Introduction to Probability and Statistics Eleventh Edition

Chapter 9

Large-Sample Tests of Hypotheses

Example 1

- A homeowner randomly samples 64 homes similar to her own and finds that the average selling price is $252,000 with a standard deviation of $15,000. Is this sufficient evidence to conclude that the average selling price is greater than $250,000?

- This is the kind of problem we are interested in. We see how to solve it a little bit later.

Hypothesis Testing for Population Parameters

- Consider a statement about a population parameter that can be true or false. E.g.
  - "μ is equal to 10",
  - "p is at least 0.5".
- We want to get evidence that this statement is **false**.
- How can we do that?
- Assume the statement is true (this becomes what we call the **null hypothesis (H_0).**
- We call the negation of the null hypothesis the **alternative hypothesis (H_a).**
- Make an observation using a random sample.
- If this observation would be very unlikely if H_0 is true, we take this as evidence that it is false (i.e. H_a is true).

Types of Test

- There are two basic kinds of hypothesis test.
  - Two-tailed and one-tailed.
- The reason for the names will become clear soon.
- H_0 in a **two-tailed** test is that the parameter of interest is equal to some fixed value.
  - E.g. "$\mu = 10$".
- In a **one-tailed** test, H_0 is that the parameter is either at most, or at least, some fixed value.
  - E.g. "$p \geq 0.5$".
- One-tailed tests come in two kinds.
  - **Left-tailed**: $\theta \geq \theta_0$.
  - **Right-tailed**: $\theta \leq \theta_0$.

θ is parameter of interest, θ_0 is fixed value.

Hypothesis tests for population means

E.g. We want to argue that true value μ of the population mean is bigger than μ_0?

$H_a = \mu > \mu_0$, $H_0 = \mu \leq \mu_0$.

This is a right tailed test.

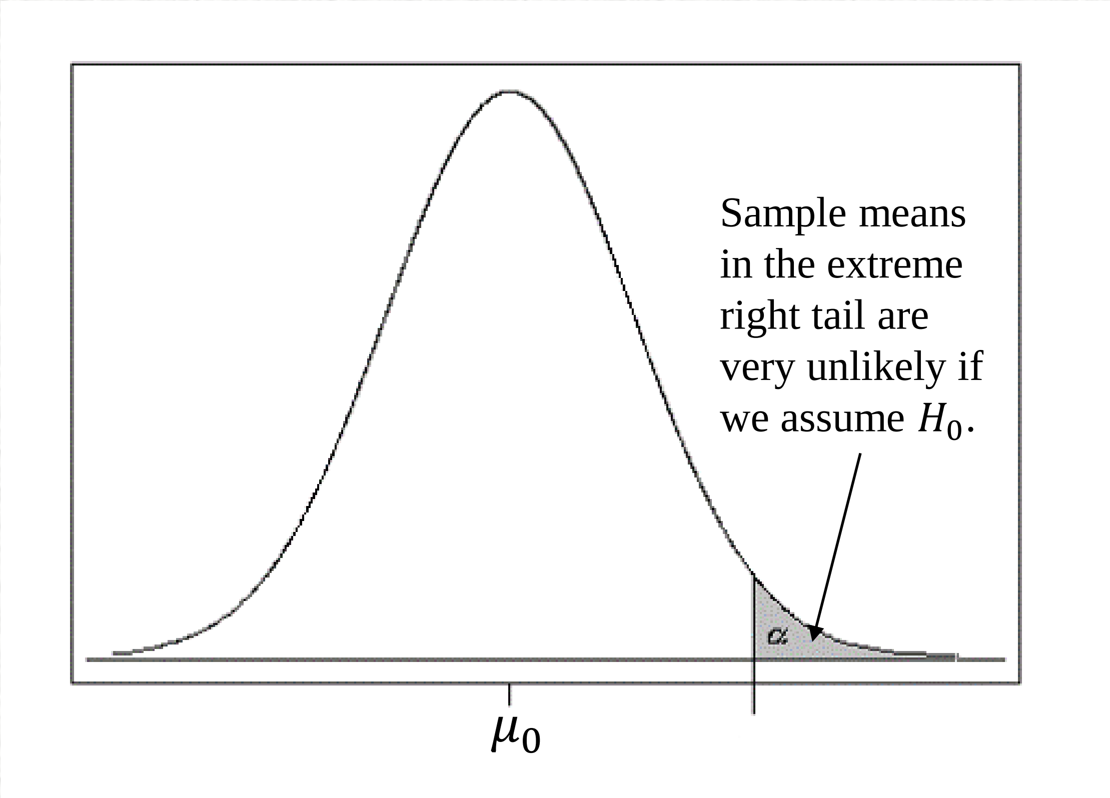
Take a random sample and calculate the sample mean $\bar{x}$.

