Unit 6 Day 2: Exercise 21

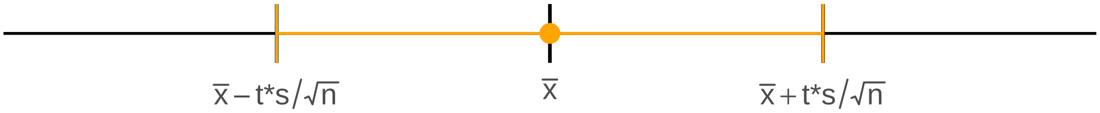
Name:

Reminder of Notation: $\bar{x}$ is the sample mean, s is the sample standard deviation, n is the sample size

Goal: Construct a 95% confidence interval for μ.

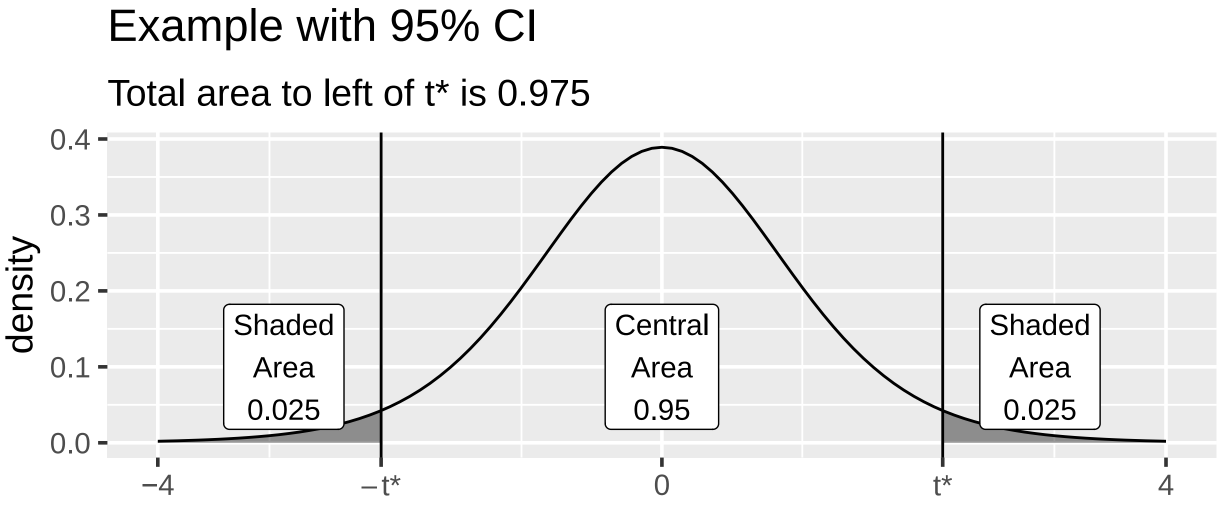
Solution:

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



Interpretation:

- The **margin of error** is $t^* \frac{s}{\sqrt{n}}$: the amount we add and subtract from $\bar{x}$.
- The **critical value** is t^*: 0.975th quantile of the t_{n-1} distribution.



In R, to look up t^*:

```r
qt(0.975, df = 10) # For a 95% CI, sample size is n = 11
```

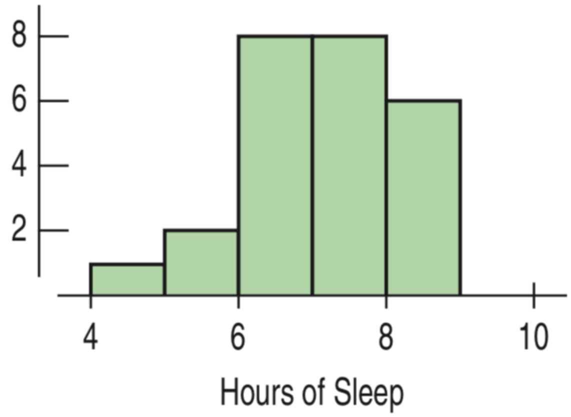
```
## [1] 2.228139
```

Important things:

- For a 95% CI, the first argument to `qt` is 0.975, not 0.95!
- The second argument to `qt` is $n - 1$.

Example 1: College student sleep

I have data on the number of hours that 25 students slept and a histogram of the 25 observed amounts that students slept.



6	6	6
7	6	7
6	6	8
7	8	5
6	8	8
7	8	7
7	5	6
7	4	7
8		

I have computed that

- sample mean $\bar{x} = 6.64$ hours
- sample standard deviation $s = 1.075$ hours

Question: What can we say about the mean amount of sleep that college students get? Let's build a 95% confidence interval for the mean amount that college students sleep in a night.

(a) What is the population parameter of interest?

(b) Check the conditions for inference with these data

- Randomization Condition: do data come from a random sample or suitably randomized experiment?

- Nearly normal condition: do data come from a distribution that is unimodal and symmetric.

