Lecture 10 - Confidence Intervals for Sample Means

Sta102/BME102

October 5, 2015

Colin Rundel

Confidence Intervals in the Real World

A small problem

Lets assume we are collecting a large sample (n=200) from a population and are measuring some numeric characteristic that has distribution D, where $E(D) = \mu$ and $Var(X) = \sigma^2$ (e.g. blood pressure of high school athletes).

We want to make some inference about the population mean, to do this we can construct a 95% confidence interval based on our observed sample average:

$$\begin{aligned} CI_{95\%} &= \bar{X} \pm Z^\star \, SE \\ &= \bar{X} \pm 1.96 \, \frac{\sigma}{\sqrt{n}} \end{aligned}$$

Anyone see a problem here?

Missing σ

When working with real samples the population standard deviation (σ) is almost never known, we address this by plugging in the sample standard deviation when calculating the standard error. However, when we do this it changes the sampling distribution.

- We estimate the standard error using the sample standard deviation, this adds uncertainty to inference process.

Missing σ

When working with real samples the population standard deviation (σ) is almost never known, we address this by plugging in the sample standard deviation when calculating the standard error. However, when we do this it changes the sampling distribution.

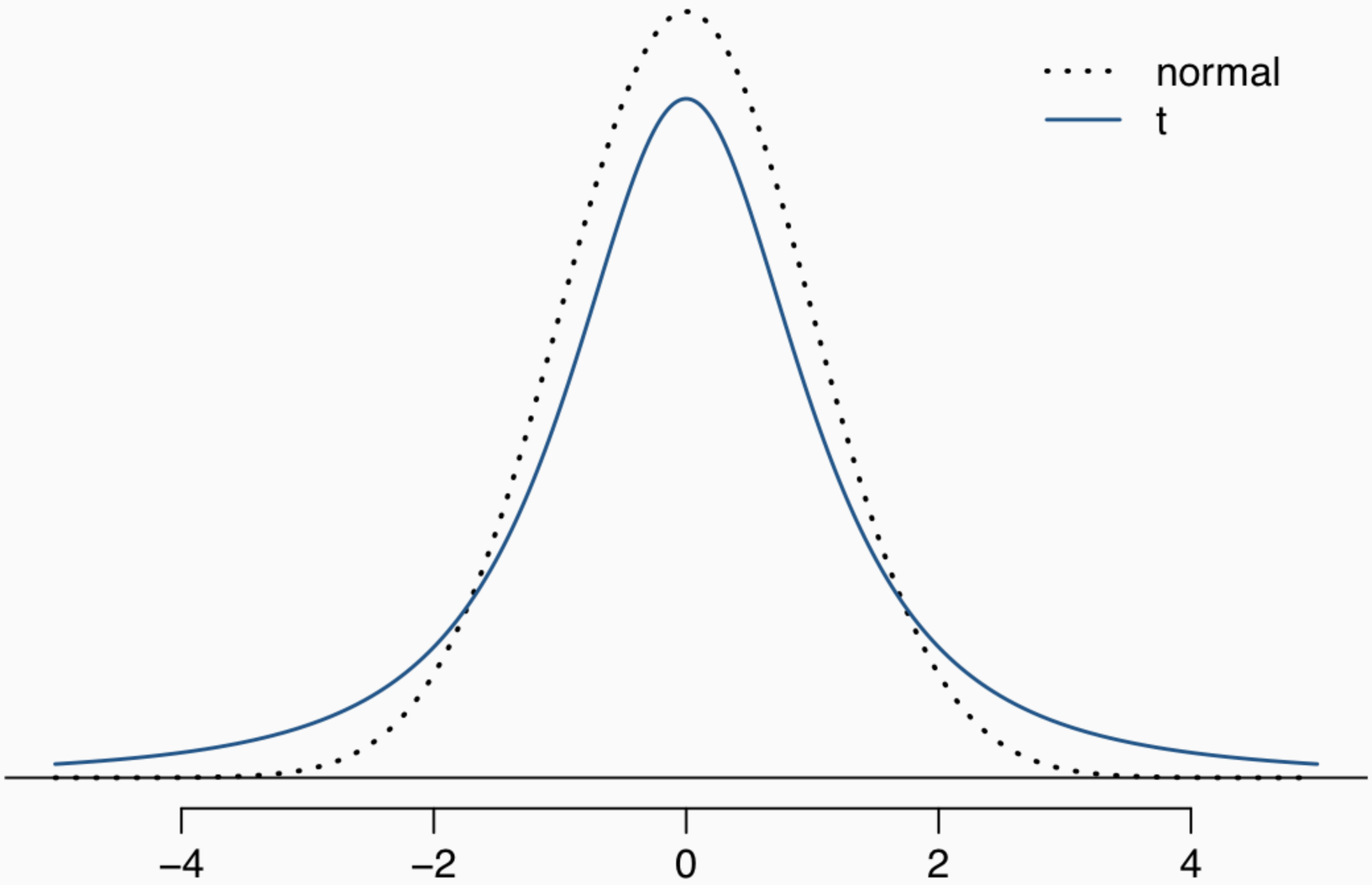
- We estimate the standard error using the sample standard deviation, this adds uncertainty to inference process.
- Our new sampling distribution is still symmetric and roughly bell shaped, but its tails are *thicker* than the normal distribution.