Is the value of $\bar{x}$ consistent with the hypothesis that $\mu \leq \mu_0$?

If our value of $\bar{x}$ should be very rare if we assume $\mu \leq \mu_0$, then this is evidence against the null hypothesis.

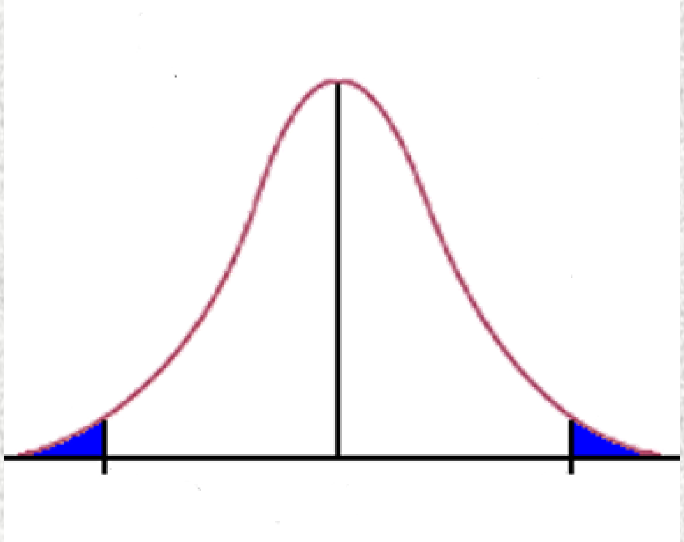
If the sample size is big enough, the CLT says the sampling distribution is normal, and we can use this to calculate probabilities.

CLT says that for large sample sizes the distribution of sample means is normal.



Two-Tailed Tests

- In the right-tailed test on the previous slide, only sample means to the right of the sampling distribution are incompatible with H_0.
  - Because sample means less than μ_0 are obviously consistent with $\mu \leq \mu_0$.



- In a two-tailed test, extreme values to the right and the left are incompatible with H_0.
  - Because here H_0 is that $\mu = \mu_0$.

Example 1 (revisited)

- A homeowner randomly samples 64 homes similar to her own and finds that the average selling price is \$252,000 with a standard deviation of \$15,000. Is this sufficient evidence to conclude that the average selling price is greater than \$250,000? Use $\alpha = .01$.

Null hypothesis

$$H_0 : \mu \leq 250{,}000$$
$$H_a : \mu > 250{,}000$$

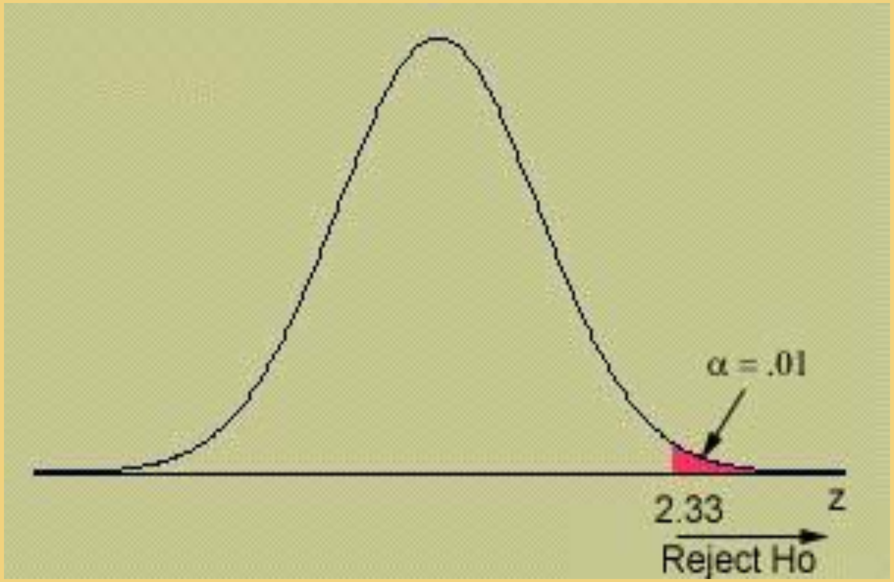
Alternative hypothesis

Test Statistic:

$$z \approx \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{252{,}000 - 250{,}000}{15{,}000/\sqrt{64}} = 1.07$$

Critical Value Approach

What is the critical value of z that cuts off exactly $\alpha = .01$ in the right-tail of the z distribution?



