Lab 3B: Foundations for inference - Confidence levels

Complete all **Exercises**, and submit answers to **Questions** on the Coursera platform. Note that the order of the choices in multiple choice questions may be different on the Duke Coursera platform than the order in this document.

If you have access to data on an entire population, say the size of every house in Ames, Iowa, it's straight forward to answer questions like, "How big is the typical house in Ames?" and "How much variation is there in sizes of houses?". If you have access to only a sample of the population, as is often the case, the task becomes more complicated. What is your best guess for the typical size if you only know the sizes of several dozen houses? This sort of situation requires that you use your sample to make inference on what your population looks like.

The data

In Part B of the lab we'll start with a simple random sample of size 60 from the population. Note that the data set has information on many housing variables, but for the first portion of the lab we'll focus on the size of the house, represented by the variable `Gr.Liv.Area`.

```
load(url("http://www.openintro.org/stat/data/ames.RData"))
population <- ames$Gr.Liv.Area
samp <- sample(population, 60)
```

Describe the distribution of your sample. What would you say is the "typical" size within your sample? Also state precisely what you interpreted "typical" to mean.

Question 1 My distribution should be similar to others' distributions who also collect random samples from this population, but it is likely not exactly the same since it's a random sample.

A) True
B) False

Confidence intervals

One of the most common ways to describe the typical or central value of a distribution is to use the mean. In this case we can calculate the mean of the sample using

```
sample_mean <- mean(samp)
```

Return for a moment to the question that first motivated this lab: based on this sample, what can we infer about the population? Based only on this single sample, the best estimate of the average living area of houses sold in Ames would be the sample mean, usually denoted as $\bar{x}$ (here we're calling it `sample_mean`). That serves as a good **point estimate** but it would be useful to also communicate how uncertain we are of that estimate. This can be captured by using a **confidence interval**.

We can calculate a 95% confidence interval for a sample mean by adding and subtracting 1.96 standard errors to the point estimate. (See Section 4.2.3 if you are unfamiliar with this formula.)

```
se <- sd(samp)/sqrt(60)
lower <- sample_mean - 1.96 * se
upper <- sample_mean + 1.96 * se
c(lower, upper)
```

This is an important inference that we've just made: even though we don't know what the full population looks like, we're 95% confident that the true average size of houses in Ames lies between the values `lower` and `upper`. There are a few conditions that must be met for this interval to be valid.

Question 2 For the confidence interval to be valid, the sample mean must be normally distributed and have standard error $s/\sqrt{n}$. Which of the following is not a condition needed for this to be true?

A) The sample is random.
B) The sample size, 60, is less than 10% of all houses.
C) The sample distribution must be nearly normal.

Confidence levels

Question 3 What does "95% confidence" mean?

A) 95% of the time the true average area of houses in Ames, Iowa, will be in this interval.
B) 95% of random samples of size 60 will yield confidence intervals that contain the true average area of houses in Ames, Iowa.
95% of the houses in Ames have an area in this interval.
C) 95% confident that the sample mean is in this interval.

In this case we have the luxury of knowing the true population mean since we have data on the entire population. This value can be calculated using the following command:

```
mean(population)
```

- **Exercise:** Does your confidence interval capture the true average size of houses in Ames?

Question 4 What proportion of 95% confidence intervals would you expect to capture

the true population mean?

A) 1%
B) 5%
c) 95%
D) 99%

Using R, we're going to recreate many samples to learn more about how sample means and confidence intervals vary from one sample to another. **Loops** come in handy here.

Here is the rough outline:

- Step 1: Obtain a random sample.
- Step 2: Calculate the sample's mean and standard deviation.
- Step 3: Use these statistics to calculate a confidence interval.
- Step 4: Repeat steps (1)-(3) 50 times.

But before we do all of this, we need to first create empty vectors where we can save the means and standard deviations that will be calculated from each sample. And while we're at it, let's also store the desired sample size as `n`.

```
samp_mean <- rep(NA, 50)
samp_sd <- rep(NA, 50)
n <- 60
```

Now we're ready for the loop where we calculate the means and standard deviations of 50 random samples.

```
for(i in 1:50){
  samp <- sample(population, n) # obtain a sample of size n = 60 from the populati
on
  samp_mean[i] <- mean(samp)    # save sample mean in ith element of samp_mean
  samp_sd[i] <- sd(samp)        # save sample sd in ith element of samp_sd
}
```

Lastly, we construct the confidence intervals.

```
lower <- samp_mean - 1.96 * samp_sd / sqrt(n)
upper <- samp_mean + 1.96 * samp_sd / sqrt(n)
```

Lower bounds of these 50 confidence intervals are stored in `lower`, and the upper bounds are in `upper`. Let's view the first interval.

```
c(lower[1],upper[1])
```

```
plot_ci(lower, upper, mean(population))
```

- **Exercise:** Does this proportion of confidence intervals that include the true population mean,

exactly equal to the confidence level? If not, explain why.

Question 5 What is the appropriate critical value for a 99% confidence level?

A) 0.01
B) 0.99
C) 1.96
D) 2.33
E) 2.58

Calculate 50 confidence intervals at the 99% confidence level. You do not need to obtain new samples, simply calculate new intervals based on the sample means and standard deviations you have already collected. Using the `plot_ci` function, plot all intervals and calculate the proportion of intervals that include the true population mean.

Question 6 We would expect 99% of the intervals to contain the true population mean.

A) True
B) False