Chapter 7

Inference for numerical data[1]

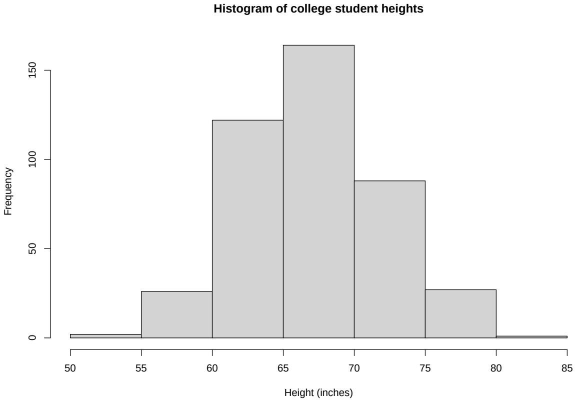
Department of Mathematics & Statistics
North Carolina A&T State University

[1] These notes use content from OpenIntro Statistics Slides by Mine Cetinkaya-Rundel.

One-sample means with the t distribution

Heights

- According to the CDC, the mean height of U.S. adults ages 20 and older is about 66.5 inches (69.3 inches for males, and 63.8 inches for females).
- In our sample data, we have a sample of 430 college students from a single college.



Summary statistics

n	$\bar{x}$	s	minimum	maximum
430	67.09	4.86	53.78	83.21

Objective: We would like to investigate if the mean height of students at this college is significantly different than 66.5 inches.

Conditions

- **Independence:** We are told to assume that cases (rows) are independent.

Conditions

- **Independence:** We are told to assume that cases (rows) are independent.

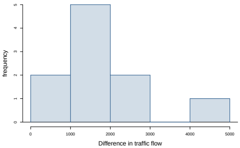
- **Sample size / skew:**

Conditions

- **Independence:** We are told to assume that cases (rows) are independent.

- **Sample size / skew:**

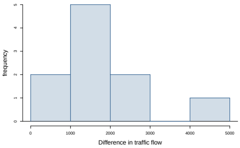
- The sample distribution does not appear to be extremely skewed, but it's very difficult to assess with such a small sample size. We might want to think about whether we would expect the population distribution to be skewed or not — probably not, it should be equally likely to have days with lower than average traffic and higher than average traffic.



- We do not know σ and n is too small to assume s is a reliable estimate for σ.

Conditions

- **Independence:** We are told to assume that cases (rows) are independent.
- **Sample size / skew:**
- The sample distribution does not appear to be extremely skewed, but it's very difficult to assess with such a small sample size. We might want to think about whether we would expect the population distribution to be skewed or not — probably not, it should be equally likely to have days with lower than average traffic and higher than average traffic.



- We do not know σ and n is too small to assume s is a reliable estimate for σ.

So what do we do when the sample size is small?

Review: what purpose does a large sample serve?

As long as observations are independent, and the population distribution is not extremely skewed, a large sample would ensure that...

- The sampling distribution of the mean is nearly normal.
- The estimate of the standard error, as $\frac{s}{\sqrt{n}}$, is reliable.

The normality condition

- The CLT, which states that sampling distributions will be nearly normal, holds true for **any** sample size as long as the population distribution is nearly normal.

The normality condition

- The CLT, which states that sampling distributions will be nearly normal, holds true for **any** sample size as long as the population distribution is nearly normal.
- While this is helpful special case, it's inherently difficult to verify normality in small data sets.

The normality condition

- The CLT, which states that sampling distributions will be nearly normal, holds true for **any** sample size as long as the population distribution is nearly normal.
- While this is helpful special case, it's inherently difficult to verify normality in small data sets.
- We should exercise caution when verifying the normality condition for small samples. It is important to not only examine the data but also think about where the data come from.
  - For example, ask: would I expect this distribution to be symmetric, and am I confident that outliers are rare?

The t distribution

- When the population standard deviation is unknown (almost always), the uncertainty of the standard error estimate is addressed by using a new distribution: the ***t* distribution**.

The t distribution

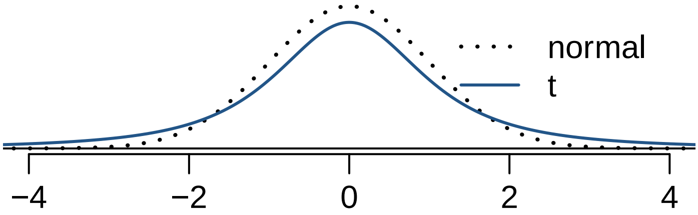
- When the population standard deviation is unknown (almost always), the uncertainty of the standard error estimate is addressed by using a new distribution: the ***t* distribution**.
- This distribution also has a bell shape, but its tails are **thicker** than the normal model's.

The t distribution

- When the population standard deviation is unknown (almost always), the uncertainty of the standard error estimate is addressed by using a new distribution: the ***t* distribution**.
- This distribution also has a bell shape, but its tails are **thicker** than the normal model's.
- Therefore observations are more likely to fall beyond two SDs from the mean than under the normal distribution.