For our example, $z = 1.07$ does not fall in the rejection region and H_0 is not rejected.

There is not enough evidence to indicate that μ is greater than $250,000.

Rejection Region: Reject H_0 if $z > 2.33$.

p-values

- A p-value is a probability calculated from the test statistic that, speaking informally, measures whether the test statistic is likely or unlikely, assuming H_0 is true.

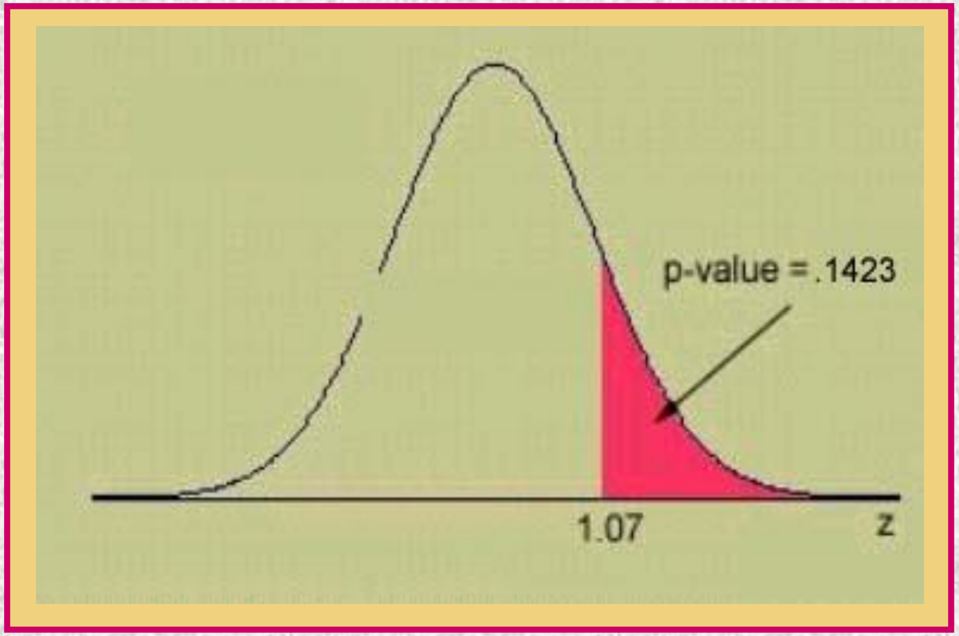
- Formally it is the probability that, assuming H_0 is true, we would observe in a random sample an effect as strong or stronger than the one observed in the sample we are looking at.

- E.g. "If there's no difference in the means of two populations, how likely would it be for the means of two random samples to be this far apart?"

p-Value Approach

- The probability that our sample results or something even more unlikely would have occurred *just by chance*, when $\mu = 250{,}000$.

$$p-\text{value} = P(Z > 1.07) = 1 - 0.8577 = 0.1423$$



Since *p*-value > $\alpha = 0.01$, H_0 is not rejected.

There is insufficient evidence to indicate that μ is greater than $250,000.

Example 2

The daily yield for a chemical plant has averaged 880 tons for several years. The quality control manager wants to know if this average has changed. She randomly selects 50 days and records an average yield of 871 tons with a standard deviation of 21 tons. At the 1% significance level what would she conclude?

