Module 10
Hypothesis Testing

Module Learning Outcomes

- Get familiarized with the symbols.
- Define a proper parameter and set up the null hypothesis and alternative hypothesis for both means and proportions.
- Calculate test statistics for both means and proportions.
- Find p-values (one-tailed and two-tailed) for both means and proportions.
- Provide an interpretation of the p-value.
- Apply the Central Limit Theorem, as needed.
- Make a proper conclusion using a level of significance.
- Understand the meaning of Type I error and Type II error, in the context.

Assumption and Review of Notation/Symbols

- One main assumption or condition in **statistical inference** is that parameters are always unknown to us.
- Here are some key notations or symbols used in this module.

Symbols	Name	Description	
μ	Population Average/Mean	The average/mean of the population	Parameter
$\bar{X}$	Sample Average/Mean	The average/mean of a sample	Statistic

σ	Population Standard Deviation	The standard deviation of the population	Parameter
s	Sample Standard Deviation	The standard deviation of a sample	Statistic
p	Population Proportion	The proportion of the population who have certain characteristic	Parameter
$\bar{p}$	Sample Proportion	The proportion of a sample who have certain characteristic	Statistic

- Note: Statistics are used to estimate unknown parameters.
- MLO: Get familiarized with the symbols.

Interval Estimation vs. Hypothesis Testing

- As the name implies, interval estimation is about estimating parameters using statistics.
- Hypothesis testing, in contrast, is to verify (or prove) if a claim is true.
- The claim is written in a statement (either in words or in symbols) in the alternative hypothesis.
- Therefore, the entire procedure of hypothesis testing is to see if we could prove its validity.

10.1 Alternative Hypothesis – Content

- The **alternative hypothesis** is a statement or sentence, and it is usually shorthanded as H_a.
- There are three things we need to have in the alternative hypothesis: (1) the parameter of interest, (2) an inequality sign, and (3) a hypothesized value.
- In most cases, parameters are expressed in the form of a symbol (like p). In other cases, you could write out the parameter in English words (i.e. no symbols).

- There are three common **inequality signs** we could use. They are listed below, together with their proper terminology.

Inequality Sign	Test	Hints / Keywords
$<$	Lower-tailed test (or One-tailed test)	Decrease, Less, Lower, Smaller, Has gone down, Worse, Faster (faster means shorter time)
$>$	Upper-tailed test (or One-tailed test)	Increase, More, Higher, Bigger/Larger, Has gone up, Better, Slower (slower means longer time)
$\neq$	Two-tailed test	Difference/Different, Same, Identical

- **Lower-tailed tests** focus on the left-hand side of the distribution (more details in the p-value section); **upper tailed tests** focus on the right-hand side of the distribution. They are collectively called the **one-tailed tests** because, well, they only deal with one side.
- **Two-tailed tests** focus on both sides of the distribution.
- As you will see in examples, the questions or the situations will give you enough hints of which one to use. Some of the hints or keywords are listed in the last column of the above table.
- The **hypothesized value** (usually denoted as μ_0 for numerical variable or p_0 for categorical variable) is simply a value we postulate or believe. Or it is directly coming from the claim. The value will be provided in the situation.

10.2 Alternative Hypothesis – Role/Purpose

- Besides the name **alternative hypothesis**, it is often referred to as the **research hypothesis** in many disciplines.
- It is because this statement represents what a researcher wants to prove in her study.
- Proving a statement is very hard because it requires the checking of all scenarios.
- For example, proving the statement "All dogs have four legs" is almost impossible because you would have to make sure that **every single dog** in the world has four legs.
- However, refuting a statement (or proving a statement is false) is much easier!
- For example, disproving the statement "All dogs have four legs" is comparatively easier because all you need is to find one counterexample that a dog does not have four legs.
- Back to statistics, as you can imagine, proving a claim (or proving some assumption about the population) is impossible.
- It is because it requires us to check "every single subject in the population".
- Instead, we would try to refute or disprove a statement that has the opposite meaning, i.e. we would want to reject the null hypothesis.
- *Note to Self #1*: Make sure we understand why we can only say "rejecting or not rejecting the null hypothesis" but not "accepting or not accepting the alternative hypothesis"?
- *Note to Self #2*: Make sure we understand why the null hypothesis and the alternative hypothesis are always opposite?

