Week 6 Worksheet

This is a worksheet for the week of Week 6. You do not need to turn this in.
This worksheet follows the materials from Lecture Slide Set 6

Confidence Intervals for One Mean

When do we want to use confidence intervals?

- Without prior estimate for population parameter, use *confidence interval to estimate reasonable values for parameter*

How do we calculate confidence intervals?

Given a sample of *n* independent observations from an approximately normal distribution, the confidence interval for population mean μ is

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

- $\underline{x}$ is the sample mean, *s* is the sample standard deviation
- t^* is the critical value for a particular confidence level and degrees of freedom
- $df = n - 1$.

How do we interpret the confidence interval?

The confidence level tells us how confident we can be that the interval we construct contains the true population mean. We say, we estimate with 95% confidence that the population mean of xyz is between (confidence interval).

How do we find the critical value t^*?

Like with our confidence intervals for *p*, typical choices for confidence levels are 90%, 95%, and 99%. To find the t^* associated with 95% confidence intervals with sample size of 30, we need to look at the t distribution with $df = n - 1 = 30 - 1 = 29$. Since we want an area of 0.95 between $-t^*$ and t^*, that leaves 1 – 0.95 = 0.05 to be split between the two tails.

Deleted: 1

Formatted: Left

R Tutorial

Show students how to use the `qt()` function in R, which is like the `qnorm()` function for proportion problems. You may need to provide a review of `qnorm` before we get started in qt:

`pnorm():`
The pnorm() function gives us a way to compute probabilities when a variable has a normal distribution. The arguments you need to send to pnorm():

- q: the quantile (value on the axis) for the normal distribution
- mean: the mean of the normal distribution (μ)
- sd: the standard deviation of the normal distribution (σ)
- lower.tail: set to 'TRUE' initially, signifying that R will compute the probability to the LEFT of q; if you would like R to compute the probability to the right of q, set lower.tail to FALSE

Example:
What percent of middle-aged men have high cholesterol (levels above 240 mg/dl)?

```
pnorm(q = 240,
mean = 222,
sd = 37,
lower.tail = FALSE)
```

[1] 0.3133111

`qnorm():`
The qnorm() function gives us a way to find the values of a normally distributed variable when you are given a probability. The arguments you need to send to qnorm():

- p: the probability or area under the curve you want to find an x-axis
- value for
- mean: the mean of the normal distribution, defaults to 0
- sd: the standard deviation of the normal distribution, defaults to 1
- lower.tail: determines whether qnorm() finds the value of the variable with area p to its left or right. If lower.tail is set to 'TRUE' (the default), the area p is to the LEFT. If lower.tail is set to 'FALSE', the area p is to the RIGHT.

Deleted: 2

Formatted: Left

Example:
The blood cholesterol levels of men age 55 to 64 are approximately normal, with mean 222 milligrams per deciliter (mg/dl) and standard deviation 37 mg/dl. Men in the 95th percentile have blood cholesterol levels of what value?

```
qnorm(p = 0.95,
      mean = 222,
      sd = 37,
      lower.tail = TRUE)
```

```
qnorm(p = 0.05,
      mean = 222,
      sd = 37,
      lower.tail = FALSE)
```

Now for `qt()`:

- p: The probability to the left by default of the quantile we wish to find. If we want the probability to the right, we should tinker with lower.tail as specified below.
- df: This is the degrees of freedom for the sample; this week, we can find df by computing n − 1, where n is the sample size.
- lower.tail: By default, this argument is set to TRUE, meaning that we want the lower tail (i.e., to shade to the left). If we don't want the lower tail, and we actually want the upper tail (i.e., to shade to the right), we should set this to FALSE.

```
qt(p, df, lower.tail = TRUE)
```

```
qt(p = 0.975, df = 29, lower.tail = TRUE)
```

```
## [1] 2.04523
```

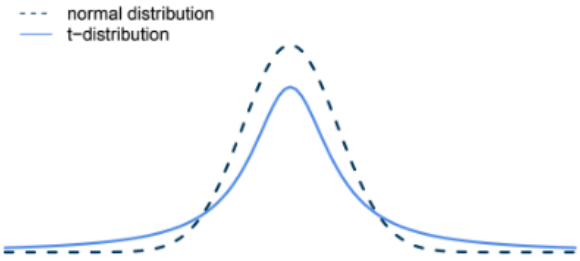
```
qt(p = 0.025, df = 29, lower.tail = FALSE)
```

```
## [1] 2.04523
```

Why are we using *t*? What happened to z*?*
We use the *t* distribution because it is the correct adjustment for the additional variability we have in the distribution of the sample mean due to estimating the population standard deviation

Deleted: 3
Formatted: Left

σ with the sample standard deviation *s*. The *t* distribution has thicker tails than the standard normal N(0, 1) distribution, and this accounts for the additional variability.



