Lecture 11 - Hypothesis Testing for Means

Sta102/BME102

Colin Rundel

February 20, 2015

Inference using CIs for sample means

When conditions for CLT are met, depending if σ is known or unknown:

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{df=n-1} \qquad \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim Z$$

Conditions (same as CLT):

- *Independent observations* - random sample, if sampling without replacement $n < 10\%$ of population
- *Sample size* - > 30 is usually reasonable, population not overly skewed or heavy/light tailed

Confidence interval:

$$\bar{X} \pm t^{\star}_{df=n-1} \frac{s}{\sqrt{n}} \qquad \bar{X} \pm Z^{\star} \frac{\sigma}{\sqrt{n}}$$

Error Rate:

$$1 - \text{Confidence level}$$

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.
- We examine how likely our data (or something more extreme) is under this assumption, and use that as evidence against the null hypothesis (and hence for 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 ask 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*.

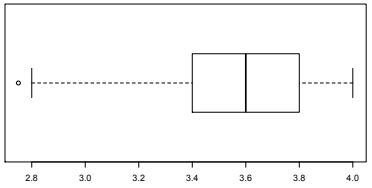
- The *p-value* is the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis was true.
- If the p-value is *low* (lower than the significance level, α, which is usually 5%) we say that it would be very unlikely to observe the data if the null hypothesis were true, and hence *reject* H_0.
- If the p-value is *high* (higher than α) 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 assumed this sample is random.
   - 147 $<$ 10% of all current Duke students.

   $\Rightarrow$ it appears reasonable to assume 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 is true (the true population mean $\mu = 3.37$).

Therefore, assuming H_0 is true,

$$T = \frac{\bar{X} - \mu}{s/n} \sim t_{df=n-1}$$

$$\begin{aligned}
\text{p-value} &= P(\bar{X} > 3.56 \text{ or } \bar{X} < 3.18) \\
&= P(\bar{X} > 3.56) + P(\bar{X} < 3.18) \\
&= P\left(T > \frac{3.56 - 3.37}{0.31/\sqrt{147}}\right) + P\left(T < \frac{3.18 - 3.37}{0.31/\sqrt{147}}\right) \\
&= P\left(T > 7.43\right) + P\left(T < -7.43\right) = 2 \times P(T < -7.43) \\
&\approx 10^{-11}
\end{aligned}$$

Drawing a Conclusion / Inference

$$p\text{-value} \approx 10^{-11}$$

If the true average GPA Duke students applied to is 3.37, there is approximately a 10^{-9} % 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.

What about a confidence interval?

We can also assess the claim of grade inflation using a confidence interval.

$$\bar{X} = 3.56 \qquad s^2 = 0.31^2 \qquad n = 147$$

We construct a 95% confidence interval using $t^*_{df=146} \approx t^*_{df=150} = 1.98$,

$$\begin{aligned} CI &= \bar{X} \pm t^* \, s/\sqrt{n} \\ &= 3.56 \pm 1.98(0.31/\sqrt{147}) \\ &= 3.56 \pm 0.05 \\ &= (3.51,\ 3.61) \end{aligned}$$

What does this tell us about claim about grade inflation given that average GPA used to be 3.37? *3.37 is not a plausible claim for the current average GPA of Duke students.*

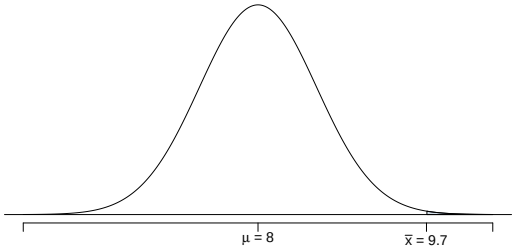
Example - College applications

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).



$$P(\bar{X} > 9.7) = P\left(T > \frac{9.7 - 8}{7/\sqrt{206}}\right) = P(T > 3.4) < 0.005$$

College Applications - Making a decision

$$p - value \approx 0.0003$$

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.

What about a confidence interval?

