HYPOTHESIS TESTING

This module examines hypothesis testing as a key method in inferential statistics. It illustrates how researchers use sample data to evaluate claims about population means through systematic steps and decision-making rules. It also addresses potential errors in decision-making and explains the role of the alpha level in managing the risk of incorrect conclusions.

The Logic of Hypothesis (Gravetter, 2021)

Researchers usually cannot observe every individual in a population because this is either impossible or impractical. Instead, they draw conclusions by collecting data from a representative sample and use this information to answer questions about the population.

Hypothesis testing is one of the most widely used techniques in inferential statistics. Although the specific details may differ across studies, the fundamental process of hypothesis testing remains consistent. This module introduces the general procedure for conducting a hypothesis test.

- **Hypothesis testing** is a statistical procedure that uses sample data to assess the validity of a hypothesis about a population.

The fundamental logic of hypothesis testing follows these steps:

1. First, formulate a hypothesis about a population. The hypothesis typically centers on a specific population parameter. For example, a researcher may propose that Filipino adults gain an average of $\mu = 7$ pounds during the holidays.
2. Next, before collecting a sample, the researcher uses the hypothesis to predict the expected values of the sample mean if the hypothesis is true. For example, if the population mean is $\mu = 7$ pounds, the sample mean should be approximately 7 pounds. Some deviation is expected, since a sample rarely represents the population perfectly.
3. Then, the researcher selects a random sample from the population. For example, the researcher may measure the weight change of a sample of $n = 200$ adults during the holiday period.
4. Finally, the researcher compares the sample data with the prediction based on the hypothesis. If the sample mean matches the prediction, the hypothesis appears reasonable. If the difference is large, the hypothesis is likely incorrect.

Researchers usually apply hypothesis testing after completing a research study. The specific steps of the test depend on the research design and the type of data. Later chapters describe different forms of hypothesis testing for different research situations. This chapter focuses on the basic elements common to all hypothesis tests. To do this, it examines the simplest case, which uses a sample mean to test a hypothesis about a population mean.

The Four Steps of a Hypothesis Test

Before the treatment, the original population had a mean tip of $\mu = 16$ percent. After the treatment, the population mean is unknown. Researchers do not know what happens to the average tip when waitresses wear red for all male customers. However, researchers have a sample of $n = 36$ customers who were served by waitresses wearing red. This sample allows researchers to draw conclusions about the unknown population. Hypothesis testing provides a step-by-step method for using sample data to answer questions about a population.

STEP 1: State the Hypotheses

Researchers often cannot observe an entire population, so they use a sample to draw inferences about it. In this example, the original population has a mean tip of μ = 16 percent, but the population mean after the treatment is unknown. A sample of n = 36 customers served by waitresses wearing red is used to draw conclusions about the population.

Hypothesis testing begins by stating two opposing hypotheses about population parameters. These hypotheses are mutually exclusive and exhaustive. Only one can be true, and the sample data determines whether the null hypothesis is rejected or not rejected.

The **null hypothesis (H_0)** states that the treatment has no effect. It predicts no change or difference in the population. In this example:

H_0: $\mu_{red\ shirt} = 16$
(Wearing red does not change the mean tip.)

The **alternative hypothesis (H_1)** states that the treatment has an effect. It predicts a change in the population mean:

H_1: $\mu_{red\ shirt} \neq 16$
(Wearing red changes the mean tip.)

This is a **nondirectional hypothesis test** because the alternative hypothesis does not specify whether tips increase or decrease.

STEP 2: Set the Criteria for a Decision

In hypothesis testing, the researcher uses sample data to evaluate the null hypothesis. The data may support the null hypothesis or suggest that it is wrong. A large difference between the sample data and the null hypothesis leads the researcher to reject it.

To guide this decision, the null hypothesis is used to predict which sample means are expected. Sample means close to the value stated in the null hypothesis are considered consistent with it. Sample means that differ greatly are considered inconsistent.

In this example, the null hypothesis states that the population mean tip is μ = 16 percent. If this is true, the sample mean should be close to 16. To decide what counts as "close" or "very different," researchers examine the **distribution of sample means** for n = 36. This distribution is centered at μ = 16 and is divided into two parts:

- Sample means that are likely if the null hypothesis is true
- Sample means that are very unlikely if the null hypothesis is true

The unlikely values fall in the extreme tails of the distribution.

The Alpha Level and the Critical Region

The boundary between likely and unlikely sample means is defined by the **alpha level (α)**, also called the level of significance. The alpha level is a small probability that defines what "very unlikely" means in a hypothesis test. Common values are α = .05, α = .01, and α = .001.

The most unlikely sample means form the **critical region**. These values are unlikely to occur if the null hypothesis is true. If the sample mean falls in the critical region, the null hypothesis is rejected.

The exact boundaries of the critical region are found using the unit normal table.