(c) Construct the 95% onfidence interval

(d) Interpret the confidence interval in the proper context

(e) Find a 90% confidence interval, and interpret the interval in the proper context

Interpretation of confidence intervals:

Here are some things you shouldn't say:

- Don't say, "90% of all students sleep between 6.272 and 7.008 hours per night." The confidence interval is about the mean sleep, not about the sleep of individual students.
- Don't say, "We are 90% confident that a randomly selected student will sleep between 6.272 and 7.008 hours per night." This false interpretation is also about individual stu- dents rather than about the mean. We are 90% confident that the mean amount of sleep is between 6.272 and 7.008 hours per night.
- Don't say, "The mean amount students sleep is 6.64 hours 90% of the time." That's about means, but still wrong. It implies that the true mean varies, when in fact it is the confidence interval that would have been different had we gotten a different sample.
- Finally, don't say, "90% of all samples will have mean sleep between 6.272 and 7.008 hours per night." That statement suggests that this interval somehow sets a standard for every other interval. In fact, this interval is no more (or less) likely to be correct than any other. You could say that 90% of all possible samples will produce intervals that actually do contain the true mean sleep. (The problem is that, because we'll never know where the true mean sleep really is, we can't know if our sample was one of those 90%.)

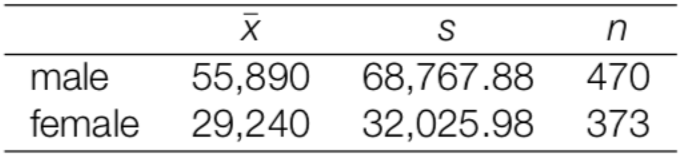
Do say:

"90% of intervals that could be found in this way would cover the true value." Or make it more personal and say, "I am 90% confident that the true mean amount that students sleep is between 6.272 and 7.008 hours per night."

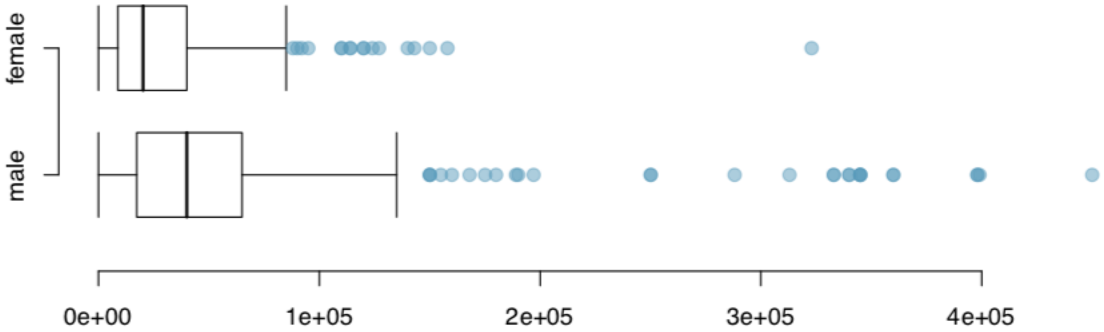
Remember: Our uncertainty is about the interval, not the true mean. The interval varies randomly. The true mean sleep is neither variable nor random —just unknown.

Example 2: Gender gap in salaries

Since 2005, the American Community Survey polls approximately 3.5 million households yearly. The following summarizes distribution of salaries of males and females from a random sample of individuals who responded to the 2012 ACS:



	$\bar{x}$	s	n
male	55,890	68,767.88	470
female	29,240	32,025.98	373



(a) What is the population parameter of interest?

(b) Check the conditions for inference with these data

- Randomization Condition: do data come from a random sample or suitably randomized experiment?

- Nearly normal condiion: do data come from a distribution that is unimodal and symmetric.

For samples of $n < 15$ in either group, you should not use these methods if the histogram or Normal probability plot shows severe skewness. For n's closer to 40, a mildly skewed histogram is OK, but you should remark on any outliers you find and not work with severely skewed data. When both groups are bigger than 40, the Central Limit Theorem starts to kick in no matter how the data are distributed, so the Nearly Normal Condition for the data matters less.

- Independent Groups Assumption: are the two groups independent?

(c) Construct the 95% onfidence interval

The formula for making the confidence interval is given below

$$\bar{x}_1 - \bar{x}_2 \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where t^* is the 0.975 quartile of the t-distribution with a degree of freedom

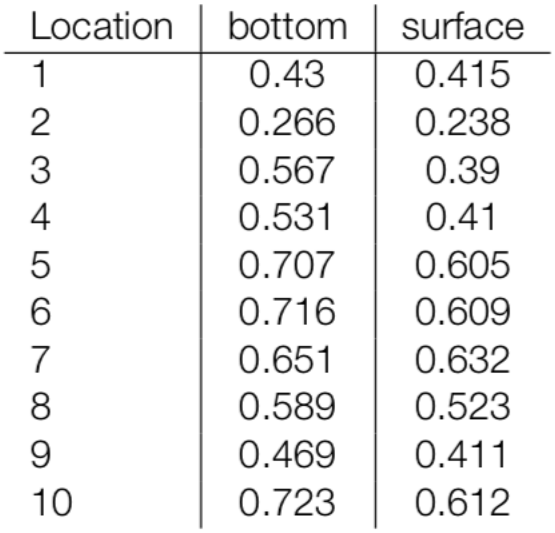
$$df = \min(n_1 - 1, n_2 - 2).$$

(d) Interpret the confidence interval in the proper context

(e) Find a 90% confidence interval, and interpret the interval in the proper context

Example 3: Zinc in water

Trace metals in drinking water affect the flavor and an unusually high concentration can pose a health hazard. Ten pairs of data were taken measuring zinc concentration in bottom water and surface water at 10 randomly sampled locations.