10.3 Null Hypothesis – Role/Purpose

- From the above, it seems that the existence of the null hypothesis is mostly ceremonial, i.e. it would be nice if it is there, but it is not necessary!
- However, the null hypothesis has its use and purpose in the entire process of hypothesis testing.
- 1) The null hypothesis is assumed to be true during the entire process of hypothesis testing, until proven otherwise at the end.
- The idea is very similar to a suspect/accused (of a serious crime) is presumed innocent throughout the entire criminal trial, until he is proven guilty. Google "Presumption of Innocence" to learn more about it.
- Without this assumption, we will be stuck half-way in the procedure (more later).
- 2) The rejection of the null hypothesis can only be achieved with overwhelming evidence, or "beyond reasonable doubt".
- If reasonable doubt remains, the null hypothesis should "get the benefit of the doubt" and not be rejected.
- Other interpretations of it include "null hypothesis is the status quo", "null hypothesis represents the current condition", or "null hypothesis says that whatever happens is due to sampling error (also called the random error)".
- Recall that what we want to do is to prove the validity of the alternative hypothesis; but we cannot do it until we are able to access the entire population of interest, which is impossible.
- Tactically, if we could refute the null hypothesis (by rejecting it), then we basically have shown that the alternative hypothesis is **more or less** true!
- Note: Although the alternative hypothesis is to be thought of first, the null hypothesis is always written first because it represents what you believe (or assume).

10.4 Null Hypothesis – Content

- Like the alternative hypothesis, the **null hypothesis** is also a statement or sentence, and it is usually shorthanded as H_0.
- The null hypothesis contains the same three things: (1) the parameter of interest, (2) an inequality sign, and (3) a hypothesized value.
- Because it has the opposite meaning to the alternative hypothesis, the only difference is the use of the "opposite" inequality signs.

Means or Numerical Variables

- All three pairs of complete hypotheses are listed below. I arbitrary choose the hypothesized value of $\mu_0 = 10$ (for numerical variables).

	Null Hypothesis	**Alternative Hypothesis**
Lower-tailed test	$H_0: \mu \geq 10$	$H_a: \mu < 10$
Upper-tailed test	$H_0: \mu \leq 10$	$H_a: \mu > 10$
Two-tailed test	$H_0: \mu = 10$	$H_a: \mu \neq 10$

Proportions or Categorical Variables

- All three pairs of complete hypotheses are listed below. I arbitrary choose the hypothesized value of p_0=0.20.

	Null Hypothesis	**Alternative Hypothesis**
Lower-tailed test	$H_0: p \geq 0.20$	$H_a: p < 0.20$
Upper-tailed test	$H_0: p \leq 0.20$	$H_a: p > 0.20$
Two-tailed test	$H_0: p = 0.20$	$H_a: p \neq 0.20$

- *Note to self #3*: Make sure we know that the equality sign or "=" always appears in the null hypothesis. This is like having the idea of *status quo* or the *benefit of the doubt*.
- *Note to self #4*: Make sure we understand that the direction of the alternative hypothesis always agrees with the name of the test.
- MLO: Define a proper parameter and set up the null hypothesis and alternative hypothesis for both means and proportions.

10.5 Significance Level (or Level of Significance)

- In a proper procedure of hypothesis testing, this is the time to think about what value of **significance level** (denoted as α) we want to use.
- In theory, it could be any value between 0 and 1, but it is generally less than 10% or 0.10.
- Common values are 1% (or 0.01), 5% (or 0.05) and 10% (or 0.10).
- Smaller value of α means the null hypothesis is harder to be rejected.
- In other words, if we do not want H_0 to be rejected that easily, we will use a smaller α.