- For α = .05, the extreme 5% of values are split evenly between both tails of the distribution. This gives boundary z-scores of ±1.96.
- For α = .01, the boundaries are ±2.58. For α = .001, the boundaries are ±3.30.

These boundary values define which sample outcomes are considered strong evidence against the null hypothesis.

STEP 3: Collect Data and Compute Sample Statistics

In Step 3, the researcher collects data after stating the hypotheses and setting the decision criteria. This order ensures an objective evaluation of the results. In this example, the researcher records tips from male customers while waitresses wear red.

The researcher then summarizes the sample data using appropriate statistics. Here, the key statistic is the **sample mean (M)**. The sample mean is compared with the value stated in the null hypothesis. This comparison forms the core of hypothesis testing.

The **standard error (σ_m)** measures the expected difference between a sample mean and the population mean. Standard error for the sample mean:

$$\sigma_M = \frac{\sigma}{\sqrt{n}}$$

Next, compute the z-score, which indicates how far the sample mean deviates from the hypothesized population mean. The z-score places the sample mean within the distribution of sample means assumed under the null hypothesis.

The formula for the z-score is:

$$z = \frac{X - \mu}{\sigma_M}$$

In words, the formula is:

z = (sample mean − hypothesized population mean) / standard error of the mean

The numerator shows the difference between the data and the hypothesis, and the denominator shows the amount of error expected between a sample mean and a population mean.

STEP 4: Make a Decision

In the final step, the researcher compares the z-score from Step 3 with the decision criteria set in Step 2. The decision depends on whether the z-score falls inside or outside the critical region.

Two outcomes are possible:

1. **Reject the null hypothesis (H_o).** If the sample mean falls in the critical region, the result is very unlikely if H_o is true. The researcher rejects H_o and concludes that there is **evidence** of a treatment effect. In the red shirt example, a sample mean of M = 17.2 with a standard error of 0.5 yields z = 2.40. With α = .05, this value is beyond ±1.96, so H_o is rejected. The conclusion is that there is evidence that wearing red affects tipping.
2. **Fail to reject the null hypothesis (H_o).** If the sample mean does not fall in the critical region, the data are close to the value predicted by H_o. The researcher fails to reject H_o and concludes that there is **no evidence** of a treatment effect. For example, a sample mean of M = 16.4 gives z = 0.80, which is not in the critical region. The conclusion is that there is no evidence that wearing red affects tipping.

In general, the treated sample is compared with the distribution of sample means expected under no treatment effect. If the treated sample looks similar to untreated samples, the treatment appears to have no effect. If it looks very different, the treatment appears to be effective.

Conclusions are stated carefully. Rejecting H_o does not prove that an effect exists. It only shows evidence for an effect. Failing to reject H_o does not prove that there is no effect. It only shows a lack of evidence.

Example

A normal population has a mean of μ = 40 and a standard deviation of σ = 8. A treatment is given to a sample of n = 16 individuals from this population. After the treatment, the sample has a mean of M = 45. A hypothesis test is conducted to evaluate whether the treatment has an effect on the population mean. The test uses an alpha level of α = .05 and is two-tailed. The task is to state the null hypothesis and determine whether the sample provides sufficient evidence to reject it and conclude that the treatment has a significant effect.

Hypotheses	**H_o: $\mu_{treatment} = 40$** (The treatment does not change the population mean.) **H_1: $\mu_{treatment} \neq 40$** (The treatment changes the population mean.)
Alpha	α = .05 (two-tailed)
Critical region	z = ± 1.96
Sample statistic	$\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{8}{\sqrt{16}} = \frac{8}{4} = 2$ $z = \frac{X - \mu}{\sigma_M} = \frac{45 - 40}{2} = \frac{5}{2} = 2.50$
Decision	Reject H_o
Result	The sample mean falls in the critical region (z = 2.50). There is evidence of a significant treatment effect.

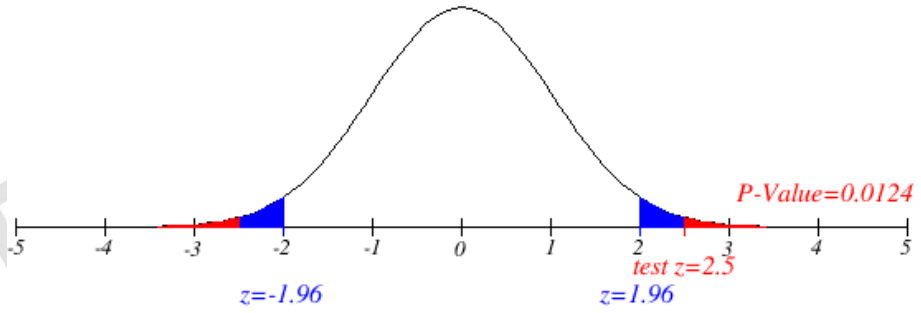

Figure 1. Graphical Illustration of a Two-Tailed z-Test

Type I and Type II Errors

Hypothesis testing is an inferential process because it uses limited information to make a general conclusion. A sample provides incomplete information about a population, yet researchers use it to draw conclusions about the population. Because of this limitation, incorrect conclusions are possible. Even when a sample usually represents the population well, it can sometimes be misleading and lead to an incorrect decision.