Missing σ

When working with real samples the population standard deviation (σ) is almost never known, we address this by plugging in the sample standard deviation when calculating the standard error. However, when we do this it changes the sampling distribution.

- We estimate the standard error using the sample standard deviation, this adds uncertainty to inference process.
- Our new sampling distribution is still symmetric and roughly bell shaped, but its tails are *thicker* than the normal distribution.
- Observations are more likely to fall beyond two SDs from the mean than with the normal distribution.

t distribution



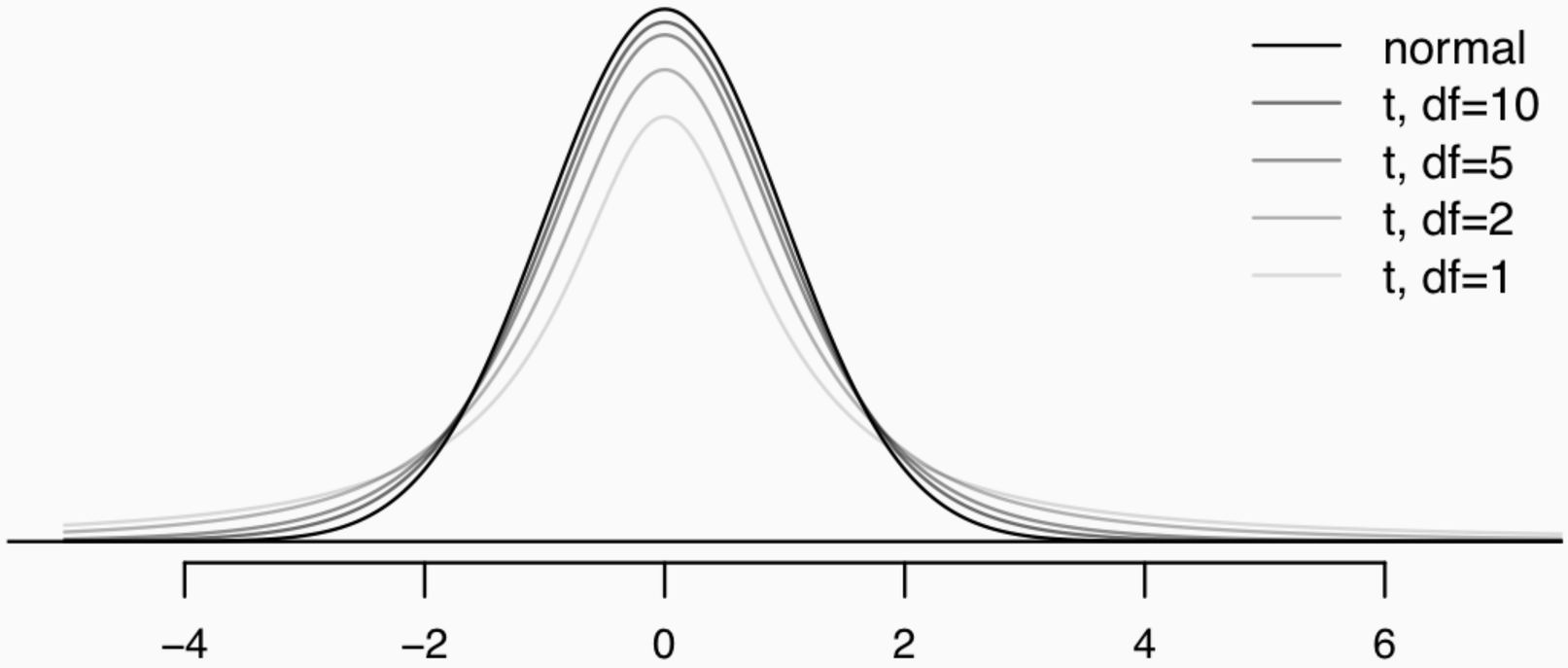
History of the t distribution

First described by William Gosset ...

- Oxford Graduate with a degree in Chemistry and Mathematics
- Hired by the Guinness Brewery in 1899
- Spent 1906 - 1907 studying with Karl Pearson
- Published "The probable error of a mean" in 1908 under the pseudonym "A. Student"
- Much of his work was promoted by R.A. Fisher

Properties of the t distribution

- is centered at zero*, like the standard normal (Z) distribution.
- has a single parameter, *degrees of freedom* (*df*), which dictates the thickness of the tails.



- as *df* increases the t distribution converges to the Z distribution.

Finding probabilities

As before we can find any probability we are interested by knowing how to calculate the area under the tail of the t distribution. For example, if we want to know $P(T_{df=19} > 1.16)$ then we can use:

Finding probabilities

As before we can find any probability we are interested by knowing how to calculate the area under the tail of the t distribution. For example, if we want to know $P(T_{df=19} > 1.16)$ then we can use:

