Lecture 11 - Hypothesis Tests for Mean(s)

Sta102/BME102

Colin Rundel

September 7, 2015

Inference using CIs for sample means

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

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

Conditions (same as CLT):

- *Independent observations* - random sample, if sampling without replacement $n < 10\%$ of population
- *Sample size* - $> 20 - 30$ is usually reasonable, population not overly skewed or heavy/light tailed

Confidence interval:

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

Null Value 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). It should be mutually exclusive to H_0.
- 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 - we will start with an example we looked at last time using a confidence interval.

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 63 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*.

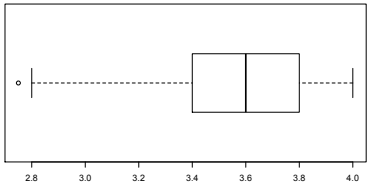
- The *p-value* is the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis was true.
- If the p-value is *low* ($< \alpha$, usually 5%) we say that it would be very unlikely to observe the data if the null hypothesis were true, and hence *reject* H_0.
- If the p-value is *high* ($> \alpha$) 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 on this data, we need the CLT and therefore we need to check the conditions:

1. *Independence:*
   - We have already assumed this sample is random.
   - $63 < 10\%$ 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 $n = 63$ so we can assume that the distribution of the sample means should be nearly normal by the CLT.



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,

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

$$\begin{aligned}
\text{p-value} &= P(\bar{X} > 3.56 \text{ or } \bar{X} < 3.18) \\
&= P(\bar{X} > 3.56) + P(\bar{X} < 3.18) \\
&= P\left(T > \frac{3.56 - 3.37}{0.31/\sqrt{63}}\right) + P\left(T < \frac{3.18 - 3.37}{0.31/\sqrt{63}}\right) \\
&= P\left(T > 4.86\right) + P\left(T < -4.86\right) = 2 \times P(T < -4.86) \\
&\approx 4.2 \times 10^{-6}
\end{aligned}$$

Drawing a Conclusion / Inference

$$p\text{-}value = 4.2 \times 10^{-6}$$

If the true average GPA of Duke students is 3.37, there is approximately a 4.2×10^{-6} chance of observing a random sample of 63 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.
- We therefore conclude that the difference between the null value of a 3.37 GPA and observed sample mean of 3.56 GPA is *not due to chance / 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 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 *greater* than the number recommended by the College Board.

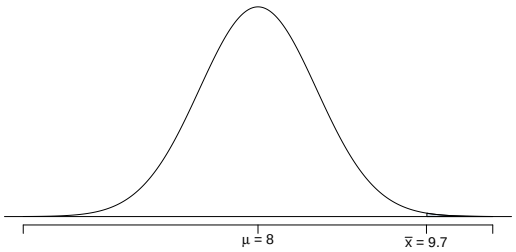
$$H_0 : \mu = 8$$
$$H_A : \mu > 8$$

Are the conditions for inference met?

Yes - Independence ✓, Nearly Normal ✓

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

$$p - value < 0.005$$

If the true average of the number of colleges Duke students applied to is 8, there is less than 0.005 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?

$$H_0 : \mu = 7$$
$$H_A : \mu < 7$$

What is the correct inference for this situation?

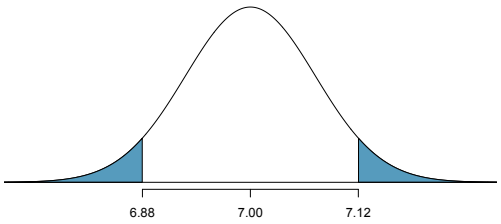
Reject H_0, the data provide convincing evidence that Duke students sleep less than 7 hours on average.

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?", how would the null and alternative hypotheses change?

$$H_0 : \mu = 7$$
$$H_A : \mu \neq 7$$

How would the p-value change?



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: Null Value Hypothesis Testing

Regardless of the sample statistic of interest, all null value hypothesis testing takes exactly the same form.

1. Set the hypotheses
2. Check assumptions and conditions
3. Calculate a *test statistic* and a p-value (draw a picture!)
4. Make a decision, and interpret it in context of the research question

Recap: Null Value Hypothesis Testing - Sample Means

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

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

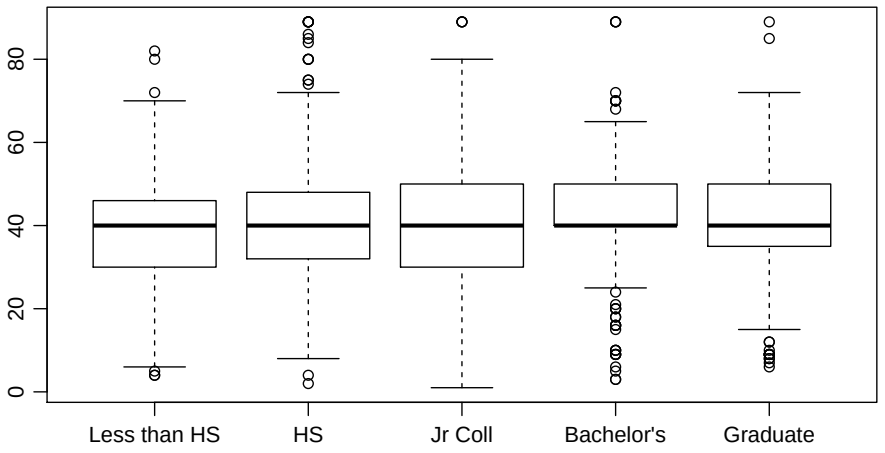
4. Make a decision, and interpret it in context of the research question
   - If p-value $< \alpha$, reject H_0, data provide strong evidence for H_A
   - If p-value $> \alpha$, do not reject H_0