Two types of errors can occur in hypothesis testing:

Type I Error

A **Type I error** occurs when a researcher rejects a null hypothesis that is actually true. In most research situations, this means the researcher concludes that there is evidence of a treatment effect when the treatment has no real effect.

This error happens because some samples differ greatly from the population by chance. If a researcher selects an extreme sample, the data may appear to show a strong effect even when no effect exists. In the tipping example, a sample of men who already tip well could still have a high average tip even if the red shirt has no effect. The researcher may then conclude that the treatment worked, which is incorrect.

A Type I error is not the result of carelessness. The researcher follows proper procedures and bases the decision on the sample data. The problem occurs because the sample gives misleading information.

Type I errors have serious consequences. Researchers may report or publish false results. Other researchers may build theories or conduct studies based on these false findings. This process wastes time and resources and adds incorrect information to scientific knowledge.

Probability of a Type I Error

The probability of a Type I error depends on the **alpha level (α)**. The alpha level defines how unlikely a sample result must be to reject the null hypothesis. If $\alpha = .05$, then 5% of all possible samples fall in the critical region when the null hypothesis is true. This means there is a 5% chance of rejecting a true null hypothesis.

- The **alpha level is the probability of committing a Type I error**. By choosing a smaller alpha level, researchers reduce the risk of making this error.

In summary, researchers reject the null hypothesis when the sample data fall in the critical region. Most of the time, this decision correctly identifies a real treatment effect. Sometimes, chance alone places the sample in the critical region. When this happens, a Type I error occurs. The risk of this error is controlled by the alpha level selected by the researcher.

Type II Error

Whenever a researcher fails to reject the null hypothesis, there is a risk of a **Type II error**. A Type II error occurs when the null hypothesis is false, but the test fails to detect the effect.

- A **Type II error** occurs when a researcher fails to reject a null hypothesis that is actually false. In most studies, this means the treatment has a real effect, but the test fails to show it.

This error often happens when the treatment effect is small. The treatment influences the sample, but the effect is not large enough to move the sample mean into the critical region. Because the sample does not differ much from the population, the decision is to fail to reject the null hypothesis, concluding that there is no evidence of an effect.

The consequences of a Type II error are usually less serious than those of a Type I error. The researcher does not report a false effect. The researcher may accept the result or repeat the study, often with a larger sample, to better detect the effect.

The probability of a Type II error does not have a single fixed value. It depends on factors such as sample size and effect size. This probability is represented by **β (beta)**.

Summary of Decisions and Errors

A hypothesis test leads to one of two decisions:

1. Reject the null hypothesis and conclude that the treatment has an effect.
2. Fail to reject the null hypothesis and conclude that there is no evidence of an effect.

Each decision carries a possible error:

- **Type I error (α):** Rejecting a true null hypothesis.
- **Type II error (β):** Failing to reject a false null hypothesis.

The risk of a Type I error is especially important because it can lead to false reports. This risk is fully controlled by the researcher through the choice of the alpha level at the start of the hypothesis test.

Selecting an Alpha Level

The **alpha level (α)** serves two key purposes in hypothesis testing. First, it defines which outcomes are considered "very unlikely," which sets the boundaries of the critical region. Second, it determines the probability of making a **Type I error**. The value of alpha chosen at the start of a hypothesis test affects both of these functions.

The main goal in selecting an alpha level is to reduce the risk of a Type I error. For this reason, alpha values are small probabilities. By convention, the largest commonly accepted value is **α = .05**. This value means there is a 5% chance of rejecting a true null hypothesis. Because Type I errors can have serious consequences, many researchers prefer smaller values, such as α = .01 or α = .001, to reduce the risk of false conclusions.

Choosing the smallest possible alpha may seem ideal, but this choice creates another problem. As alpha decreases, the hypothesis test requires stronger evidence to reject the null hypothesis. This happens because the **critical region moves farther into the tails of the distribution**.

When alpha is smaller, the sample mean must be farther from the null hypothesis value to reach the critical region. If the alpha level becomes extremely small, the risk of a Type I error becomes very low, but rejecting the null hypothesis becomes very difficult. This situation would require a very large treatment effect or a very large sample size.

Researchers, therefore, aim to balance the risk of a Type I error with the test's ability to detect real effects. Alpha levels of **.05, .01, and .001** are commonly used because they provide a reasonable balance between these two concerns.

References

Gravetter, F. J., Wallnau, L. B., Forzano, L. A., & Witnauer, J. E. (2021). *Essentials of Statistics for the Behavioral Sciences (10th ed.).* Cengage.