- Using R:

```
1-pt(1.16,df=19)
```

```
## [1] 0.1302092
```

Finding probabilities

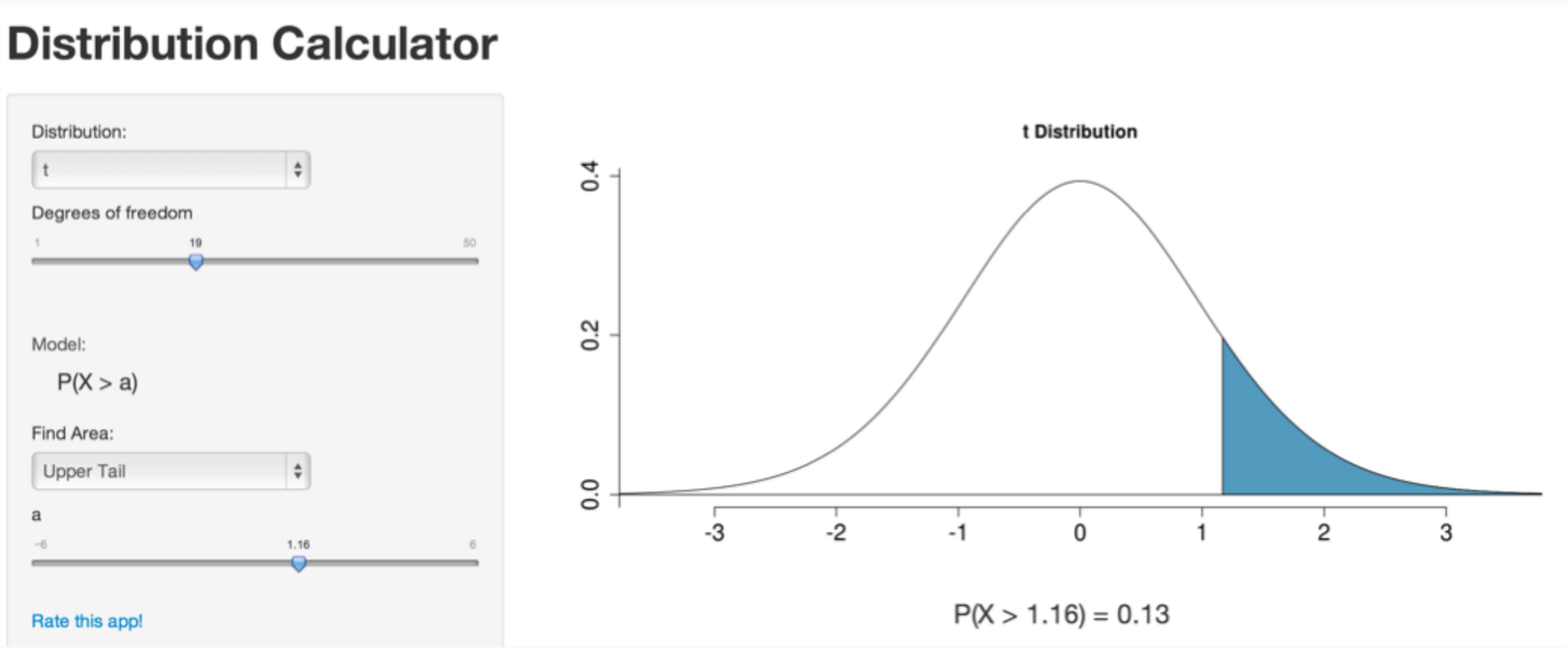
As before we can find any probability we are interested by knowing how to calculate the area under the tail of the t distribution. For example, if we want to know $P(T_{df=19} > 1.16)$ then we can use:

- Using R:

```
1-pt(1.16,df=19)
```

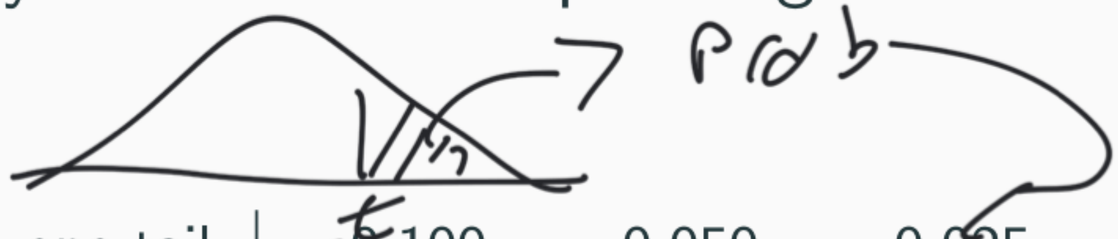
```
## [1] 0.1302092
```

- Using a web applet (`http://bit.ly/dist_calc`):



Finding Probabilities - t table

Locate the T value on the appropriate df row, obtain the probability from the corresponding column heading (one or two tail).



one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	1	3.08	6.31	12.71	31.82	63.66
	2	1.89	2.92	4.30	6.96	9.92
	3	1.64	2.35	3.18	4.54	5.84
	⋮	⋮	⋮	⋮	⋮	
	17	1.33	1.74	2.11	2.57	2.90
	18	1.33	1.73	2.10	2.55	2.88
	19	1.33	1.73	2.09	2.54	2.86
	20	1.33	1.72	2.09	2.53	2.85
	⋮	⋮	⋮	⋮	⋮	
	400	1.28	1.65	1.97	2.34	2.59
	500	1.28	1.65	1.96	2.33	2.59
	∞	1.28	1.64	1.96	2.33	2.58