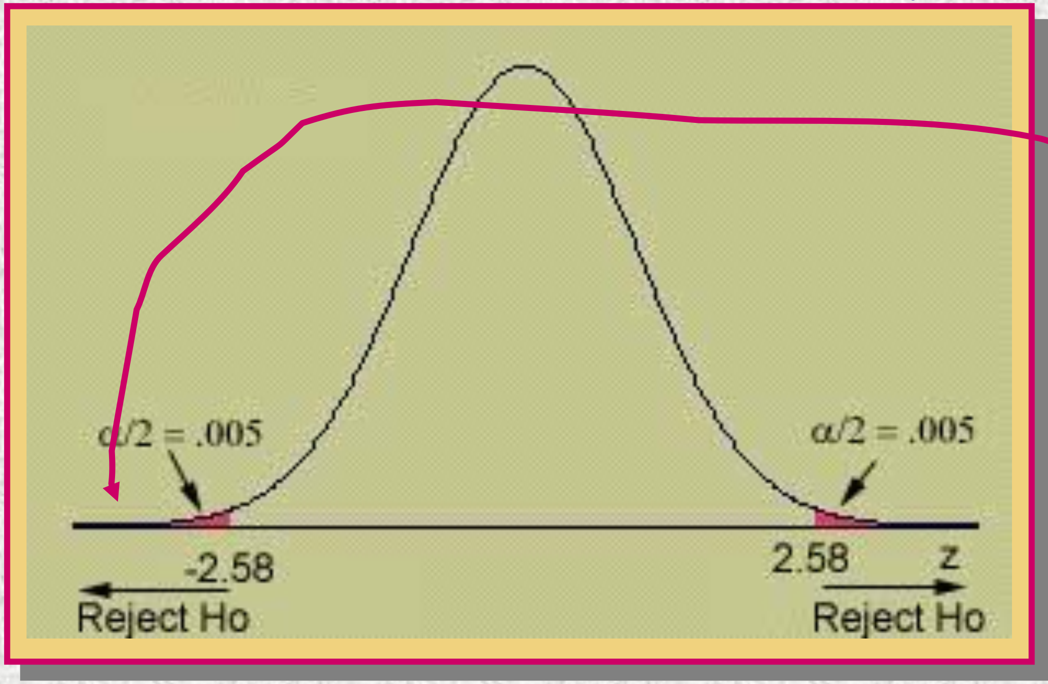
$$H_0 : \mu = 880$$
$$\text{vs. } H_a : \mu \neq 880$$

Test Statistic :

$$z \approx \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{871 - 880}{21/\sqrt{50}} = -3.03$$

Critical Value Approach

If $\alpha = 0.01$, what is the critical value of z that cuts off exactly $\alpha/2 = 0.01/2 = 0.005$ in the tail of the z distribution?



For our example,

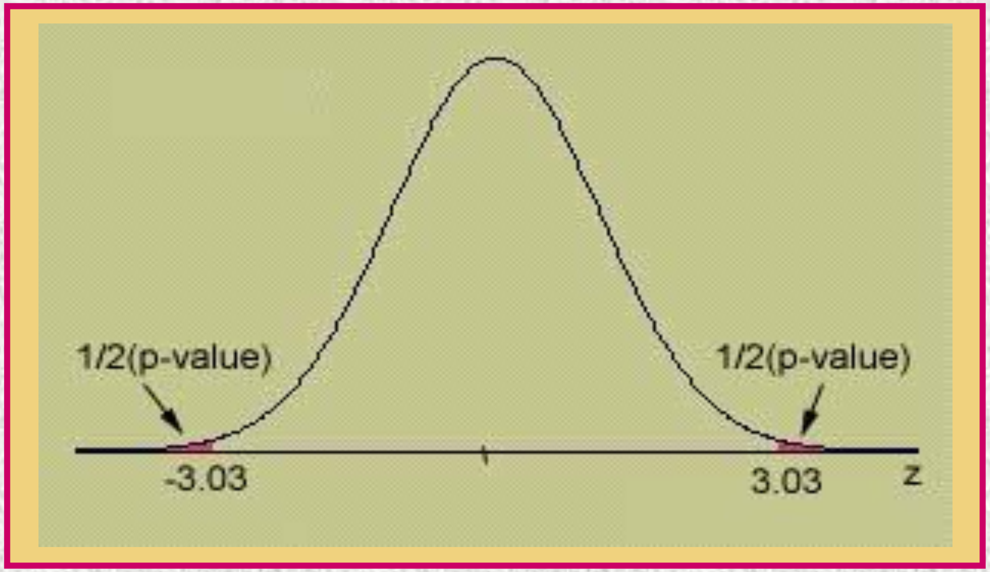
$z = -3.03$ falls in the rejection region and H_0 is rejected at the 1% significance level. Thus, the average yield is different from 880 tons.

Rejection Region: Reject H_0 if $z > 2.58$ or $z < -2.58$.

p –Value Approach

What is the probability that this test statistic or something even more extreme (far from what is expected if H_0 is true) could have happened *just by chance*?

$$p-\text{value} = P(Z<-3.03)+P(Z>3.03)$$
$$= 2P(Z<-3.03) = 2(0.0012) = 0.0024$$



Since our *p*-value = 0.0024 is less than α=0.01, we reject H_0 and conclude that the average yield has changed.

