Lecture 6: Hypothesis Testing I

Taeyong Park

Carnegie Mellon University in Qatar

February 19, 2018

Next Week

- Midterm Exam on Feb 28 during class time **at 1185**.
  - Part 1 (paper-and-pencil type) and Part 2 (R questions involving data).
    - ★ Closed book; Closed R script; No googling.
    - ★ Two cheat sheets of A4-size paper allowed.
- Exam Review on Feb 26.
- Problem Set 2 assigned on Feb 21 and due Feb 26.

This Week

- Hypothesis testing for population means and for population proportions.
- Lab sessions.
- In-class exercise quiz.
- Problem Set 2.

Hypothesis testing: The big picture

- The goal of hypothesis testing is to see if the data agree with some prediction we make based on our theory.

Hypothesis

In statistics, a hypothesis is a statement about a population. It is usually a prediction that a parameter describing some characteristic of a variable that takes a particular numerical value or falls in a certain range of values.

- To test a hypothesis, we take our data and conduct a *significance test* or *hypothesis test*: Does the data support my hypothesis?

Hypothesis testing: The big picture

Confidence intervals

- You want to know μ.
- You use a sample mean $\overline{y}$ and a theoretical distribution of $\overline{y}$ to provide an interval estimate for μ.
- The confidence level determines the length of C.I. and how much confidence you have about your estimation.

Hypothesis testing: The big picture

Hypothesis testing

- You want to know μ.
- You formulate your hypothesis about μ based on your theory. **[Research hypothesis]**
- You formulate another hypothesis that negates your research hypothesis. **[Null hypothesis]**
- Evaluate whether the sample data *you have observed* would be very unlikely if the **null hypothesis** were true.
- If very unlikely, you infer that your research hypothesis is true.
  - Proof by contradiction: It starts by assuming that the opposite proposition is true, and then shows that such an assumption leads to a contradiction.
- Use the significance level ($\alpha = 1 - \text{conf. level}$) to determine the "unlikeliness."

Hypothesis testing: The big picture

To summarize,

- Assume the null hypothesis is true. $\rightarrow$ See if you can reject the null hypothesis based on the data at hand. $\rightarrow$ If you reject, you will infer that your research hypothesis is true.

Why not try to prove your research hypothesis is true directly?

Hypothesis testing: The big picture

- Science and hypothesis testing are based on **falsification**: you never know if there isn't one more experiment that will prove your hypothesis wrong.
  - "All swans are white cannot be proved true by any number of observations of white swan –we might have failed to spot a black swan somewhere– but it can be shown false by a single authentic sighting of a black swan. Scientific theories of this universal form, therefore, can never be conclusively verified, though it may be possible to falsify them." (Karl Popper).
  - "A thousand scientists can't prove me right, but one can prove me wrong." (Albert Einstein).
- We can't prove a hypothesis but we can disprove it.

Hypothesis testing: Specifics

- A business travel magazine wants to classify transatlantic gateway airports according to the mean rating for the population of business travelers.
  - A rating scale: $1 - 10$; Rating > 7: superior service airports.
  - Consider a sample for London's Heathrow Airport: surveyed 60 business travelers; $\overline{y} = 7.25$ and $s = 1.052$.

- The magazine hypothesizes that Heathrow is a superior service airport and wants to test it.
  - Why would it want to **test**?

Step 1 of 4: Make some assumptions about your data

- To get some leverage on the problem, we have to make some assumptions about the data and where it came from.
  - Type of data.
  - Sample size.
  - Population distribution.
  - Sample method (i.e., randomization).

Step 1 of 4: Make some assumptions about your data

- To get some leverage on the problem, we have to make some assumptions about the data and where it came from.
  - Type of data.
  - Sample size.
  - Population distribution.
  - Sample method (i.e., randomization).

- Heathrow Airport example:
  - **Type of data**: Numerical data.
  - **Sample size**: Large enough for the Central Limit Theorem.
  - **Population distribution**: Given the size and numerical data, a normal distribution.
  - **Sampling method**: Assumes that the business travelers were selected at random.

Step 2 of 4: Formulate the null and alternative hypotheses

- Heathrow Airport example:
  - **Research hypothesis, also called alternative hypothesis**: Heathrow Airport is a superior service airport.
    - ★ $H_a : \mu > 7$
  - **Null hypothesis**: Heathrow Airport is not a superior service airport.
    - ★ $H_0 : \mu \leq 7$
  - $\overline{y} = 7.25$ is an estimate, casting doubt on the null hypothesis. But we must verify that if $\overline{y} = 7.25$ is **significantly** greater than 7 in the **statistical** sense.

