative is left-tailed. We would construct

$$T_n = -\frac{\hat{\mu}_{Y|1} - 600}{\text{s.e.}(\hat{\mu}_{Y|1})} = \frac{600 - \hat{\mu}_{Y|1}}{\text{s.e.}(\hat{\mu}_{Y|1})}.$$

If, say, $\hat{\mu}_{Y|1}$ from our sample is significantly below 600, T_n will be large. We reject H_0 for $T_n > z_{1-\alpha}$. At $\alpha = 0.05$, this means $T_n > 1.645$. Equivalently, we could compute the p-value: if the observed T_n is, for instance, 2.0, then the p-value $= 1 - \Phi(2.0) \approx 0.0228$. This is below 0.05, so we reject H_0 and conclude the average is significantly less than \$600. On the other hand, if $\hat{\mu}_{Y|1}$ was above 600, then T_n would likely be small (and possibly even negative, in which case certainly T_n is not > 1.645), and we would not reject H_0.

5.6 Hypothesis Tests and Confidence Intervals

There is a close duality between hypothesis testing and confidence intervals. In fact, constructing a confidence interval for a parameter can be viewed as performing hypothesis tests for all possible parameter values and collecting those values for which the test would *not* reject. This idea is formalized as follows:

Consider testing

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

at significance level α, for each possible value $\tilde{\theta}_0 \in \Theta$. For each $\tilde{\theta}_0$, we imagine plugging it into H_0 and performing the corresponding two-sided test. Now define

$$C_n = \{\tilde{\theta}_0 \in \Theta : \text{the test fails to reject } H_0 : \theta = \tilde{\theta}_0 \text{ at level } \alpha\}.$$

In other words, C_n is the set of all parameter values that are *consistent with the data at the α significance level* (i.e. that would not be rejected by a level-α test).

It turns out that C_n is exactly a $(1-\alpha)$ **confidence interval** for θ. This is the reasoning behind the common teaching that "we reject $H_0 : \theta = \theta_0$ at level α if and only if θ_0 lies outside the $(1-\alpha)$ confidence interval for θ."

Illustration. Suppose our test statistic for $H_0 : \theta = \tilde{\theta}_0$ is

$$T_n = \left| \frac{\hat{\theta}_n - \tilde{\theta}_0}{\text{s.e.}(\hat{\theta}_n)} \right|,$$

and we reject H_0 when $T_n > z_{1-\alpha/2}$ (this is the two-sided test we discussed earlier). For a given candidate value $\tilde{\theta}_0$, the condition for *not* rejecting H_0 is

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

This condition can be rewritten as:

$$\begin{aligned}\left|\frac{\hat{\theta}_n - \tilde{\theta}_0}{\text{s.e.}(\hat{\theta}_n)}\right| \leq z_{1-\alpha/2} &\iff -z_{1-\alpha/2} \leq \frac{\hat{\theta}_n - \tilde{\theta}_0}{\text{s.e.}(\hat{\theta}_n)} \leq z_{1-\alpha/2} \\ &\iff -z_{1-\alpha/2}\,\text{s.e.}(\hat{\theta}_n) \leq \hat{\theta}_n - \tilde{\theta}_0 \leq z_{1-\alpha/2}\,\text{s.e.}(\hat{\theta}_n) \\ &\iff -z_{1-\alpha/2}\,\text{s.e.}(\hat{\theta}_n) + \hat{\theta}_n \leq \tilde{\theta}_0 \leq z_{1-\alpha/2}\,\text{s.e.}(\hat{\theta}_n) + \hat{\theta}_n.\end{aligned}$$

The last line describes exactly the interval

$$\Big[\hat{\theta}_n - z_{1-\alpha/2}\,\text{s.e.}(\hat{\theta}_n), \quad \hat{\theta}_n + z_{1-\alpha/2}\,\text{s.e.}(\hat{\theta}_n)\Big].$$

Thus

$$C_n = \Big\{\tilde{\theta}_0 : \tilde{\theta}_0 \in [\hat{\theta}_n \pm z_{1-\alpha/2}\,\text{s.e.}(\hat{\theta}_n)]\Big\} = [\,\hat{\theta}_n \pm z_{1-\alpha/2}\,\text{s.e.}(\hat{\theta}_n)\,],$$

which is exactly the two-sided $(1-\alpha)$ confidence interval for θ that we derived in the previous chapter.

This confirms the duality: the confidence interval consists exactly of those values that would *not* be rejected by a two-sided hypothesis test at the corresponding level. In practice, this means we can draw conclusions from confidence intervals in lieu of performing explicit hypothesis tests. For example, if a 95% CI for θ is [2.1, 5.3], then $H_0 : \theta = 0$ is clearly rejected at the 5% level (since 0 is not in the interval). On the other hand, if the question of interest is whether θ is positive, we can see the entire 95% CI is positive, which implies $H_0 : \theta \leq 0$ would be rejected at 5

Remark 5.7. This duality holds generally under mild conditions: any confidence set can be seen as the inversion of a family of tests. While we demonstrated it for a symmetric two-sided interval, one-sided hypothesis tests similarly correspond to one-sided confidence bounds. For instance, the set of $\tilde{\theta}_0$ not rejected by a right-tailed test $H_0 : \theta \leq \tilde{\theta}_0$ vs $H_1 : \theta > \tilde{\theta}_0$ is of the form $(-\infty,\ \hat{\theta}_n - z_{1-\alpha}\,\text{s.e.}(\hat{\theta}_n)]$, which is a one-sided $(1-\alpha)$ confidence bound (lower bound) for θ. Thus, constructing confidence intervals is often a more informative way to summarize hypothesis tests for all possible values.

Summary

In this chapter, we reviewed the framework of hypothesis testing, which complements estimation in statistical inference:

- We learned how to formulate null and alternative hypotheses to translate substantive questions into statements about parameters.
- We defined test statistics and decision rules, and understood the types of errors (Type I and II) that can occur.
- We usually fix a significance level α (Type I error rate) and determine a critical value to control α. We introduced the concept of the p-value as an evidence measure against H_0.
- We derived tests for two-sided hypotheses (testing equality) and one-sided hypotheses (testing inequalities) using large-sample Z-statistics. We showed how to compute critical values (e.g. $z_{0.975} \approx 1.96$ for a two-sided 5% test, $z_{0.95} \approx 1.645$ for a one-sided 5% test) and how to calculate p-values for each case.
- Finally, we discussed the duality between confidence intervals and hypothesis tests: a $(1 - \alpha)$ confidence interval is the set of parameter values that would not be rejected at level α.

Equipped with these statistical tools, we are prepared to tackle causal inference questions. We can formulate causal parameters of interest (the "estimands"), identify them under certain assumptions (using our probability theory knowledge), estimate them from data, and then use confidence intervals and hypothesis tests to draw conclusions about causal effects with quantifiable uncertainty.