We can also assess this claim using a confidence interval.

$$\bar{X} = 9.7 \qquad s^2 = 7^2 \qquad n = 206$$

We construct a 95% confidence interval using $t^*_{df=205} \approx t^*_{df=200} = 1.97$,

$$\begin{aligned} CI &= \bar{X} \pm t^* \, s/\sqrt{n} \\ &= 9.7 \pm 1.97(7/\sqrt{206}) \\ &= 9.7 \pm 0.96 \\ &= (8.74,\ 10.66) \end{aligned}$$

What does this tell us about claim about Duke Students applying to 8 schools on average? *8 is not a plausible claim for the average number of applications of Duke students.*

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.

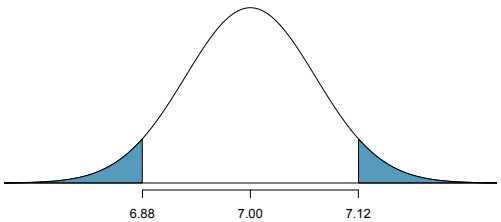
What are the hypotheses being tested?
What is the correct inference for this situation?

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:



p-value
$= 0.0485 \times 2$
$= 0.097$

Fail to reject H_0!

What about a confidence interval?

Once again, we can also assess this claim using a confidence interval.

$$\bar{X} = 6.88 \qquad s^2 = 0.94^2 \qquad n = 169$$

We construct a 95% confidence interval using $t^*_{df=168} \approx t^*_{df=150} = 1.98$,

$$\begin{aligned} CI &= \bar{X} \pm t^* \, s/\sqrt{n} \\ &= 6.88 \pm 1.98(0.94/\sqrt{169}) \\ &= 6.88 \pm 0.14 \\ &= (6.74,\ 7.02) \end{aligned}$$

What does this tell us about claim about Duke Students get to 7 hours of sleep a night on average? *7 is a plausible claim for the average hours of sleep a night for Duke students.*

Recap: Hypothesis testing for a population mean

1. Set the hypotheses
   - $H_0 : \mu =$ *null value*
   - $H_A : \mu <$ or $>$ or $\neq$ *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!)

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \qquad T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

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 $H_A : \mu > 49.5$.

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

$$T_{n=100} = \frac{50 - 49.5}{\frac{2}{\sqrt{100}}} = \frac{50 - 49.5}{\frac{2}{10}} = \frac{0.5}{0.2} = 2.5, \quad \text{p-value} = 0.007$$

$$T_{n=10000} = \frac{50 - 49.5}{\frac{2}{\sqrt{10000}}} = \frac{50 - 49.5}{\frac{2}{100}} = \frac{0.5}{0.02} = 25, \quad \text{p-value} \approx 0$$

As n increases - $SE \downarrow$, $Z \uparrow$, p-value $\downarrow$

Example - Sample Size 2

Suppose $\bar{X} = 50$, $s = 2$, $H_0 : \mu = 49.9$, and $H_A : \mu > 49.9$.

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

$$T_{n=100} = \frac{50 - 49.9}{\frac{2}{10}} = \frac{0.1}{0.2} = 0.5, \quad \text{p-value} = 0.309$$

$$T_{n=10000} = \frac{50 - 49.9}{\frac{2}{100}} = \frac{0.1}{0.02} = 5, \quad \text{p-value} = 2.87 \times 10^{-7}$$

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 and Confidence Intervals 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 for HTs

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	✓	*Type 1 Error*
	H_A true	*Type 2 Error*	✓

- 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 cirumstances?

- 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 will incorrectly reject it at most 5% of the time.
- In other words, when using a 5% significance level there is about 5% chance of making a Type 1 error.

$$P(\text{Type 1 error}) = P(\text{Reject } H_0 | H_0 \text{ is true}) = \alpha$$

- This is why we prefer small values of α – increasing α 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,* α
	H_A true	*Type 2 Error,* β	*Power,* $1 - \beta$

- Type 1 error is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level)
- Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β (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 α and β 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, but
  - 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.
- Therefore, β must depend on the *effect size* (δ) in some way

To increase power / decrease β: increase n, increase δ, or increase α