Class Activity 11

1. Find the appropriate rejection regions for the large-sample test statistic z in these cases:
   a) A left-tailed test at the 1% significant level.
   b) A two-tailed test with $\alpha = 0.01$.
   c) Suppose that the observed value of the test statistic was $z = -2.41$. For the rejection regions constructed in part a) and b), draw the appropriate conclusion for the tests. If appropriate, give a measure of the reliability of your conclusion.

Class Activity 11

2. A random sample of n = 35 observations from a quantitative population produced a mean $\bar{x}$ = 2.4 and a standard deviation s = 0.29. Suppose your research objective is to show that the population mean μ exceeds 2.3.

   a) Give the null and alternative hypotheses for the test.
   b) Locate the rejection region for the test using a 5% significance level.
   c) Find the standard error of the mean.
   d) Do the data provide sufficient evidence to indicate that μ > 2.3?
   e) Calculate the p-value for the test statistic you just calculated.
   f) Use the p-value to make a conclusion at the 5% significance level. Compare it with the conclusion from part d).

Class Activity 11

3. A random sample of 100 observations from a population produced a sample mean of 26.8 and a sample standard deviation of 6.5. Use the p-value approach to determine whether the population mean is different from 28. Explain your conclusion.

Class Activity 11

3. A random sample of 100 observations from a population produced a sample mean of 26.8 and a sample standard deviation of 6.5. Use the p-value approach to determine whether the population mean is different from 28. Explain your conclusion.

If the significance level (α) is not provided in the question, always use the p-value approach.

Class Activity 11

4. **Potency of an Antibiotic.** A drug manufacturer claimed that the mean potency of one of its antibiotics was 80%. A random sample of n = 100 capsules were tested and produced a sample mean of $\bar{x}$ = 79.7% with a standard deviation of s = 0.8%. Does the data present sufficient evidence to refute* the manufacturer's claim? Let α = 0.05.

*Refute means 'prove to be false'.

Class Activity 11

5. **Hamburger Meat.** Each package of ground beef in a small tray, intended to hold 1 pound of meat. A random sample of 35 packages in the small tray produced weight measurements with an average of 1.01 pounds and a standard deviation of 0.18 pounds.

   a) If you were the quality control manager and wanted to make sure that the average amount of ground beef was indeed 1 pound, what hypotheses would you test?

   b) Find the p-value for the test and use it to perform the test in part a).

   c) How would you, as the quality control manager, report the results of your study to a consumer interest group?

Statistical Significance

- The critical value approach and the *p*-value approach always produce identical results.
- The *p*-value approach is often preferred because it gives more information.
  - Computer printouts usually calculate *p*-values
- Both approaches involve setting a limit for **statistical significance**. E.g. $\alpha = 0.05$ (If the H_0 is true, 5% of possible results will cause us to reject it).
- Test results not in the rejection region are considered to not be statistically significant.
  - This means they are not unusual enough to reject H_0.
- The lower the p-value associated with an experiment result the stronger the evidence that we should reject H_0.

Levels of Significance

- Sometimes people have a naming scheme for significance levels:
- If the *p*-value is less than **.01** the results are **highly significant** (reject H_0).
- If the *p*-value is between **.01** and **.05** the results are **statistically significant** (reject H_0).
- If the *p*-value is greater than **.05** the results are **not statistically significant** (do not reject H_0).
- **This is just an example of how we might talk about *p*-values. Not everyone uses this system.**

p-values in science

- p-values are very important in experimental science.
- If you try to read a modern science paper, there's a good chance you'll see results reported using a p-value, or an equivalent.
- In fields like biology or medicine, the significance level 0.05 is the standard.
  - This is because in medicine it is very hard to set up precise experiments, so researchers have to do the best they can with noisy data.
- In particle physics, significance levels are set much lower.
  - This is because particle physicists use very expensive equipment to repeat experiments many times, so they can demand more precision.
  - For example, when they reported the discovery of the Higgs boson in 2012, physicists used the equivalent of a p-value of 0.0000003.

Problems with p-values?