Step 3 of 4: Calculate a test statistic

- Assume the null hypothesis is true.
- Calculate a test statistic that summarizes how much our sample observation differs from what we would have expected to observe *if the null hypothesis were true*. (i.e., what H_0 predicts.)
- Test statistic $= \dfrac{\text{Sample observation (estimate)} - \mu_{H_0}}{\text{Standard error}_{H_0}}$.
- The test statistic measures the number of standard deviations the observed data is away from what H_0 predicts.

Step 3 of 4: Calculate a test statistic

Heathrow Airport example:

- Test statistic $= \dfrac{\overline{y} - \mu_{H_0}}{\sigma_{H_0}} = \dfrac{7.25 - 7}{\sigma_{H_0}}$.
- σ_{H_0} should be estimated by $\dfrac{S}{\sqrt{n}} = \dfrac{1.052}{\sqrt{60}}$.
  - $\overline{y} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$
- $\therefore$ Test statistic $= 1.84$.

Step 3 of 4: Calculate a test statistic

- We are going to assess **how unlikely the test statistic is**.
  - If it is very ulikely to see such a value, it means the observed data is far away from what H_0 predicts and we will reject the null hypothesis: *proof by contradiction*.

- It's important to configure **the distribution of the test statistic** to evaluate unlikeliness.

Step 3 of 4: Calculate a test statistic

- If the sampling distribution of $\overline{y}$ is a normal distribution, test statatistic $= \frac{\overline{y}-\mu_0}{\sigma_{\mu_0}}$ follows the standard normal distribution.
- But, mostly, hypothesis tests about a population mean are based on the t distribution to account for an increased error by estimating σ.
- Test statistic $= \dfrac{\overline{x}-\mu_0}{\sigma_{\mu_0}} = 1.84$ follows the t distribution.
- The test statistic is a t-value, or t-score.
- t-test.

Step 4 of 4: Assess how unusual the test statistic value is

- Assess how unusual a test statistic value is **given the distribution of that test statistic**.
- How unusual? $\Leftrightarrow$ How far out in the tail of the distribution.

- Two approaches: p-value and critical value.

Step 4 of 4: Assess how unusual the test statistic value is

1. The p-value approach

p-value

A p-value is a measure of surprise: The probability that we would observe the test statistic or a value even more extreme if *the null hypothesis were true*.

- $p = Pr(t \geq test.stat)$ in one-tailed test, where $test.stat > 0$.
- $p = Pr(t \leq -test.stat)$ or $Pr(t \geq test.stat) = 2Pr(t \geq test.stat)$ in two-tailed, where $test.stat > 0$.

Step 4 of 4: Assess how unusual the test statistic value is

1. The p-value approach

   - A small p-value more strongly contradicts the null hypothesis. Why?

Step 4 of 4: Assess how unusual the test statistic value is

1. The p-value approach

- A small p-value more strongly contradicts the null hypothesis. Why?
- $p = 0.0001$ if H_0 were true $\rightarrow$ **very unlikely**. Thus, the smaller the p-value is, the more likely we reject the hypothesis that H_0 is true.

- Significance level 0.05. $\Rightarrow$ Reject the null if $p < 0.05$.
- Significance level 0.01. $\Rightarrow$ Reject the null if $p < 0.01$.

Step 4 of 4: Assess how unusual the test statistic value is

Heathrow Airport example:

- Test statistic $= \dfrac{\overline{x} - \mu_0}{\sigma_{\mu_0}} = \dfrac{7.25 - 7}{\sigma_{\mu_0}} = \dfrac{7.25 - 7}{1.052/\sqrt{60}} = 1.84$ is a t-score following the t distribution.

- Using a t table to find a probability associated with 1.84, where df = 60-1 = 59. $\Rightarrow$ Between 0.05 and 0.025.
- In R, run `pt(1.84, df=59, lower.tail=F)`. $\Rightarrow$ $p = 0.0354$.

Step 4 of 4: Assess how unusual the test statistic value is

$p = 0.0354 \Rightarrow$ Reject the null?

