Applied Statistical Decision Making
Hypothesis Testing

Monia Ranalli
monia.ranalli@uniroma2.it

Objectives

- understand the logic of **hypothesis testing**
- learn how to **set up hypotheses**
- learn how to conduct hypothesis testing for **population proportion** using the p-**value approach**
- learn how to conduct hypothesis testing for **population mean** using the p-**value approach**
- understand the **connection between confidence intervals and hypothesis tests**
- define **Type I and Type II errors** and explain **which error** is **more important to minimize**

Introduction to Hypothesis Testing I

A hypothesis, is a statement about a population, typically of the form that a certain parameter takes a particular numerical value or falls in a certain range of values. To perform a statistical test there are certain steps to follow

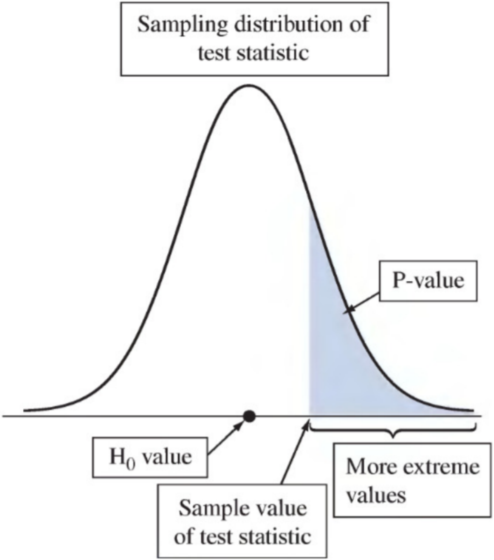
- **1. Assumptions**. Each hypothesis test makes certain assumptions about some characteristics of the distribution of the population and the sampling distribution of a statistic.
- **2. Hypothesis**. Each hypothesis test includes two hypothesis about the population:
  - the null hypothesis, H_0, which is a statement of a particular parameter value;
  - the alternative hypothesis, H_a or H_1, which is a statement of a range of alternative values in which the parameter may fall.

  H_0 is assumed to be true until there is evidence to suggest otherwise. Formally, we have to decide, on the basis of the information contained in the sample, if H_0 have to be rejected or not.

Introduction to Hypothesis Testing II

- **3. Test Statistic**. It is calculated under the assumption H_0 is true, and incorporates a measure of standard error and assumptions (conditions) related to the sampling distribution. In practice, tt describes (measures) how far the point estimate falls from the parameter value given in H_0.
- **4. p-value**. It is found by using the test statistic to calculate the probability of the sample data producing such a test statistic or one more extreme. The **rejection region** is found by using alpha to find a critical value; the rejection region is the area that is more extreme than the critical value. The **smaller** the p-value, the **stronger** the **evidence** the data provide against H_0. That is, a small p-value indicates a small likelihood of observing the sampled results if H_0 were true.

Introduction to Hypothesis Testing III



Introduction to Hypothesis Testing IV

- **5. Conclusion**. Based on the p-value, make a decision about H_0, we decide to either reject H_0 or decide to fail to reject H_0. *Notice we do not make a decision where we will accept H_0.* Before seeing the sample data, we decide how small the p-value would need to be to reject H_0. This cutoff point is called the **significance level**. In practice, the most common significance level is 0.05 (or 0.01 if the sample size is very large).

$$\text{If } p\text{-value} < \alpha \Rightarrow \text{REJECT } H_0$$

  When we reject H_0 we say the results are statistically significant. Following this decision, we want to summarize our results into an overall conclusion for our test.

Hypotheses

Hypothesis tests are about a parameter value : π for proportion or μ for mean. There are **three types** of H_1:

1. The population parameter is **not equal** to a certain value. Referred to as a "**two-sided test**".

$$H_1 : \pi \neq \pi_0 \quad H_1 : \mu \neq \mu_0$$

2. The population parameter is **less than** a certain value. Referred to as a "**left-tailed test**".

$$H_1 : \pi \leq \pi_0 \quad H_1 : \mu \leq \mu_0$$

3. The population parameter is **greater than** a certain value. Referred to as a "**right-tailed test**".

$$H_1 : \pi \geq \pi_0 \quad H_1 : \mu \geq \mu_0$$

For all three alternatives, the null hypothesis H_0 is the population parameter is equal to that certain value.