- Recently, many problems with research in several branches of science have been discovered.
  - https://statmodeling.stat.columbia.edu/2016/09/21/what-has-happened-down-here-is-the-winds-have-changed/
- Results that seemed to have been established by experiments now look doubtful.
- Part of the problem is that p-values are a target to aim at, and scientists can accidentally bias their analysis to hit it.
- Another problem is that people often subtly misunderstand what p-values tell them.
  - https://royalsocietypublishing.org/doi/full/10.1098/rsos.140216
- The point is that hypothesis tests are not magic. To move towards the truth scientists have to be very careful with how they use statistics.
- This is one reason why science is so hard.

Hypothesis tests for other statistics.

- We've seen how to do hypothesis tests for population means.
- We can also use the same techniques for related things like binomial proportions and the difference between means.
- We will see how this works by looking at examples.

Example 3

Average Daily Intakes	Men	Women
Sample size	50	50
Sample mean	756	762
Sample Std Dev	35	30

- Is there a difference in the average daily intakes of dairy products for men versus women? Use $\alpha = .05$.

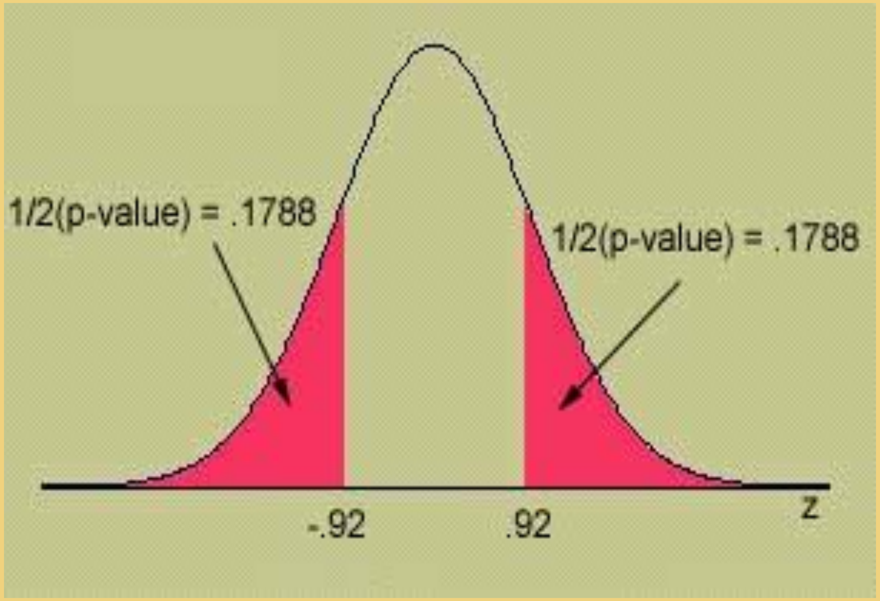
H_0: $\mu_1 - \mu_2 = 0$ (same) vs. H_a: $\mu_1 - \mu_2 \neq 0$ (different)

Test Statistics:

$$z \approx \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{756 - 762}{\sqrt{\frac{35^2}{50} + \frac{30^2}{50}}} = -0.92$$

p-Value Approach

$$p-\text{value} = P(Z < -0.92) + P(Z > 0.92)$$
$$= 2(0.1788) = 0.3576$$



Since p-value > $\alpha = .05$, H_0 is not rejected.

There is insufficient evidence to indicate that men and women have different average daily intakes.

Example 4

- Regardless of age, about 20% of American adults participate in fitness activities at least twice a week. A random sample of 100 adults over 40 years old found 15 who exercised at least twice a week. Is this evidence of a decline in participation after age 40? Use $\alpha = .05$.

