

# Lecture 8 - Hypothesis Testing

Sta102/BME102

Colin Rundel

February 12, 2014

# Hypothesis testing framework

- We start with a *null hypothesis* ( $H_0$ ) that represents the status quo.
- We develop an *alternative hypothesis* ( $H_A$ ) that represents our research question (what we're testing for).
- We conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation or theoretical methods.
- If the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, we stick with the null hypothesis. If they do, then we reject the null hypothesis in favor of the alternative.

We'll formally introduce the hypothesis testing framework using an example on testing a claim about a population mean.

## Example - Grade inflation?

In 2001 the average GPA of students at Duke University was 3.37. Last semester Duke students in a Stats class were surveyed and asked for their current GPA. This survey had 147 respondents and yielded an average GPA of 3.56 with a standard deviation of 0.31.

Assuming that this sample is random and representative of all Duke students, do these data provide convincing evidence that the average GPA of Duke students has *changed* over the last decade?

# Setting the hypotheses

- The *parameter of interest* is the average GPA of current Duke students.
- There may be two explanations why our sample mean is higher than the average GPA from 2001.
  - The true population mean has changed.
  - The true population mean remained at 3.37, the difference between the true population mean and the sample mean is simply due to natural sampling variability.
- We start with the assumption that nothing has changed.

$$H_0 : \mu = 3.37$$

- We test the claim that average GPA has changed.

$$H_A : \mu \neq 3.37$$

## Making a decision - p-values

We would know like to make a decision about whether we think  $H_0$  or  $H_A$  is correct, to do this in a principled / quantitative way we calculate what is known as a *p-value*.

- The *p-value* is the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis was true.
- If the p-value is *low* (lower than the significance level,  $\alpha$ , which is usually 5%) we say that it would be very unlikely to observe the data if the null hypothesis were true, and hence *reject  $H_0$* .
- If the p-value is *high* (higher than  $\alpha$ ) we say that it is likely to observe the data even if the null hypothesis were true, and hence *do not reject  $H_0$* .
- We never accept  $H_0$  since we're not in the business of trying to prove it. We simply want to know if the data provide convincing evidence against  $H_0$ .

# Conditions for inference

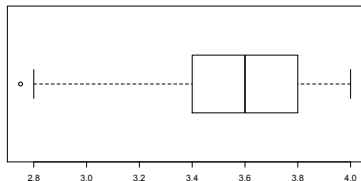
In order to perform inference using this data set, we need to use the CLT and therefore we must make sure that the necessary conditions are satisfied:

## 1. *Independence*:

- We have already assume this sample is random.
- $147 < 10\%$  of all current Duke students.

⇒ we are safe assuming that GPA of one student in this sample is independent of another.

## 2. *Sample size / skew*: The distribution appears to be slightly skewed (but not extremely) and $n > 30$ so we can assume that the distribution of the sample means is nearly normal.



# Calculating the p-value

*p-value*: probability of observing data at least as favorable to  $H_A$  as our current data set (a sample mean greater than 3.56 or less than 3.18), if in fact  $H_0$  was true (the true population mean was 3.37).

$$\begin{aligned} & P(\bar{x} > 3.56 \text{ or } \bar{x} < 3.18 \mid \mu = 3.37) \\ &= P(\bar{x} > 3.56 \mid \mu = 3.37) + P(\bar{x} < 3.18 \mid \mu = 3.37) \\ &= P\left(Z > \frac{3.56 - 3.37}{0.31/\sqrt{147}}\right) + P\left(Z < \frac{3.18 - 3.37}{0.31/\sqrt{147}}\right) \\ &= P(Z > 7.43) + P(Z < -7.43) \\ &= 10^{-13} \approx 0 \end{aligned}$$

# Drawing a Conclusion / Inference

$$p\text{-value} \approx 10^{-13}$$

- If the true average GPA Duke students applied to is 3.37, there is approximately a  $10^{-11}$  % chance of observing a random sample of 147 Duke students with an average GPA of 3.56.
- This is a very low probability for us to think that a sample mean of 3.56 GPA is likely to happen simply by chance.
- Since p-value is *low* (lower than 5%) we *reject  $H_0$* .
- The data provide convincing evidence that Duke students average GPA has changed since 2001.
- The difference between the null value of a 3.37 GPA and observed sample mean of 3.56 GPA is *not due to chance* or sampling variability.



