22. HYPOTHESIS TESTING

Often, we need to make decisions based on incomplete information. Do the data support some belief ("hypothesis") about the value of a population parameter?

- Is OJ Simpson guilty? (DNA Fingerprinting).

- Does Viagra work?

- Does a market analyst have any skill at forecasting the price of Netscape?

- Is the climate changing?

- Does a high fiber diet help prevent colon cancer?

- Are students in this class able to tell the difference between Diet Coke and Diet Pepsi?

- Do "4 oz" bags of M&Ms actually contain 4 ounces of product, on average?

- Do McDonalds "Quarter Pounders" actually weigh less than 1/4 pound, before cooking, on the average?

Hypothesis testing provides answers to questions like these.

It must be emphasized, however, that our conclusions may be wrong, because of sampling variability.

Here, we focus attention on methods of making decisions about a population mean, μ.

Eg: Based on a sample of 50 “Quarter Pounders,” we want to decide if the mean weight μ of all “Quarter Pounders” is actually less than 0.25 pounds.

• Statistical hypothesis tests are set up like criminal court cases: the defendant is presumed innocent until proven guilty beyond a reasonable doubt.

Thus, we will start with the assumption that μ is in fact 0.25. This is called the **null hypothesis**, and is written as H_0: $\mu = 0.25$. This represents the “status quo”, since McDonald’s has been marketing these hamburgers as “quarter pounders” for years.

•In most applications, it is hoped that the data will provide enough evidence to allow us to **reject** the null hypothesis in favor of the **alternative hypothesis**, which is in this case H_A: $\mu < 0.25$.

The alternative hypothesis is sometimes called the research hypothesis, since it is often the hypothesis that the researcher wishes to prove. In the above example, a demonstration that $\mu < 0.25$ would be newsworthy. You (the researcher) might be able to win a Pulitzer prize for your investigative reporting.

• The evidence is provided by a single number called a **test statistic**, obtained from the observed data.

Based on the value of the test statistic, we will decide either to reject, or not to reject, the null hypothesis. Specifically, we reject H_0 whenever the test statistic lies in a pre-specified range of values called the **rejection region**, or **critical region**.

As in the courtroom analogy, there are two different kinds of incorrect decisions, or errors.

- **Type I Error**: Rejecting H_0 when H_0 is true.

- **Type II Error**: Not rejecting H_0 when H_0 is false.

Generally, we try to formulate the hypotheses so that a Type I error is considered more serious than a Type II error.

Using statistical theory, we can control the probability of making a Type I error. This probability is denoted by α and called the **significance level** of the test.

We are free to choose the value of α. Clearly it is desirable for α to be small, but there is a tradeoff here: the smaller α is, the larger the probability of a Type II error will be.

- It is customary to use either $\alpha = 0.05$ or $\alpha = 0.01$.

Examples: What are H_0, H_A, Type I errors and Type II errors in the following situations:

- Murder Trials

- Smoke Alarms

- Pregnancy Tests

- Boy Who Cried Wolf.

Format For Hypothesis Tests

(1) Before examining the data, formulate H_0, H_A, and select α.

(2) Examine the data, evaluate the test statistic, and draw your conclusion.

To ensure the validity of the test results, H_0 and H_A must be formulated *before* the data are examined. The null hypothesis always takes the form H_0: $\mu = \mu_0$, so that μ_0 is the value of the population mean if H_0 is true.

Depending on what we are trying to prove, we take H_A as either:

- H_A: $\mu \neq \mu_0$ (A **two-sided** alternative)
- H_A: $\mu < \mu_0$ (A **one-sided** alternative)
- H_A: $\mu > \mu_0$ (A **one-sided** alternative).

Suppose for now that σ is known. Suppose also that either the population is normally distributed or that the sample size is large enough for the Central Limit Theorem to hold.

Then the test statistic is $z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$.

The z statistic provides a standardized measure of the difference between the hypothesized value of μ and the actual value of the sample mean $\bar{x}$.