Finding probabilities - upper tail

Using the table below find:

$$P(T_{df=19} \geq 1.16) > 0.1$$

$$0.025 \quad P(T_{df=19} > 2.25) > 0.01$$

one tail	0.100	0.050	0.025	0.010	0.005
two tails	0.200	0.100	0.050	0.020	0.010
df 17	1.33	1.74	2.11	2.57	2.90
18	1.33	1.73	2.10	2.55	2.88
19	1.33	1.73	2.09	2.54	2.86
20	1.33	1.72	2.09	2.53	2.85

Finding probabilities - upper tail

Using the table below find:

$0.05 > P(T_{df=19} < -2) =$ $P(T > 2) > 0.025$

one tail	0.100	0.050	0.025	0.010	0.005
two tails	0.200	0.100	0.050	0.020	0.010
df 17	1.33	1.74	2.11	2.57	2.90
18	1.33	1.73	2.10	2.55	2.88
19	1.33	1.73	2.09	2.54	2.86
20	1.33	1.72	2.09	2.53	2.85

Finding probabilities - two tails

Using the table below find:

$$0.2 > P(T_{df=19} < -1.5 \text{ or } T_{df=19} > 1.5) > 0.1$$

	one tail	0.100	0.050	0.025	0.010	0.005
	two tails	0.200	0.100	0.050	0.020	0.010
df	17	1.33	1.74	2.11	2.57	2.90
	18	1.33	1.73	2.10	2.55	2.88
	19	1.33	1.73	2.09	2.54	2.86
	20	1.33	1.72	2.09	2.53	2.85

CLT vs. t

From the Central Limit Distribution we have,

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Since σ is unknown we must use s which results in the following

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

Implications of t distribution for Confidence intervals

Confidence intervals are always of the form

$$\text{point estimate} \pm CV \times SE$$

Implications of t distribution for Confidence intervals

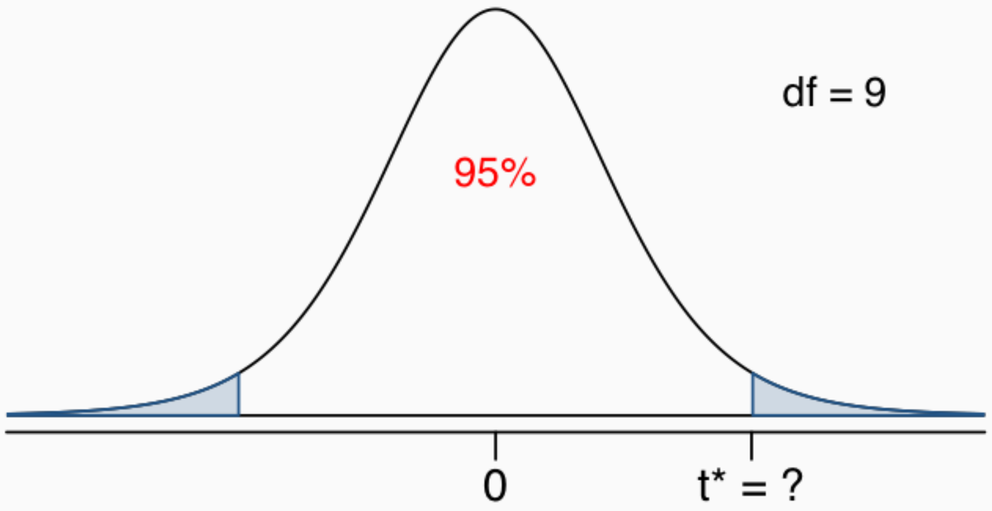
Confidence intervals are always of the form

$$\text{point estimate} \pm CV \times SE$$

If our point estimate is a sample mean and σ is unknown, then our sample mean follows a t distribution (and not a Z distribution), the critical value is then given by $t^\star_{df}$ (as opposed to a $Z^\star$) and the SE is $s/\sqrt{n}$ (and not $\sigma/\sqrt{n}$).