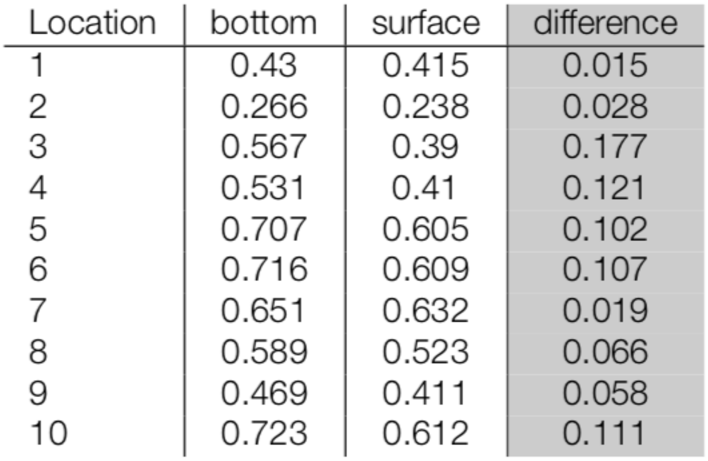
Location	bottom	surface
1	0.43	0.415
2	0.266	0.238
3	0.567	0.39
4	0.531	0.41
5	0.707	0.605
6	0.716	0.609
7	0.651	0.632
8	0.589	0.523
9	0.469	0.411
10	0.723	0.612

(a) What is the population parameter of interest?

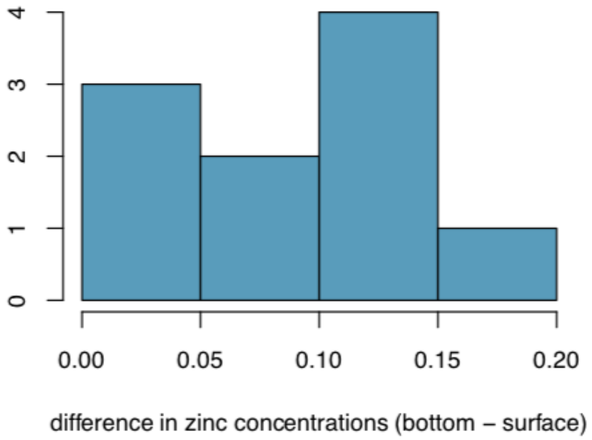
(b) How is this example different from Example 2?

This is an example of **paired data**:

- We have two measurements on each location (these are **not independent**!)
- We are interested in the **difference** between these measurements
- These **differences are independent** across different locations



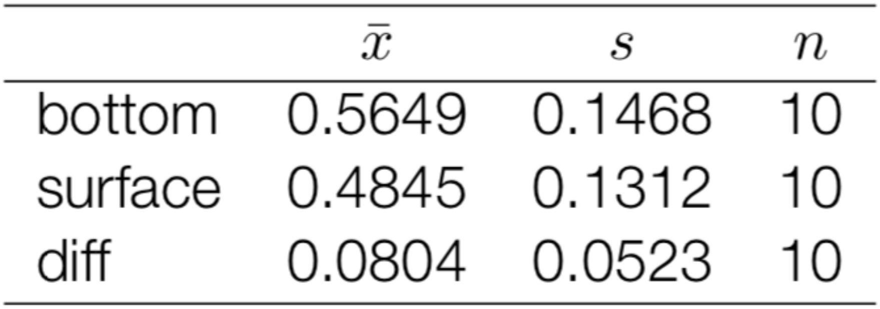
Location	bottom	surface	difference
1	0.43	0.415	0.015
2	0.266	0.238	0.028
3	0.567	0.39	0.177
4	0.531	0.41	0.121
5	0.707	0.605	0.102
6	0.716	0.609	0.107
7	0.651	0.632	0.019
8	0.589	0.523	0.066
9	0.469	0.411	0.058
10	0.723	0.612	0.111



(c) Check the conditions for inference with these data

- Randomization Condition: do data come from a random sample or suitably randomized experiment?

- Nearly normal condition: do data come from a distribution that is unimodal and symmetric.



	$\bar{x}$	s	n
bottom	0.5649	0.1468	10
surface	0.4845	0.1312	10
diff	0.0804	0.0523	10

(d) Construct the 95% onfidence interval

The formula for making the confidence interval is given below

$$\bar{x}_{\text{diff}} \pm t^* \sqrt{\frac{s_{\text{diff}}^2}{n_{\text{diff}}}}$$

where t^* is the 0.975 quartile of the t-distribution with a degree of freedom

$$df = \min(n_{\text{diff}} - 1).$$

(e) Interpret the confidence interval in the proper context

(f) Find a 99% confidence interval, and interpret the interval in the proper context.