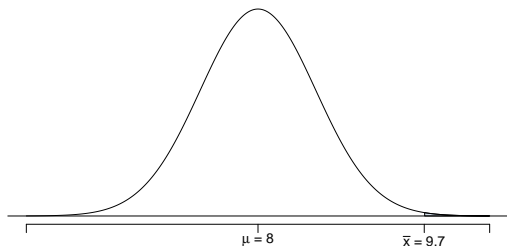
## Example - College applications

A similar survey asked how many colleges each student had applied to. 206 students responded to this question and the sample yielded an average of 9.7 college applications with a standard deviation of 7. The College Board website states that counselors recommend students apply to roughly 8 colleges. What would be the correct set of hypotheses to test if these data provide convincing evidence that the average number of colleges Duke students apply to is *higher* than the number recommended by the College Board. Are the conditions for inference met?

<http://www.collegeboard.com/student/apply/the-application/151680.html>

# College Applications - p-value

*p-value*: probability of observing data at least as favorable to  $H_A$  as our current data set (a sample mean greater than 9.7), if in fact  $H_0$  was true (the true population mean was 8).



$$P(\bar{x} > 9.7 \mid \mu = 8) = P\left(Z > \frac{9.7 - 8}{7/\sqrt{206}}\right) = P(Z > 3.4) = 0.0003$$

# College Applications - Making a decision

$$p - \text{value} \approx 0.0003$$

- If the true average of the number of colleges Duke students applied to is 8, there is only 0.03% chance of observing a random sample of 206 Duke students who on average apply to 9.7 or more schools.
- This is a pretty low probability for us to think that a sample mean of 9.7 or more schools is likely to happen simply by chance.
- Since p-value is *low* (lower than 5%) we *reject  $H_0$* .
- The data provide convincing evidence that Duke students average apply to more than 8 schools.
- The difference between the null value of 8 schools and observed sample mean of 9.7 schools is *not due to chance* or sampling variability.

## Example - Sleep

A poll by the National Sleep Foundation found that college students average about 7 hours of sleep per night. A sample of 169 Duke students (you!) yielded an average of 6.88 hours, with a standard deviation of 0.94 hours. Assuming that this is a random sample representative of all Duke students, a hypothesis test was conducted to evaluate if Duke students on average sleep *less than* 7 hours per night. The p-value for this hypothesis test is 0.0485. Which of the following is correct?

What are the hypotheses being tested?

What is the correct inference for this situation?

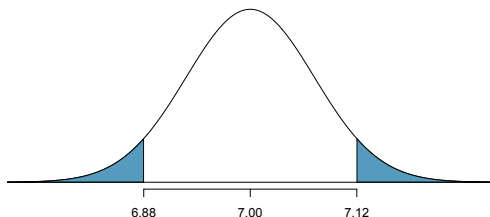
## Two-sided hypothesis test

If the research question had been “Do the data provide convincing evidence that the average amount of sleep Duke students get per night is *different* than the national average?”, the alternative hypothesis would be different.

$$H_0 : \mu = 7$$

$$H_A : \mu \neq 7$$

- Hence the p-value would change, as well as our decision to reject:



$$\begin{aligned} \text{p-value} &= 0.0485 \times 2 \\ &= 0.097 \end{aligned}$$

*Fail to reject  $H_0$ !*

# Recap: Hypothesis testing for a population mean

- 1 Set the hypotheses
  - $H_0 : \mu = \text{null value}$
  - $H_A : \mu < \text{or } > \text{ or } \neq \text{ null value}$
- 2 Check assumptions and conditions
  - Independence: random sample/assignment, 10% condition when sampling without replacement
  - Normality: nearly normal population or  $n \geq 30$ , no extreme skew
- 3 Calculate a *test statistic* and a p-value (draw a picture!)