α	10%	5%	1%
How hard to reject null hypothesis?	Easiest	Medium / No preference	Hardest

- *Note to self #5*: Make sure we know that a 5% significance level would be used, when we are not sure what value to use or the question never mentions its value.
- More about its use and definition (related to Type I Error) later.

Data Collection

- In the proper procedure of hypothesis testing, this is the time we go out and collect the data using the Simple Random Sampling Method.
- In other words, our data comes from a random sample and it is representative of the population.

10.6 Test Statistic Calculations

Means or Numerical Variables

- Using the equation of the sampling distribution of $\bar{X}$, we can standardize the sample mean into a Z-score, i.e.

$$Z = \frac{value - mean}{SD} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

- However, both μ and σ are unknown because they are both parameters.
- The unknown parameter (μ) will take the value of the hypothesized value (μ_0) because the null hypothesis is assumed to be true. Otherwise, we would be stuck and cannot proceed.
- In addition, we can always use the statistic (s) to estimate the unknown parameter (σ).
- Hence, we have

$$T.S. = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

Proportions or Categorical Variables

- Using the equation of the sampling distribution of $\bar{p}$, we can standardize the given sample proportion into a Z-score, i.e.:

$$Z = \frac{value - mean}{SD} = \frac{\bar{p} - p}{\sqrt{p(1-p)/n}}$$

- The unknown parameter (p) will take the value of the hypothesized value (p_0) because the null hypothesis is assumed to be true. Otherwise, we would be stuck and cannot proceed.
- Hence, the test statistic becomes:

$$T.S. = \frac{\bar{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

- Calculate test statistics for both means and proportions.

10.7 Finding p-value

- The **p-value** is defined as the probability that the observed statistic, or some more extreme values, will happen again, provided that the null hypothesis is assumed to be true.
- Note that we are again assuming that the null hypothesis is true here.

- The interpretation of the p-value is the same as the interpretation of the probabilities that we calculated in Sampling Distribution module, plus the meaning of the null hypothesis is assumed to be true.

A. Means or Numerical Variables

- To find the p-value, we just need to find the "tail probability" under the Z distribution.
- Which tail? It depends on the "direction" of the alternative hypothesis.

	Alternative Hypothesis	**p-value**
Lower-tailed test	$H_a\!: \mu < 10$	Area to the *left* of T.S.
Upper-tailed test	$H_a\!: \mu > 10$	Area to the *right* of T.S.
Two-tailed test	$H_a\!: \mu \neq 10$	Both sides

- Here, the p-value could be exact or approximate, depending on the situation.
- When the variable of interest (X) has a Normal distribution, then the p-value is **exact**.
- When the variable of interest (X) does not have a Normal distribution or known to have a skewed distribution, then the p-value is a good **approximation** when the **sample size is at least 30**, i.e. $n \geq 30$, according to the Central Limit Theorem (or CLT).

B. Proportions or Categorical Variables

- Again, it depends on the side alternative hypothesis points to.

	Alternative Hypothesis	**p-value**
Lower-tailed test	$H_a\!: p < 0.20$	Area to the *left* of T.S.
Upper-tailed test	$H_a\!: p > 0.20$	Area to the *right* of T.S.
Two-tailed test	$H_a\!: p \neq 0.20$	More of that later

- Keep in mind that the p-value is always an approximation, but the approximation is good enough when:

  1) $\boldsymbol{n \geq 30}$, 2) $\boldsymbol{n \times p_0 \geq 5}$, and 3) $\boldsymbol{n \times (1 - p_0) \geq 5}$

- Note that p_0 (the hypothesized value) is used instead of p (the parameter) in these three conditions.
- MLO: Find p-values for both means and proportions.
- MLO: Provide an interpretation of the p-value.
- MLO: Apply the Central Limit Theorem, as needed.