Example - GSS

The General Social Survey (GSS) is an annual Census Bureau survey covering demographic, behavioral, and attitudinal questions. To facilitate time-trend studies many of the questions have not changed since 1972. Below is an excerpt from the 2010 survey. The variables are number of hours worked per week and highest educational attainment.

	degree	hrs1
1	BACHELOR	55
2	BACHELOR	45
3	JUNIOR COLLEGE	45
⋮		
1172	HIGH SCHOOL	40

Exploratory analysis



What can we say about the relationship between educational attainment and hours worked per week?

Collapsing levels

- Say we are only interested the difference between the number of hours worked per week by college and non-college graduates.
- We can combine the levels of education into:
  - `hs or lower` ← less than high school or high school
  - `coll or higher` ← junior college, bachelor's, and graduate
- Here is how you can do this in R:

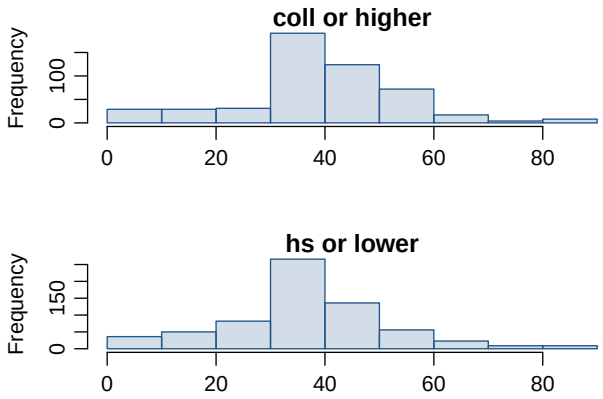
```
# create a new empty variable
gss$edu = NA

# if statements to determine levels of new variable
gss$edu[gss$degree == "LESS THAN HIGH SCHOOL" |
        gss$degree == "HIGH SCHOOL"] = "hs or lower"
gss$edu[gss$degree == "JUNIOR COLLEGE" |
        gss$degree == "BACHELOR" |
        gss$degree == "GRADUATE"] = "coll or higher"

# make sure new variable is categorical
gss$edu = as.factor(gss$edu)
```


Exploratory analysis - another look

	$\bar{x}$	s	n
coll or higher	41.8	15.14	505
hs or lower	39.4	15.12	667



Parameter and point estimate

We want to construct a 95% confidence interval for the average difference between the number of hours worked per week by Americans with a college degree and those with a high school degree or lower. What are the parameter of interest and the point estimate?

- *Parameter of interest:* Average difference between the number of hours worked per week by *all* Americans with a college degree and those with a high school degree or lower.

$$\mu_c - \mu_{hs}$$

- *Point estimate:* Average difference between the number of hours worked per week by *sampled* Americans with a college degree and those with a high school degree or lower.

$$\bar{x}_c - \bar{x}_{hs}$$

Difference of Means and the CLT

We can think about our observations as being samples from two distributions D_x and D_y,

$$X_1,\ X_2,\ \ldots,\ X_{n_x} \sim D_x$$
$$Y_1,\ Y_2,\ \ldots,\ Y_{n_y} \sim D_y.$$

We now want to know what the distribution of $\bar{x} - \bar{y}$ will be so that we can perform inference.

From our work with a single sample means, we know that the CLT tells us that both

$$\bar{x} \sim N(E(D_x),\ Var(D_x)/n_x),$$
$$\bar{y} \sim N(E(D_y),\ Var(D_y)/n_y),$$

Difference of Means and the CLT (cont.)

Proposition - the sum or difference of normal RVs is also normally distributed. (Not terribly hard to prove, but requires more probability theory than we've covered).

This proposition then tells us that

$$\bar{x} - \bar{y} \sim N\left(E(\bar{x} - \bar{y}),\ Var(\bar{x} - \bar{y})\right),$$

where

$$E(\bar{x} - \bar{y}) = E(\bar{x}) - E(\bar{y}) = \mu_x + \mu_y$$
$$Var(\bar{x} - \bar{y}) = Var(\bar{x}) + Var(\bar{y}) = \frac{\sigma_x^2}{n_x} + \frac{\sigma^2}{n_y}$$

Did I make any assumptions here?
Yes - variance result requires that $\bar{x}$ and $\bar{y}$ are independent. We call this independence between groups.