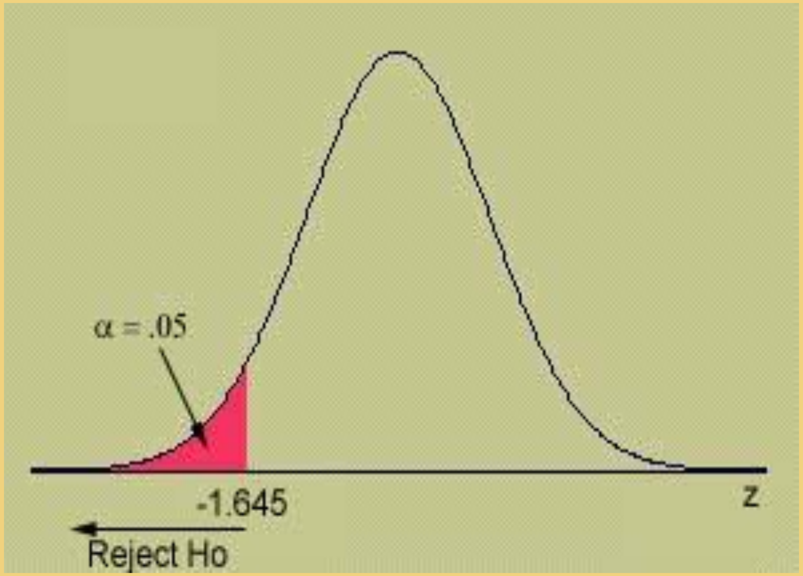
$$H_0: p \geq 0.2$$
$$H_a: p < 0.2$$

Test Statistic:

$$z \approx \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} = \frac{0.15 - 0.2}{\sqrt{\frac{0.2(0.8)}{100}}} = -1.25$$

Critical Value Approach

What is the critical value of z that cuts off exactly $\alpha = .05$ in the left-tail of the z distribution?



For our example, $z = -1.25$ does not fall in the rejection region and H_0 is not rejected.

There is not enough evidence to indicate that p is less than 0.2 for people over 40.

Rejection Region: Reject H_0 if $z < -1.645$.

Example 5

Youth Soccer	Male (1)	Female (2)
Sample size	80	70
Played soccer	65	39

- Compare the proportion of male and female college students who said that they had played on a soccer team during their K-12 years using a 1% test of hypothesis.

$$H_0: p_1 - p_2 = 0 \text{ (same)} \quad \text{vs.} \quad H_a: p_1 - p_2 \neq 0 \text{ (different)}$$

$$\text{Calculate} \quad \hat{p}_1 = \frac{65}{80} = 0.81, \quad \hat{p}_2 = \frac{39}{70} = 0.56$$

We want to calculate the test statistic, which is formally given by the equation $\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$.

Example 5

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$$

- We don’t know the values of p_1 and p_2, but we can estimate them with $\hat{p}_1$ and $\hat{p}_2$.
- But remember, we are assuming as H_0 that $p_1 = p_2 = p$ for some p.
- If this assumption is true, we can get better estimates for the population proportion by combining the samples.

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{104}{150} = 0.69$$

$\hat{p}_2 = 39 / 70 = .56$

- This gives a better estimate for p_1 and p_2 which we can use in the test statistic formula.

Example 5

Youth Soccer	Male (1)	Female (2)
Sample size	80	70
Played soccer	65	39

Test Statistic :

$$z = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.81 - 0.56}{\sqrt{0.69(0.31)\left(\frac{1}{80} + \frac{1}{70}\right)}} = 3.30$$

$$p\text{-value} = P(Z > 3.30) + P(Z < -3.30) = 2(0.0005) = 0.001$$

Since the *p*-value is less than $\alpha = .01$, H_0 is rejected. The results are highly significant. There is evidence to indicate that the rates of participation are different for boys and girls.

Class Activity 11

6. Independent random samples of 36 and 45 observations are drawn from two quantitative populations, 1 and 2, respectively. The sample data summary is shown here:

	Sample 1	Sample 2
Sample size	36	45
Sample mean	1.24	1.31
Sample standard deviation	0.2366	0.2324

Do the data present sufficient evidence to indicate that the mean for the population 1 is smaller than the mean for population 2? Use one of the two methods of testing presented in this section, and explain your conclusion.

Class Activity 11

7. A new teaching method is trialled before an exam. From 46 students taught with the new method, 39 passed. From 82 students taught with the old method, 66 passed.

   a) Does this give enough evidence to conclude that the new method gives different results (either better or worse)? Use a hypothesis test at the 10% significance level by calculating a p-value.

   b) Construct a 90% confidence interval to estimate the difference in pass rates $p_1 - p_2$.

   c) Use your confidence interval to answer the question from part a). Is your new answer the same as the old one?

Video Links

- Simple hypothesis testing
  https://www.youtube.com/watch?v=5D1gV37bKXY
- An introduction to hypothesis testing
  https://www.youtube.com/watch?v=tTeMYuS87oU
- Rejection region approach
  https://www.youtube.com/watch?v=60x86lYtWI4&index=3&list=PLvxOuBpazmsNo893xlpXNfMzVpRBjDH67
- P-value approach
  https://www.youtube.com/watch?v=m6sGjWz2CPg&list=PLvxOuBpazmsNo893xlpXNfMzVpRBjDH67&index=4