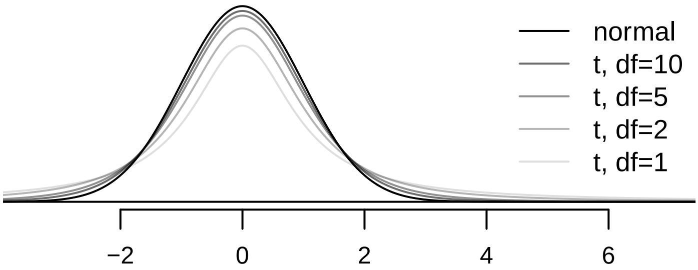
The *t* distribution

- When the population standard deviation is unknown (almost always), the uncertainty of the standard error estimate is addressed by using a new distribution: the ***t* distribution**.
- This distribution also has a bell shape, but its tails are **thicker** than the normal model's.
- Therefore observations are more likely to fall beyond two SDs from the mean than under the normal distribution.
- Extra thick tails are helpful for resolving our problem with a less reliable estimate the standard error (since n is small).



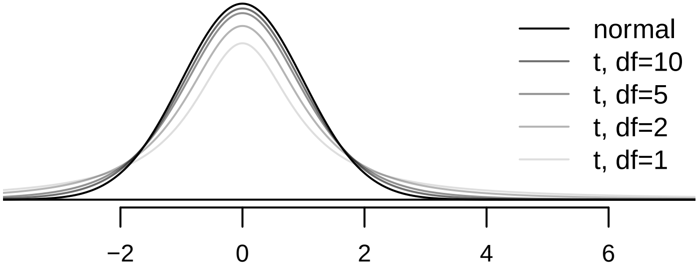
The t distribution

- Always centered at zero, like the standard normal (z) distribution.
- Has a single parameter: **degrees of freedom** ($\mathbf{df}$).



The t distribution

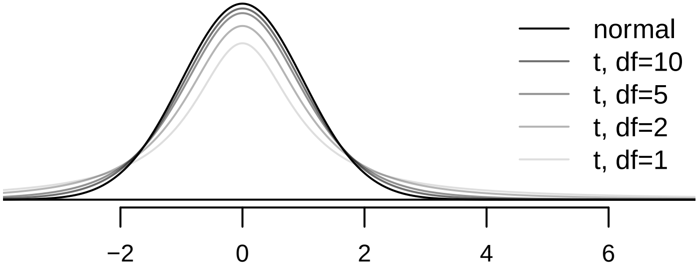
- Always centered at zero, like the standard normal (z) distribution.
- Has a single parameter: **degrees of freedom** ($\mathbf{df}$).



What happens to shape of the t distribution as df increases?

The t distribution

- Always centered at zero, like the standard normal (z) distribution.
- Has a single parameter: **degrees of freedom** ($\mathbf{df}$).



What happens to shape of the t distribution as df increases?

Approaches normal.

Back to the student heights survey

n	$\bar{x}$	s	minimum	maximum
430	67.09	4.86	53.78	83.21

Objective: We would like to investigate if the mean height of students at this college is significantly different than 66.5 inches.

Hypotheses

What are the hypotheses for testing for the mean of college student heights being different from 67 inches?

A) $H_0 : \mu = 66.5$
$H_A : \mu \neq 66.5$

B) $H_0 : \mu = 66.5$
$H_A : \mu > 66.5$

C) $H_0 : \mu = 66.5$
$H_A : \mu < 66.5$

D) $H_0 : \mu \neq 66.5$
$H_A : \mu > 66.5$

Hypotheses

What are the hypotheses for testing for the mean of college student heights being different from 66.5 inches?

A) $H_0 : \mu = 66.5$
$H_A : \mu \neq 66.5$

B) $H_0 : \mu = 66.5$
$H_A : \mu > 66.5$

C) $H_0 : \mu = 66.5$
$H_A : \mu < 66.5$

D) $H_0 : \mu \neq 66.5$
$H_A : \mu > 66.5$

Finding the test statistic

The test statistic for inference on sample mean is the T statistic with $df = n - 1$.

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

Finding the test statistic

The test statistic for inference on sample mean is the T statistic with $df = n - 1$.

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

in context...

$$\text{point estimate} = \bar{x} = 67.09$$

Finding the test statistic

The test statistic for inference on sample mean is the T statistic with $df = n - 1$.

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

in context...

$$\text{point estimate} = \bar{x} = 67.09$$

$$SE = \frac{s}{\sqrt{n}} = \frac{4.86}{\sqrt{430}} = 0.234$$

Finding the test statistic

The test statistic for inference on sample mean is the T statistic with $df = n - 1$.