Expected Value and Variance of the Difference

Checking assumptions & conditions

1. *Independence:*
   1. *Independence within groups:*
      - Both the college graduates and those with HS degree or lower are sampled randomly.
      - 505 $<$ 10% of all college graduates and 667 $<$ 10% of all students with a high school degree or lower.

      We can assume that the number of hours worked per week by one college graduate in the sample is independent of another, and the number of hours worked per week by someone with a HS degree or lower in the sample is independent of another as well.
   2. *Independence between groups:*
      Since the sample is random, the college graduates in the sample are independent of those with a HS degree or lower.
2. *Sample size / skew:*
   Both distributions look reasonably symmetric, and the sample sizes are large, therefore we can assume that the sampling distribution of number of hours worked per week by college graduates and those with HS degree or lower are nearly normal. Hence the sampling distribution of the average difference will be nearly normal as well.

Confidence interval for difference between two means

- All confidence intervals have the same form:

$$point\ estimate \pm ME$$

- Always, $ME = critical\ value \times SE\ of\ point\ estimate$
- In this case the point estimate is $\bar{x} - \bar{y}$
- Since the population σ for the difference is unknown, the critical value is $t^\star$. We will define $df = \min(n_x - 1,\ n_y - 1)$ which is wrong (but in the conservative direction).
- So the only new concept is the standard error of the difference between two means...

$$SE = \sqrt{Var(\bar{x} - \bar{y})} = \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} \approx \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$$

Let's put things in context

Calculate the standard error of the average difference between the number of hours worked per week by college graduates and those with a HS degree or lower.

	$\bar{x}$	s	n
college or higher	41.8	15.14	505
hs or lower	39.4	15.12	667

$$SE = \sqrt{\frac{s_c^2}{n_c} + \frac{s_{hs}^2}{n_{hs}}} = \sqrt{\frac{15.14^2}{505} + \frac{15.12^2}{667}} = 0.89$$

Confidence interval for the difference (cont.)

Estimate (using a 95% confidence interval) the average difference between the number of hours worked per week by Americans with a college degree and those with a high school degree or lower.

$$\bar{x}_c = 41.8 \qquad \bar{x}_{hs} = 39.4 \qquad SE = 0.89$$

$$df = \min(505 - 1,\ 667 - 1) = 504 \qquad t^\star_{df=504} = 1.96$$

$$\begin{aligned}(\bar{x}_c - \bar{x}_{hs}) \pm t^\star \times SE_{(\bar{x}_c - \bar{x}_{hs})} &= (41.8 - 39.4) \pm 1.96 \times 0.89 \\ &= 2.4 \pm 1.74 \\ &= (0.66, 4.14)\end{aligned}$$

We are 95% confident that college grads work on average between 0.66 and 4.14 more hours per week than those with a HS degree or lower.

Setting the hypotheses

If instead we wanted to conduct a hypothesis, what would the hypotheses be for testing if there is a difference between the average number of hours worked per week by college graduates and those with a HS degree or lower?

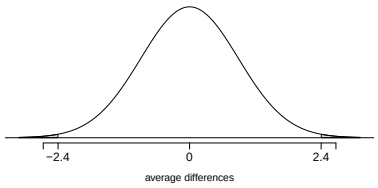
H_0: $\mu_c = \mu_{hs} \rightarrow \mu_c - \mu_{hs} = 0$
There is no difference in the average number of hours worked per week by college graduates and those with a HS degree or lower. Any observed difference between the sample means is due to natural sampling variation (chance).

H_A: $\mu_c \neq \mu_{hs} \rightarrow \mu_c - \mu_{hs} \neq 0$
There is a difference in the average number of hours worked per week by college graduates and those with a HS degree or lower.

Calculating the test-statistic and the p-value

H_0: $\mu_c - \mu_{hs} = 0$
H_A: $\mu_c - \mu_{hs} \neq 0$

$$\bar{x}_c - \bar{x}_{hs} = 2.4,\ SE_{\bar{x}_c - \bar{x}_{hs}} = 0.89$$



$$\begin{aligned}T &= \frac{(\bar{x}_c - \bar{x}_{hs}) - (\mu_c - \mu_{hs})}{SE} \\ &= \frac{2.4}{0.89} = 2.70 \\ P(T > 2.70) &= 1 - 0.9965 = 0.0035 \\ p - value &= 2 \times P(T > 2.70) = 0.007\end{aligned}$$

Reject H_0 - the data provide convincing evidence of a difference between the average number of hours worked per week by college graduates and those with a HS degree or lower.

Inference using difference of two means

- For sufficiently large sample size (of both groups), the distribution of the difference between the sample means has a $SE \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1}}$ and follows a T distribution with $df = \min(n_1 - 1,\ n_2 - 1)^*$.
- Conditions:
  - independence within groups (often verified by a random sample, and if sampling without replacement, $n < 10\%$ of population)
  - independence between groups
  - Sample sizes (n_1 and n_2) large enough relative to skew and or think/thin tails in either sample.
- Hypothesis testing:

$$T = \frac{\text{point estimate} - \text{null value}}{SE}$$

- Confidence interval:

$$\text{point estimate} \pm t^\star \times SE$$