$$\bar{X} \pm t^\star_{df} \times \frac{s}{\sqrt{n}}$$

Finding the critical t ($t^\star$)



$n = 10$, $df = 10 - 1 = 9$

$t^\star$ is at the intersection of row $df = 9$ and two tails column 0.05.

	one tail	0.100	0.050	0.025	0.010	0.005
	two tails	0.200	0.100	0.050	0.020	0.010
df	6	1.44	1.94	2.45	3.14	3.71
	7	1.41	1.89	2.36	3.00	3.50
	8	1.40	1.86	2.31	2.90	3.36
	9	1.38	1.83	2.26	2.82	3.25
	10	1.37	1.81	2.23	2.76	3.17

Constructing a CI

We would like to calculate a 95% confidence interval for the average rental price of an apartment in Durham. We sample craigslist and find

$$\begin{aligned}\text{Rent} = \{&625, 733, 895, 929, 775, 1349, 599, 749, 1020, 799,\\ &705, 665, 1282, 1143, 1209, 500, 1495, 1076, 975, 879\}\end{aligned}$$

Constructing a CI

We would like to calculate a 95% confidence interval for the average rental price of an apartment in Durham. We sample craigslist and find

$$\text{Rent} = \{625, 733, 895, 929, 775, 1349, 599, 749, 1020, 799,$$
$$705, 665, 1282, 1143, 1209, 500, 1495, 1076, 975, 879\}$$

$\bar{X} = 920.1 \qquad s = 271 \qquad n = 20 \qquad SE = s/\sqrt{n} = 60.6$

$\bar{x} \pm t^*_{df=19} SE$

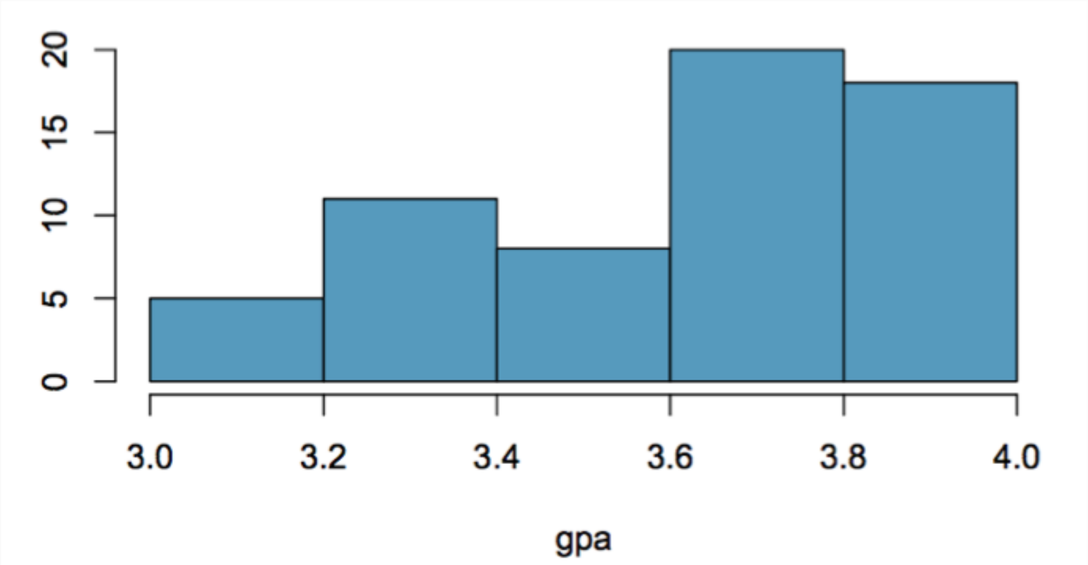
$920.1 \pm 2.09\ (60.6)$

$(793.45,\ 1046.75)$

Confidence Intervals as Inference

Example - Grade Inflation

In 2001 the average GPA of students at Duke University was 3.37. Last semester 63 introductory statistics students reported their GPA on an in class survey. The mean was 3.58, and the standard deviation 0.53. A histogram of the data is shown below.