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

in context...

$$\text{point estimate} = \bar{x} = 67.09$$

$$SE = \frac{s}{\sqrt{n}} = \frac{4.86}{\sqrt{430}} = 0.234$$

$$T = \frac{67.09 - 66.5}{0.234} = 2.52$$

Finding the test statistic

The test statistic for inference on sample mean is the T statistic with $df = n - 1$.

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

in context...

$$\text{point estimate} = \bar{x} = 67.09$$

$$SE = \frac{s}{\sqrt{n}} = \frac{4.86}{\sqrt{430}} = 0.234$$

$$T = \frac{67.09 - 66.5}{0.234} = 2.52$$

$$df = 430 - 1 = 429$$

Note: Null value is 66.5 because in the null hypothesis we set $\mu = 66.5$.

Finding the p-value

- The p-value is, once again, calculated as the area tail area under the t distribution.

Finding the p-value

- The p-value is, once again, calculated as the area tail area under the t distribution.
- Using a web app: https://gallery.shinyapps.io/dist_calc/

Finding the p-value

- The p-value is, once again, calculated as the area tail area under the t distribution.
- Using a web app:
  https://gallery.shinyapps.io/dist_calc/
- Or when these aren't available, we can use a t-table.

Conclusion of the test

What is the conclusion of this hypothese test?

Conclusion of the test

What is the conclusion of this hypothese test?

We saw that the p-value was extremely low. Thus, we reject the null hypothesis. Based on the p-value, we conclude that the survey provide strong evidence that the mean of the college students height is different from the mean height of U.S. adults over 20.

What is the difference?

- We concluded that there is a difference in the mean heights of the college students compared to the mean height of U.S. adults

What is the difference?

- We concluded that there is a difference in the mean heights of the college students compared to the mean height of U.S. adults
- But it would be more interesting to find out what exactly this difference is.

What is the difference?

- We concluded that there is a difference in the mean heights of the college students compared to the mean height of U.S. adults
- But it would be more interesting to find out what exactly this difference is.
- We can use a confidence interval to estimate this difference.

Confidence interval for a sample mean

- Confidence intervals are always of the form

$$\text{point estimate} \pm ME$$

Confidence interval for a sample mean

- Confidence intervals are always of the form

$$\text{point estimate} \pm ME$$

- ME is always calculated as the product of a critical value and SE.

Confidence interval for a sample mean

- Confidence intervals are always of the form

$$\text{point estimate} \pm ME$$

- ME is always calculated as the product of a critical value and SE.

- $ME = t^* \times SE$

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

Finding the critical $t(t^*)$

- We want to find the 95% confidence interval.

Constructing a CI for a small sample mean

Which of the following is the correct calculation of a 95% confidence interval for the heights of the college students?

$$\bar{x} = 67.09 \quad s = 4.86 \quad n = 430 \quad SE = 0.234$$

A) $66.5 \pm 1.96 \times 0.234$

B) $67.09 \pm 1.97 \times 0.234$

C) $67.09 \pm -2.26 \times 0.234$

D) $66.5 \pm 2.26 \times 4.86$

Constructing a CI for a small sample mean

Which of the following is the correct calculation of a 95% confidence interval for the heights of the college students?

$$\bar{x} = 67.09 \quad s = 4.86 \quad n = 430 \quad SE = 0.234$$

A) $66.5 \pm 1.96 \times 0.234$

B) $67.09 \pm 1.97 \times 0.234 \rightarrow (66.63, 67.55)$

C) $67.09 \pm -2.26 \times 0.234$

D) $66.5 \pm 2.26 \times 4.86$

Synthesis

Does the conclusion from the hypothesis test agree with the findings of the confidence intereval?

Synthesis

Does the conclusion from the hypothesis test agree with the findings of the confidence intereval?

Yes, the hypothesis test found a significant difference, and the CI does not contain the null value of 66.5.

Recap: Inference using the t-distribution

- If σ is unknown, use t-distribution with $SE = \frac{s}{\sqrt{n}}$.

Recap: Inference using the t-distribution

- If σ is unknown, use t-distribution with $SE = \frac{s}{\sqrt{n}}$.
- Conditions:
  - Independence of observations (often verified by random sample, and if sampling w/o replacement, n $<$ 10% of population).
  - No extreme skew.

Recap: Inference using the t-distribution

- If σ is unknown, use t-distribution with $SE = \frac{s}{\sqrt{n}}$.
- Conditions:
  - Independence of observations (often verified by random sample, and if sampling w/o replacement, n $<$ 10% of population).
  - No extreme skew.
- Hypothesis testing:

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}, \text{where } df = n - 1$$

Recap: Inference using the t-distribution

- If σ is unknown, use t-distribution with $SE = \frac{s}{\sqrt{n}}$.
- Conditions:
  - Independence of observations (often verified by random sample, and if sampling w/o replacement, n $<$ 10% of population).
  - No extreme skew.
- Hypothesis testing:

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}, \text{where } df = n - 1$$

- Confidence interval: point estimate $\pm\, t^{*}_{df} \times SE$