$$H_0 : \pi = \pi_0 \quad H_1 : \mu = \mu_0$$

Examples

- "Are the majority of students at Tor Vergata from Rome?" This example is about a population proportion (π). Here the value π_0 is 0.5 since more than 0.5 constitute a majority. The hypthoses set up would be a **right-tailed test**:

$$H_0 : \pi = 0.5 \text{ vs. } H_1 : \pi > 0.5$$

- A tire company claims that the mean lifetime is more than or equal to 42.000 miles. A consumer test agency wants to see whether the mean lifetime of a particular brand of tires is less than 42.000 miles. This example is about a population mean (μ). Here the value of μ_0 is 42.000. The hypthoses set up would be a **left-tailed test**:

$$H_0 : \mu = 42.000 \text{ vs. } H_1 : \mu < 42.000$$

Examples

- The length of a certain lumber from a national home building store is supposed to be 8.5 feet. A builder wants to check whether the shipment of lumber she receives has a mean length different from 8.5 feet. This example is about a population mean (μ). Here the value of μ_0 is 8.5. The hypthoses set up would be a **two-tailed test**:

$$H_0 : \mu = 8.5 \text{ vs. } H_1 : \mu \neq 0.5$$

- A political news company believes the national approval rating for the current president has fallen below 40%. This example is about a population proportion (π). Here the value of π is 0.4. The hypthoses set up would be a **a left-tailed test:**

$$H_0 : \pi = 0.4 \text{ vs. } H_1 : \pi < 0.4$$

Test Statistic - Proportion

- **Assumptions**
  1. The variable is categorical in 0-1 coding.
  2. The sample size is sufficiently large that the sampling distribution of the sample proportion is approximately normal [$n\pi_0$ and $n(1-\pi_0)$ are at least 15].
- **Test Statistic**

$$Z_{obs} = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)}}\sqrt{n}$$

It has a *standard normal distribution* when the null is true.

Example

Referring back to the first example above, say we take a random sample of 500 Tor Vergata students and find that 278 are from Rome. Can we conclude that the proportion is larger than 0.5?

Is $278/500 = 0.556$ much bigger than 0.5? What is much bigger? This depends on the standard deviation of $\hat{\pi}$ under the null hypothesis.

$$\hat{\pi} - \pi_0 = 0.556 - 0.5 = 0.056$$

The standard deviation of $\hat{\pi}$, if the null hypotheses is true (e.g. when $\pi_0 = 0.5$) is:

$$\sqrt{\frac{\pi_0(1-\pi_0)}{n}} = \sqrt{\frac{0.5(1-0.5)}{500}}$$

It follows that

$$Z_{obs} = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)}}\sqrt{n} = \frac{0.556 - 0.5}{\sqrt{0.5(1-0.5)}}\sqrt{500} = 2.504$$

How do we determine whether to reject the null hypothesis? It depends on the level of significance α, and the probability the sample data would produce the observed result.

Test Statistic - Mean

- **Assumptions**
  1. The variable is quantitative.
  2. The population distribution is approximately normal.
  3. The population distribution is not approximately normal, but the sample size is sufficiently large that the sampling distribution of the sample mean is approximately normal.
- **Test Statistic**

$$T_{obs} = \frac{\bar{x} - \mu_0}{S}\sqrt{n}$$

Under the null it is distributed as a *Student t* with $n-1$ degree of freedom (if assumption 2. is true). If n is *large enough*, it has a *standard normal distribution* even if the population is not approximately normal.

Example

In the lumber example above, the mean length of the lumber is supposed to be 8.5 feet. A builder wants to check whether the shipment of lumber she receives has a mean length different from 8.5 feet. If the builder observes that the sample mean of 61 pieces of lumber is 8.3 feet with a sample standard deviation of 1.2 feet. What will she conclude?

Is 8.3 very different from 8.5? This depends on the standard deviation of $\bar{X}$:

$$T_{obs} = \frac{\bar{x} - \mu_0}{S}\sqrt{n} = \frac{8.3 - 8.5}{1.2}\sqrt{61} = -1.3$$

How do we determine whether to reject the null hypothesis? It depends on the level of significance α, and the probability the sample data would produce the observed result.

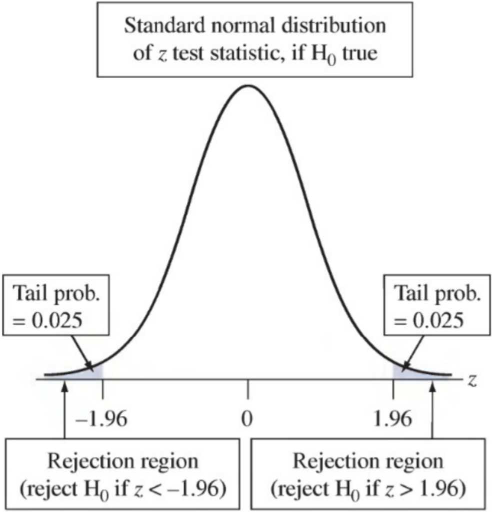
Type I and Type II Errors

How do we determine whether to reject the null hypothesis? It begins with the **level of significance** α, which **is the probability of the Type I error.** Due to the *sampling variability*, when conducting hypothesis testing *two types of mistakes* may be made:

	H_0 true	H_0 false
Accept H_0	Correct Decision	Type error II
Reject H_0	Type error I	Correct Decision

Level of Significance α

The significance level α is the probability of the Type I error



Optimal Test

The two error probabilities are closely (inversely) related. It follows that it is not possible to minimize both simultaneously.