Step 4 of 4: Assess how unusual the test statistic value is

$p = 0.0354 \Rightarrow$ Reject the null?

- At the 0.05 significance level, reject the null.
  - The test statistic would be unlikely to be observed if H_0 were true. But we observed, and hence reject H_0.
  - 7.25 is significantly greater than 7.
- At the 0.01 significance level, cannot reject the null.
  - Given this more strict significance level, the test statistic now would not be unlikely to observe.

Step 4 of 4: Assess how unusual the test statistic value is

2. Critical value approach

- The critical value is the largest value of the test statistic (in absolute term) that will result in the rejection of the null hypothesis.
- It's determined by the level of significance.

Step 4 of 4: Assess how unusual the test statistic value is

2. Critical value approach

- Heathrow Airport example:
  - 5% (0.05) significance level for 59 degrees of freedom: `qt(0.05, df=59, lower.tail=F)` $= 1.67$.
  - $1.84 >$ the critical value. $\Rightarrow$ Reject the null hypothesis at the 5% significance level.
  - 1% (0.01) significance level for 59 degrees of freedom: `qt(0.01, df=59, lower.tail=F)` $= 2.391$.
  - $1.84 <$ the critical value. $\Rightarrow$ Cannot reject the null hypothesis at the 1% significance level.

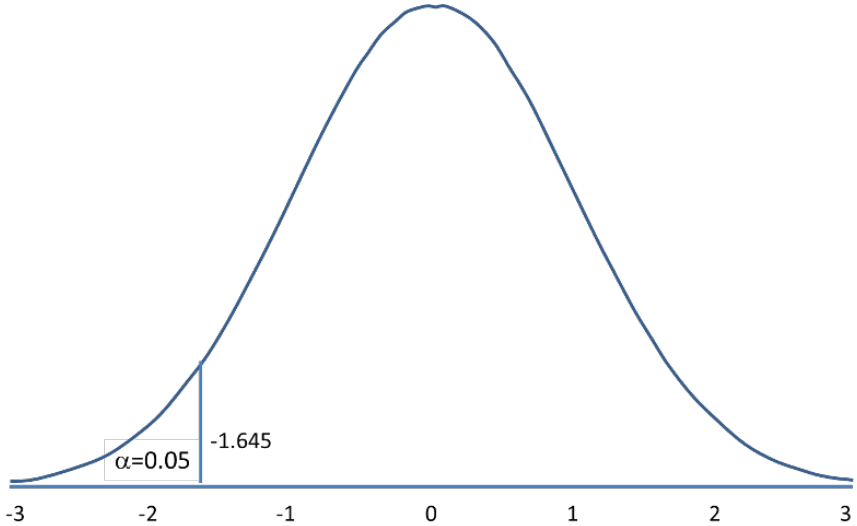
Exercise

- The Federal Trade Commission of the US suspects that Hilltop Coffee cans' filling weight is **less than 3** pounds that is the weight on the label on a large can of Hilltop Coffee. If they reach such a conclusion, the FTC will charge a label violation. To test their research hypothesis, the FTC drew a sample of 36 cans. The sample mean is 2.92 and the sample s.d. is 0.18. Perform a hypothesis test at the 5% significance level.