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 and a half?

$\bar{x} = 3.58 \quad s = 0.53 \quad n = 63$

$$\bar{x} \pm t^*_{df=62} \; s/\sqrt{n}$$

$$3.58 \pm t^*_{df=60} \left(\frac{0.53}{\sqrt{63}}\right)$$

$$\| \; 2$$

$(3.45, 3.71)$

Claim $\mu = 3.37$ is not plausible.

← PA ↑ since 2001

Example - Fair Dice

Imagine you are going to roll a die 100 times and record the average value of the rolls, under what circumstances should you conclude that the die is not fair at a 95% confidence level? Hint - be careful with your choice of critical value.

$\mu = 3.5$

$E\left(\frac{D_1 + D_2 + \dots + D_{100}}{100}\right)$

$= 3.5$

D - Face value of a Gerked die

$E(D) = 3.5 \quad Var(D) = 2.92$

$Var\left(\frac{D_1 + D_2 + \dots + D_{100}}{100}\right) = \frac{Var(D)}{100}$

$\bar{X} \sim N\left(\mu = 3.5, \sigma^2 = \frac{2.92}{100}\right)$

Expect CI to contain $\mu = 3.5$

$\Rightarrow$ not fair if CI doesn't contain 3.5

$$\bar{X} \pm Z^* SE$$

$$\left(\bar{X} - 1.96\sqrt{\frac{2.92}{100}},\ \bar{X} + 1.96\sqrt{\frac{2.92}{100}}\right)$$

$$\bar{x} - 1.96\left(\sqrt{\frac{2.92}{100}}\right) \leq 3.5 \Rightarrow 3.165$$

$$\bar{x} + 1.96\left(\sqrt{\frac{2.92}{100}}\right) \geq 3.5 \Rightarrow 3.835$$

Example - Z vs t

Your friend has collected some data as part of a summer REU - they collected tadpoles from a local different stream and measured their lengths. From the stream they were able collect 50 tadpoles which had an average length 2.3 cm and a standard deviation of 0.2 cm.

They argue that since it is well know that the distribution of tadpole lengths is normal they should be able to use the Z distribution when constructing their confidence intervals for the average lengths. Are they correct? If not, how serious a mistake are they making? (Construct the CIs both ways for both steams and compare)

Use t, σ unknown

Correct t

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

$$2.3 \pm 2.01 \frac{0.2}{\sqrt{50}} = (2.24\cancel{9}, 2.35\cancel{5})$$

Incorrect z

$$\bar{x} \pm 2^* \frac{\sigma}{\sqrt{n}}$$

$$2.3 \pm 1.96 \frac{0.2}{\sqrt{50}} = (2.24\cancel{8}, 2.355)$$

Recap: Inference using CIs for sample means

If σ is unknown, then $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ has a t distribution with $df = n - 1$ when the CLT holds.

Recap: Inference using CIs for sample means

If σ is unknown, then $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ has a t distribution with $df = n - 1$ when the CLT holds.

Conditions (same as CLT):

- independence of observations (often verified by a random sample, and if sampling without replacement, $n < 10\%$ of population)
- sample size is large or population not overly skewed or heavy/light tailed

Recap: Inference using CIs for sample means

If σ is unknown, then $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ has a t distribution with $df = n-1$ when the CLT holds.

Conditions (same as CLT):

- independence of observations (often verified by a random sample, and if sampling without replacement, $n < 10\%$ of population)
- sample size is large or population not overly skewed or heavy/light tailed

Confidence interval:

$$\bar{X} \pm t^\star_{df} \frac{s}{\sqrt{n}}, \text{ where } df = n-1$$