How do we choose the best test?

There are two approaches:

- the best test **minimizes a weighted mean of the two error probabilities**
- **Neyman-Pearson approach**: given an upper bound for the probability of a type I error, the best test minimizes the probability of a type II error for each possible value of the parameter in the region of the alternative hypothesis

We follow the second approach

Decision Making in Hypothesis Testing

There are **two methods for making a statistical decision**

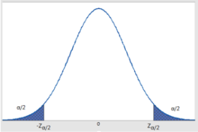
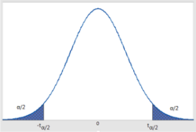
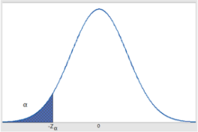
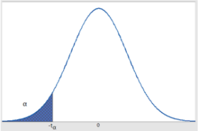
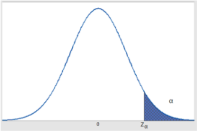
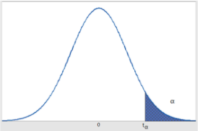
- **rejection region approach**
- **p-value (or probability value) approach**

Glossary

- **Test statistic**. The sample statistic one uses to either reject H_0 or not to reject H_0.
- **Critical values**. The values of the test statistic that separate the rejection and non-rejection regions.
- **Rejection region**. The set of values for the test statistic that leads to rejection of H_0.
- **Non-rejection region**. The set of values not in the rejection region that leads to non-rejection of H_0.
- **p-value**. The p-value (or probability value) is the probability that the test statistic equals the observed value or a more extreme value under the assumption that the null hypthothesis is true.

The logic of hypothesis testing $\rightarrow$ reject H_0 if the sample data are not consistent H_0, i.e. if the observed test statistic is more extreme in the direction of the alternative hypothesis than one can tolerate. The critical values are the boundary values obtained corresponding to the preset α level.

Rejection Region Approach to Hypothesis Testing

One Proportion Z-test	One Mean t-test
 Two-Tailed Reject H_0 if $\lvert Z^* \rvert \geq Z_{\alpha/2}$	 Two-Tailed Reject H_0 if $\lvert t^* \rvert \geq t_{\alpha/2}$
 Left-Tailed Reject H_0 if $Z^* \leq Z_\alpha$	 Left-Tailed Reject H_0 if $t^* \leq t_\alpha$
 Right-Tailed Reject H_0 if $Z^* \geq Z_\alpha$	 Right-Tailed Reject H_0 if $t^* \geq t_\alpha$

p-value Approach to Hypothesis Testing

	Right-tailed	**Left-tailed**	**Two-tailed**
π case - p-value	$P(Z > Z^\star)$	$P(Z < Z^\star)$	$2 \times P(Z >\mid Z^\star \mid)$
μ case - p-value	$P(t > t^\star)$	$P(t < t^\star)$	$2 \times P(t >\mid t^\star \mid)$

Summary

Six Steps to Conducting a Statistical Test

1. The null and alternative hypotheses
2. Level of significance α
3. Test statistics
4. Compute the p-value
5. Check whether to reject the null hypothesis by comparing p-value to α
6. Conclusion in words

Example: Tor Vergata Students from Rome

We take a random sample of 500 Tor Vergata students and find that 278 are from Roma. Can we conclude that the proportion is larger than 0.5 at a 5% level of significance?

Rejection Region Approach

1. Can we use the z-test? The answer is yes since the hypothesized value π_0 is 0.5 and we can check that: $n\pi_0 = 500 \times 0.5 = 250 \geq 15$ and $n(1-\pi_0) = 500 \times (1-0.5) = 250 \geq 15$.
   Set up the hypotheses: $H_0 : \pi = 0.5$ vs. $H_1 : \pi > 0.5$
2. **Decide on the significance level,** α. $\rightarrow \alpha = 0.05$.
3. **Compute the value of the test statistic**.

$$Z^\star = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)}}\sqrt{n} = \frac{0.556 - 0.5}{\sqrt{0.5(1-0.5)}}\sqrt{500} = 2.504$$

4. **Find the appropriate critical values for the test using the z-table**. The critical value is 1.645 $\Rightarrow$ the rejection region is given by: $Z^\star < 1.645$
5. **Check whether the value of the test statistic falls in the rejection region**. The observed Z-value is 2.504 - this is our test statistic. Since $Z^\star$ falls within the rejection region, we reject H_0.
6. **State the conclusion in words**. With a test statistic of 2.504 and critical value of 1.645 at a 5% level of significance, we have enough statistical evidence to reject the null hypothesis. We conclude that a majority of the students are from Rome.