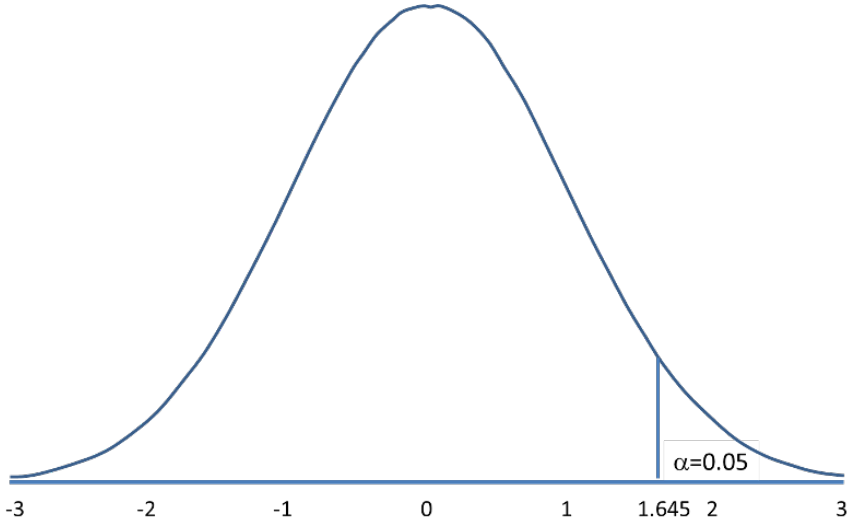
One-tailed test vs. Two-tailed test

	One-tailed test		Two-tailed test
	Lower tail test	Upper tail test	
Pop. Mean	$H_0 : \mu \geq \mu_0$ $H_a : \mu < \mu_0$	$H_0 : \mu \leq \mu_0$ $H_a : \mu > \mu_0$	$H_0 : \mu = \mu_0$ $H_a : \mu \neq \mu_0$

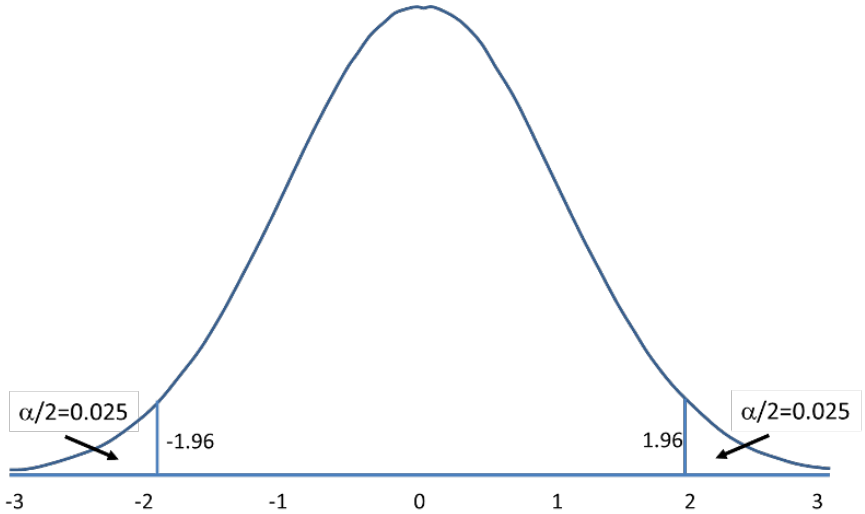
One-tailed test vs. Two-tailed test



Lower tail (5% significance level)