C. Understanding p-values

- What do we do with p-values? What do they mean?
- Let us think about it heuristically.
- If we knew the population proportion (p) has a value of 0.20 and we got a sample proportion ($\bar{p}$) of 0.21, then what we observed (from the sample) is rather similar to what we believe (in the population). Hence, we probably do not abandon our original belief (in H_0) by not rejecting the null hypothesis.
- On the other hand, if we knew the population proportion (p) has a value of 0.20 and we got a sample proportion ($\bar{p}$) of 0.50, then what we observed (from the sample) is rather

different from what we believe (in the population). Hence, we probably will abandon our original belief (in H_0) by rejecting the null hypothesis.

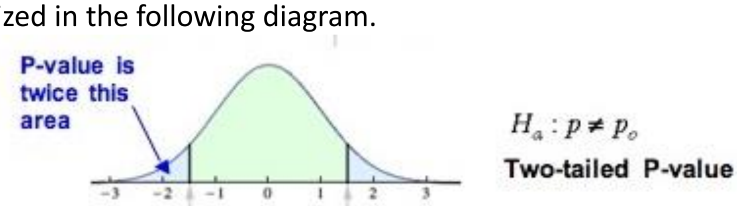
- The following table shows how p-value can be used to make conclusions. Focus on the last two columns.

Situation	$\lvert\bar{p} - p_0\rvert$	$T.S. = \frac{\lvert\bar{p} - p_0\rvert}{SE}$	p-value	Reject H_0?
$\bar{p}$ close to p_0	Small	Small	Large	No
$\bar{p}$ different from p_0	Large	Large	Small	Yes

- So, that is easy! Small value of p-value leads to the rejection of the null hypothesis, while large value of p-value would not reject the null hypothesis.
- But how large is large enough and how small is small enough?
- We will need a pre-determined "standard" to compare to and it is the significance level (α).

D. Two-tailed p-values

- Finding the p-value of two-tailed tests is simply multiplying the one-tailed probability by 2.
- In short, if the corresponding p-value is bigger than one (or 100%), you have got the wrong tail and please use the other tail!
- It can be visualized in the following diagram.



- Conclusion can then be drawn similarly using the p-value, in comparison with the level of significance.
- MLO: Find two-tailed p-values.

10.8 Conclusion

- Okay! I know that when p-value is smaller than the significance level, then I will have to reject the null hypothesis. Should I just say that the null hypothesis is wrong?
- Not quite! Recall one of the main purposes of null hypothesis is to serve as a "scapegoat" – to be sacrificed (or to be rejected) so that we could sort of prove the alternative hypothesis.
- But let us not forget that we have not literally "proven the alternative hypothesis directly". We merely are able to reject the null hypothesis, which means null hypothesis is more or less wrong.
- And because the null hypothesis is more or less wrong, the opposing statement alternative hypothesis is then more or less correct, or more or less proven.
- Of course, we will not use "more or less" in any scientific writing as it does not sound very confident. Instead, we use a significantly better word – **significance**!

- In other words, we have the following three components in the conclusion.

<table>
<tr><td rowspan="2">1)</td><td>Since p-value $< \alpha$,</td><td>Since p-value $\geq \alpha$,</td></tr>
<tr><td colspan="2">Explanation: This piece is to tell readers what significance you have used. Of course, you might say “are they blind?” Haven’t we provided the value at the beginning! Yes, you are correct. But mind you that most people will not read the entire research paper or procedure. In fact, most researchers only read the conclusion or summary of other’s research paper. That is why!</td></tr>
<tr><td rowspan="2">2)</td><td>we have enough statistical evidence to reject the null hypothesis …</td><td>we do not have enough statistical evidence to reject the null hypothesis …</td></tr>
<tr><td colspan="2">Explanation: This piece is to tell readers what immediate conclusion about the null hypothesis. At first sight, it seems unnecessary because we will have a proper conclusion coming up. True! But then again, please understand that most searchers will stop reading here if they have a pretty good idea of what the study is about.</td></tr>
<tr><td rowspan="2">3)</td><td>… and conclude that the alternative hypothesis is <u>significantly</u> true.
Note: Make sure you write H_a out in words.</td><td>… and conclude that the alternative hypothesis is not <u>significantly</u> true.
Note: Make sure you write H_a out in words.</td></tr>
<tr><td colspan="2">Explanation: This piece is to lay out the complete conclusion, in layperson’s terms. And as you can imagine, this is for general public (or non-researchers) to read.</td></tr>
</table>