Example: Tor Vergata Students from Rome

We take a random sample of 500 Tor Vergata students and find that 278 are from Roma. Can we conclude that the proportion is larger than 0.5 at a 5% level of significance?

p-value Approach

1. Can we use the z-test? The answer is yes since the hypothesized value π_0 is 0.5 and we can check that: $n\pi_0 = 500 \times 0.5 = 250 \geq 15$ and $n(1-\pi_0) = 500 \times (1-0.5) = 250 \geq 15$.
   Set up the hypotheses: $H_0 : \pi = 0.5$ vs. $H_1 : \pi > 0.5$
2. **Decide on the significance level,** α. $\rightarrow \alpha = 0.05$.
3. **Compute the value of the test statistic**.

$$Z^\star = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)}}\sqrt{n} = \frac{0.556 - 0.5}{\sqrt{0.5(1-0.5)}}\sqrt{500} = 2.504$$

4. **Compute the appropriate p-value based on our alternative hypothesis**. The p-value is $P(Z > Z^\star) = P(Z > 2.504) = 0.0062$.
5. **Check whether the value of p-value is less than** α. Since p-value $= 0.0062 < 0.05$ (the α value), we reject the null hypothesis.
6. **State the conclusion in words**. With a test statistic of 2.504 and p-value of 0.0062, we reject the null hypothesis at a 5% level of significance. We conclude that a majority of the students are from Rome.

Example: Length of Lumber

The mean length of the lumber is supposed to be 8.5 feet. Check whether the shipment of lumber has a mean length different from 8.5 feet by conducting a test at a 1% level of significance. Recall that: $\bar{x} = 8.3$, $n = 61$ and $s = 1.2$.

Rejection Region Approach

1. Can we use the t-test? The answer is yes since the sample size of 61 is sufficiently large.Set up the hypotheses: $H_0 : \mu = 8.5$ vs. $H_1 : \mu \neq 8.5$
2. **Decide on the significance level,** α. $\rightarrow \alpha = 0.01$.
3. **Compute the value of the test statistic**.

$$t^{\star} = \frac{\bar{x} - \mu_0}{S}\sqrt{n} = \frac{8.3 - 8.5}{1.2}\sqrt{61} = -1.3$$

4. **Find the appropriate critical values for the test using the t-table**. The critical value is 2.660 (d.f.=60) $\Rightarrow$ the rejection region is given by: $t^{\star} < -2.660$ and $t^{\star} > 2.660$
5. **Check whether the value of the test statistic falls in the rejection region**. $t^{\star}$ does not fall within the rejection region: we fail to reject H_0.
6. **State the conclusion in words**. With a test statistic of -1.3 and critical value of ± 2.660 at a 1% level of significance, we do not have enough statistical evidence to reject the null hypothesis. We conclude that there is not enough statistical evidence that indicates that the mean length of lumber differs from 8.5 feet.

Example: Length of Lumber

The mean length of the lumber is supposed to be 8.5 feet. Check whether the shipment of lumber has a mean length different from 8.5 feet by conducting a test at a 1% level of significance. Recall that: $\bar{x} = 8.3$, $n = 61$ and $s = 1.2$.

p-value Approach

1. Can we use the t-test? The answer is yes since the sample size of 61 is sufficiently large.Set up the hypotheses: $H_0 : \mu = 8.5$ vs. $H_1 : \mu \neq 8.5$
2. **Decide on the significance level,** α. $\rightarrow \alpha = 0.01$.
3. **Compute the value of the test statistic**.

$$t^{\star} = \frac{\bar{x} - \mu_0}{S}\sqrt{n} = \frac{8.3 - 8.5}{1.2}\sqrt{61} = -1.3$$

4. **Compute the appropriate p-value based on our alternative hypothesis**. The p-value is $2 \times P(t >| t^{\star} |) = P(t > 1.3) = 0.1986$.
5. **Check whether the value of p-value is less than** α. The p-value exceeds our 1% level of significance for the test $\rightarrow$ we fail to reject the null hypothesis.
6. **State the conclusion in words**. With a test statistic of - 1.3 and p-value of 0.2, we fail to reject the null hypothesis at a 1% level of significance since the p-value would exceed our significance level. We conclude that there is not enough statistical evidence that indicates that the mean length of lumber differs from 8.5 feet.