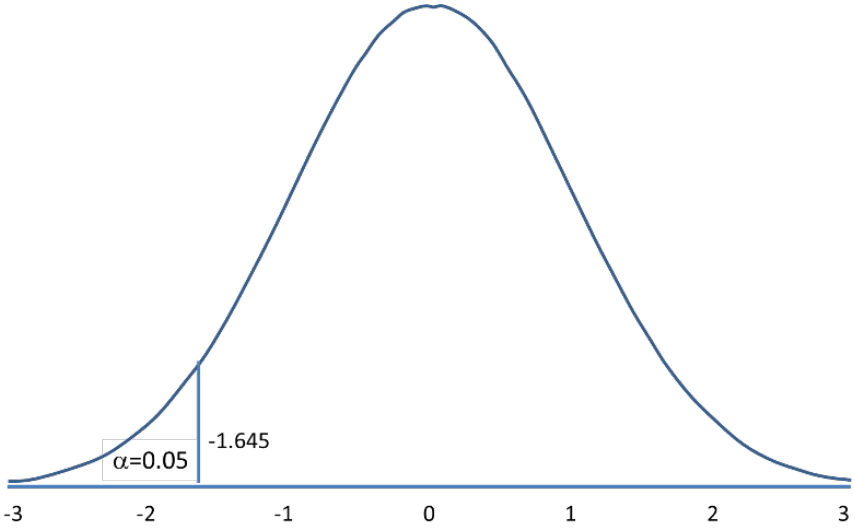


Upper tail

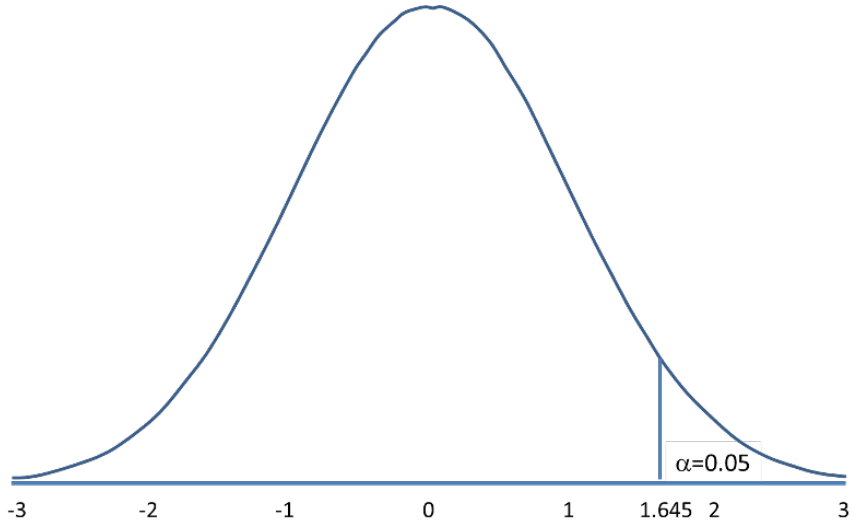


Two-tailed (5% significance level)

One-tailed test vs. Two-tailed test



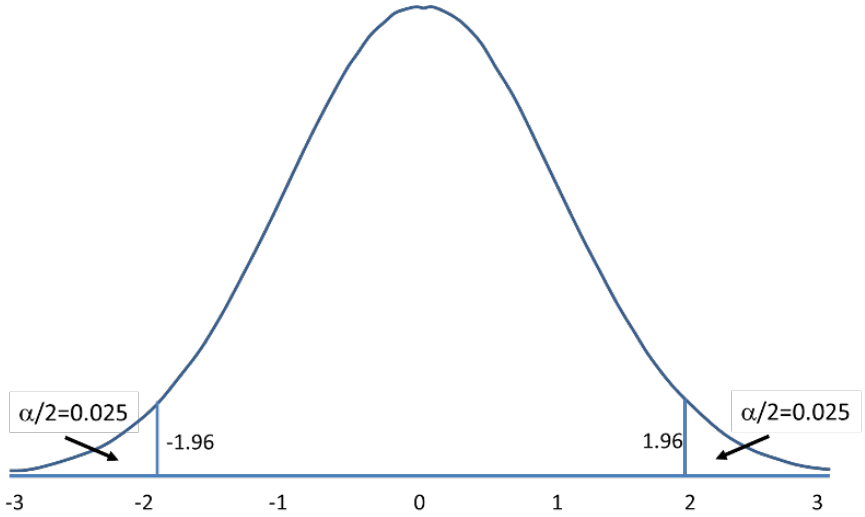
Lower tail (5% significance level)



Upper tail

- $p = Pr(t \geq test.stat)$ in upper one-tailed test.
  - p-value in R? `pt(test.stat, df, lower.tail=F)`
  - critical value in R?
- $p = Pr(t \leq test.stat)$ in lower one-tailed test.
  - p-value in R?
  - critical value in R?

One-tailed test vs. Two-tailed test



Two-tailed (5% significance level)

- $p = Pr(t \leq -test.stat)$ or $Pr(t \geq test.stat) = 2Pr(t \geq test.stat)$, where $test.stat > 0$.
- $p = Pr(t \leq test.stat)$ or $Pr(t \geq -test.stat) = 2Pr(t \leq test.stat)$, where $test.stat < 0$.
  - p-value in R?
  - critical value in R?

Choice of one-tailed test vs. two-tailed test?

- When you predict the direction of an effect, one-tailed.
- Through one-tailed tests, you are more likely to reject the null provided the same significance level.
  - Given the same test.stat, $p = Pr(t \geq test.stat)$ vs. $p = 2Pr(t \geq test.stat)$.
- In practice, two-tailed tests are more common.
  - Two-tailed tests reflect an objective approach to research that recognizes that an effect could go in either direction.
  - Two-tailed tests are more conservative in terms of the likelihood of rejecting the null hypothesis.

Exercise

- The Federal Trade Commission of the US suspects that Hilltop Coffee cans' filling weight **is not equal to** 3 pounds that is the weight on the label on a large can of Hilltop Coffee. To test their research hypothesis, the FTC drew a sample of 36 cans. If they reach such a conclusion, the FTC will decide to investigate Hilltop Coffee more thoroughly. If they don't, they will let Hilltop Coffee continue sell the item. The sample mean is 2.92 and the sample s.d. is 0.18. Perform a hypothesis test at the 5% significance level.