As a result, the critical value using the t distribution ($t^* = 2.045$) is larger than the critical value using the N(0, 1) distribution ($z^* = 1.960$). Since $t^* > z^*$ at a particular confidence level, confidence intervals using the t distribution are wider than confidence intervals using the N(0, 1) distribution.

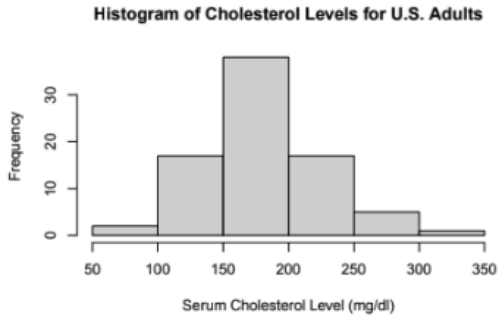
Deleted: 4

Formatted: Left

Example 1

The Centers for Disease Control and Prevention (CDC) reported that the **mean** serum total cholesterol level for adults aged 20 and older in the United States was **191 mg/dL** in 2015-2018.6 Recently, a researcher recruited a **random sample of 80** U.S. adults aged 20 and older to **estimate** the mean serum cholesterol level.

R code




```
summary(chol_levels)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   72.13  153.85  174.54  180.03  202.30  311.92
```

```
sd(chol_levels)
```

```
## [1] 45.19351
```

Step 1: Check conditions

The researcher plans to calculate a 95% confidence interval. Let μ represent the mean serum cholesterol level of U.S. adults aged 20 and over.

1. Independence: The independence condition is satisfied because the researcher took a random sample. Check the remaining condition for computing a confidence interval.
2. Normality: The histogram is unimodal and looks to be fairly symmetric, suggesting that the serum cholesterol levels came from a population with a normal distribution.

Deleted: 5

Formatted: Left

Step 2: Calculate the Confidence Interval
Use R to calculate the critical value:

```
qt(0.025, df = 79, lower.tail = FALSE)
```

```
## [1] 1.99045
```

$$df = n - 1 = 80 - 1 = 79$$

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}} = 180.03 \pm 1.990 \frac{45.19351}{\sqrt{80}} = 180.03 \pm 10.055 = (169.975, 190.085)$$

Step 3: Interpretation

What is wrong with these interpretations?

Interpretation 1: We estimate with 95% confidence that the sample mean serum cholesterol level of U.S. adults aged 20 and over is between 169.975 and 190.085 mg/dl.
It's the population mean, not the sample mean.

Interpretation 2: 95% of the serum cholesterol levels of U.S. adults aged 20 and over is between 169.975 and 190.085 mg/dl.
The interval gives us an estimate for the population mean, not for the cholesterol levels of individuals.

Interpretation 3: There is a 95% chance that the population mean serum cholesterol level of U.S. adults aged 20 and over is between 169.975 and 190.085 mg/dl.
There is no chance about it. The population mean is either in the CI or it isn't.

Correct Interpretation:
We estimate with 95% confidence that the population mean serum cholesterol level of US adults aged 20 and over is between 169.975 and 190.085 mg/dl.

How does Confidence Interval and Hypothesis Test relate?
There is a relationship between confidence intervals and hypothesis tests. Consider a 95% confidence interval and a test of $H_0 : \mu = \mu_0$ and
$H_A: \mu \neq \mu_0$ at the $\alpha = 0.05$ significance level.

- Any μ_0 value within the 95% confidence interval is a value that would result in a p-value > 0.05, meaning we would fail to reject the null hypothesis
- Any μ_0 value outside of the 95% confidence interval is a value that would result in a p-value < 0.05, meaning we would reject the null hypothesis

Deleted: 6
Formatted: Left

Example 2

Would you reject or fail to reject H_0: $\mu = 170$ and H_A: $\mu \neq 170$ at the $\alpha = 0.05$ significance level?

Since 170 is in the confidence interval, it is a reasonable value for the population mean serum cholesterol level. We would fail to reject the null hypothesis.

Would you reject or fail to reject H_0: $\mu = 195$ and H_A: $\mu \neq 195$ at the $\alpha = 0.05$ significance level?

Since 195 is not in the confidence interval, we don't think it is a reasonable value for the population mean serum cholesterol level. We would not fail to reject the null hypothesis.

Deleted: 7

Formatted: Left