If H_0 is true, then Z is (approximately) standard normal. (Why?)

In hypothesis testing, we simply check whether z is “too large” or “too small” to have plausibly come from a standard normal distribution. If so, then we reject H_0.

How To Perform the Test

• If z lies in the rejection region (given below), reject H_0. Otherwise, do not reject H_0.

• The rejection region for a level α test depends on the nature of the alternative hypothesis.

If H_A is two-sided, we get a **two-tailed test**.

If H_A is one-sided, we get a **one-tailed test**.

(Note that there are two kinds of one-tailed tests).

H_A	Rejection Region
$\mu \neq \mu_0$	$\|z\| > z_{\alpha/2}$
$\mu < \mu_0$	$z < -z_\alpha$
$\mu > \mu_0$	$z > z_\alpha$

For example, to test H_0: $\mu = \mu_0$ against the two-sided alternative H_A: $\mu \neq \mu_0$, we reject H_0 if either $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$.

The dividing lines, in this case $-z_{\alpha/2}$ and $z_{\alpha/2}$, are called **critical values**. This does give a level α test, since

$$\begin{aligned}\text{Prob(Type I Error)} &= \text{Prob(Reject } H_0 \mid H_0 \text{ is True)}\\ &= \text{Prob (} |Z| > \mathrm{z}_{\alpha/2} \mid H_0 \text{ is true)}\\ &= \text{Prob (|Std Normal RV|} > \mathrm{z}_{\alpha/2}) = \alpha \ .\end{aligned}$$

• Explain why the above 1-tailed test criteria are also correct.

There are four critical values worth memorizing:

$$z_{0.05} = 1.645,\ z_{0.025} = 1.96,\ z_{0.01} = 2.326,\ z_{0.005} = 2.576\ .$$

These correspond to $\alpha = 5\%$ (one-tailed and two-tailed), and to $\alpha = 1\%$ (one-tailed and two-tailed).

Eg 1: For the "Quarter Pounders" example, test H_0: $\mu = 0.25$ against H_A: $\mu < 0.25$ at the 5% level of significance, assuming that $n = 50$, $\sigma = 0.035$ and $\bar{x} = 0.24$.

Eg 2: The Domino's Pizza closest to NYU advertises that their average delivery time to NYU is at most 20 minutes. A sample of 32 delivery times had an average of 24 minutes. The true standard deviation of delivery times is 10 minutes. Is there sufficient evidence to reject Domino's Pizza's claim at the 1% level of significance?

Eg 3: In the Pepsi example, for 100 "2-liter" bottles, the average amount of Pepsi filled by the machine was $\bar{x} = 1.985$ liters. The population standard deviation is known to be 0.05 liters. If the expected amount (per bottle) of Pepsi filled by the machine, denoted by μ, is not equal to 2, then the machine is "out of control", and must be shut down for repairs. Test whether the machine is out of control, at the 5% level of significance.

The Confidence Interval Approach To Hypothesis Testing For Two-Tailed Tests

The 95% confidence interval for the mean in the situation of Example 3 was (1.975, 1.995), which does not contain the null hypothesis value of 2 liters. This would seem to indicate that a mean value of 2 liters is inconsistent with the available information, and indeed we rejected the null hypothesis at level 0.05 in Eg 3.

In general, it can be shown that the confidence interval consists of the values of μ_0 which would *not* be rejected in a 2-tailed hypothesis test of $H_0 : \mu = \mu_0$ versus $H_A : \mu \neq \mu_0$.

Thus, if the hypothesized mean falls in the confidence interval, then we *cannot* reject H_0 in favor of H_A.

On the other hand, if the hypothesized mean does not fall in the confidence interval, then we *can* reject H_0 in favor of H_A.

These properties allow us, if we wish, to perform a 2-tailed test at level α by simply examining the $(1-\alpha)100\%$ confidence interval and seeing if the null value falls in the interval.