- *Note to self #6*: Make sure we have all three components in the conclusion. These three cascading layers of conclusion are for three different groups of audience or readers.
- *Note to self #7*: Make sure we know that the idea of “more or less” has to be included in the conclusion because of sampling error. We are only getting one sample. Therefore, no matter how well the sample was drawn, it still has some chance of not representing the population well.
- *Note to self #8*: Make sure you know why the “equality sign” always goes to the Right Hand Side of the conclusion (above) where null hypothesis is not rejected. It is because the rejection of the null hypothesis can only be achieved with overwhelming evidence, or “beyond reasonable doubt”. If reasonable doubt remains, the null hypothesis should not be rejected. That is why!
- MLO: Make a proper conclusion using a level of significance.

10.9 Type I Error and Type II Error

- Recall that the null hypothesis is always assumed to be true, until proven otherwise.
- However, we are never sure if the assumption is correct or not.
- Therefore, there are two scenarios: (1) H_0 is true or (2) H_0 is not true.
  Note: We do not use the word “false” because the two words “true” and “false” are not really opposite to each other.

- Also recall that the conclusion is always made based on a single random sample from the population.
- Because of the sampling error, different samples could lead to one of the two possible conclusions: (1) H_0 is rejected and (2) H_0 is not rejected.
- Therefore, we have four combinations – two of them are good and two of them are bad.

		Truth	
		H_0 is true	**H_0 is not true**
Conclusion	**Reject H_0**	Type I Error	
	Do not reject H_0		Type II Error

- **Type I error** is defined as "H_0 is rejected when H_0 is in fact true".
- **Type II error** is defined as "H_0 is not rejected when H_0 is not true".
  Note: Please pay special attention to the choice of words here. It is rather awkward, but it is the proper way to do so.
- Keep in mind that they are both errors but we are not sure when and if they will happen. And if so, how serious. So, we will keep them in check by controlling their happening.
- Therefore, we define
  α = P(committing type I error) and **β = P(committing type II error)**
- Since they are both probabilities of committing error, we would like to keep them both to minimal.
- But the fact is that when one is kept to a small value, the other will go up, and vice versa.
- So, we could only control one error to a relatively "small" value, like under 10%.
- *Note to self #9*: Make sure you understand why type I error is generally considered to be a more serious error.
- Therefore, the value of α is kept to a value less than 10% and the value of β will not be controlled.
- Smaller value of α means we have a smaller type I error or less likely a true null hypothesis would be rejected.
- MLO: Understand the meaning of Type I error and Type II error, in the context.

Applications of Hypothesis Testing

- The concept of hypothesis testing is used well beyond the scope of this course.
- From a business/marketing point of view, how does a new and improved product fare when it is compared to a similar but older product that had a 25% market share?
- From a health point of view, should the FDA approve a new drug if it has 80% recovery rates from patients?
- From sports point of view, should my hockey team go for "five-hole" at the opposing goalie in shoot-outs because he yields 45% of his goals through "five-hole"? FYI: Going for "five-hole" in hockey is to shoot the puck between the two legs of the goalie.

- From social media point of view, should Instagram (for example) pull up ads about runner shoes after a user has looked up runner shoes 3 times in the past week? – This is a Statistics problem.
- Or how many times after a user searches for runner shoes should Instagram (for example) pull up an ad about runner shoes – This is a Machine Learning problem.
- Also, we have seen a problem set example about one COVID-19 test having 2% false-positive rate. The "**false-positive**" has the same meaning as "Type I error" – reject the null hypothesis when it is true. And "**false-negative**" has the same meaning as "Type II error".