$$Z = \frac{\bar{x} - \mu}{SE}, \text{ where } SE = \frac{s}{\sqrt{n}}$$

- 4 Make a decision, and interpret it in context of the research question
  - If p-value  $< \alpha$ , reject  $H_0$ , data provide strong evidence for  $H_A$
  - If p-value  $> \alpha$ , do not reject  $H_0$

## Example - Sample Size

Suppose  $\bar{x} = 50$ ,  $s = 2$ ,  $H_0 : \mu = 49.5$ , and  $H_A : \mu \geq 49.5$ .

Will the p-value be lower if  $n = 100$  or  $n = 10,000$ ?

# Statistical vs. Practical Significance

- Real differences between the point estimate and null value are easier to detect with larger samples
- However, very large samples will result in statistical significance even for tiny differences between the sample mean and the null value (*effect size*), even when the difference is not practically significant
- This is especially important to research: if we conduct a study, we want to focus on finding meaningful results (we want observed differences to be real but also large enough to matter).
- The role of a statistician is not just in the analysis of data but also in planning and design of a study.

*“To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of.” – R.A. Fisher*



# Decision errors

- Hypothesis tests are not flawless.
- In the court system innocent people are sometimes wrongly convicted and the guilty sometimes walk free.
- Similarly, we can make a wrong decision in statistical hypothesis tests as well.
- The difference is that we have the tools necessary to quantify how often we make errors in statistics.

## Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject $H_0$	reject $H_0$
Truth	$H_0$ true	✓	<i>Type 1 Error</i>
	$H_A$ true	<i>Type 2 Error</i>	✓

- A *Type 1 Error* is rejecting the null hypothesis when  $H_0$  is true.
- A *Type 2 Error* is failing to reject the null hypothesis when  $H_A$  is true.
- We (almost) never know if  $H_0$  or  $H_A$  is true, but we need to consider all possibilities.

# Hypothesis Test as a trial

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

$H_0$  : Defendant is innocent

$H_A$  : Defendant is guilty

Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty
- Declaring the defendant guilty when they are actually innocent

Which error do you think is the worse error to make?

“better that ten guilty persons escape than that one innocent suffer”

– William Blackstone

# Type 1 error rate

- As a general rule we reject  $H_0$  when the p-value is less than 0.05, i.e. we use a *significance level* of 0.05,  $\alpha = 0.05$ .
- This means that, for those cases where  $H_0$  is actually true, we do not want to incorrectly reject it more than 5% of those times.
- In other words, when using a 5% significance level there is about 5% chance of making a Type 1 error.

$$P(\text{Type 1 error}) = \alpha$$

- This is why we prefer to small values of  $\alpha$  – increasing  $\alpha$  increases the Type 1 error rate.

## Filling in the table...

		Decision	
		fail to reject $H_0$	reject $H_0$
Truth	$H_0$ true	$1 - \alpha$	Type 1 Error, $\alpha$
	$H_A$ true	Type 2 Error, $\beta$	Power, $1 - \beta$

- Type 1 error is rejecting  $H_0$  when you shouldn't have, and the probability of doing so is  $\alpha$  (significance level)
- Type 2 error is failing to reject  $H_0$  when you should have, and the probability of doing so is  $\beta$  (a little more complicated to calculate)
- *Power* of a test is the probability of correctly rejecting  $H_0$ , and the probability of doing so is  $1 - \beta$
- In hypothesis testing, we want to keep  $\alpha$  and  $\beta$  low, but there are inherent trade-offs.

## Type 2 error rate

If the alternative hypothesis is actually true, what is the chance that we make a Type 2 Error, i.e. we fail to reject the null hypothesis even when we should reject it?

- The answer is not obvious.
- If the true population average is very close to the null hypothesis value, it will be difficult to detect a difference (and reject  $H_0$ ).
- If the true population average is very different from the null hypothesis value, it will be easier to detect a difference.
- Clearly,  $\beta$  depends on the *effect size* ( $\delta$ )