MSDS 5043 - Assignment 3

Olivia Samples, Lipscomb University

10/8/2019

Sampling from Ames, Iowa

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 the previous lab, "Sampling Distributions", we looked at the population data of houses from Ames, Iowa. The data set is available on CANVAS. In this 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.

```
ames = read.csv("/Users/osamples/Desktop/MSDS5043/ames.csv", head = TRUE, sep =",")
dim(ames)
```

```
## [1] 2930   82
```

```
population <- ames$Gr.Liv.Area
set.seed(123)
samp <- sample(population, 60)
```

Question 1

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.

If we're interested in estimating the average size of the house, our best single guess is the sample mean.

```
mean(population)
```

```
## [1] 1499.69
```

```
summary(samp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     808    1061    1413    1534    1816    3395
```

```
IQR(samp)
```

```
## [1] 754.75
```

Here we have that our sample distribution has an IQR range of 635.25 and a sample mean of 1624. The median (1413) is less than the mean (1534) which means that the sample distribution is skewed to the left. We have that our sample mean is the "typical" size of a house within our sample, because it is the average size. In comparison to the population mean, the sample mean (1624) is greater than the population mean (1499.69). Depending on which 60 homes we select, the estimate could be a bit above or a bit below the true

population mean. In general, though, the sample mean turns out to be a pretty good estimate of the average living area.

Question 2

Pull a second set of samples. Would you expect its distribution to be identical to the first one? Would you expect it to be similar? Why or why not?

We will expect the sample mean to be similar, but not identical, because it is nearly impossible for the random sample to measure the exact same 60 house sizes. This difference allows us to see the variability that we should expect.

```
set.seed(456)
samp2 <- sample(population, 60)
summary(samp2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     790    1149    1470    1510    1769    2775
```

And so, we have that the samp mean 1534 does not equal the samp2 mean 1510, but they are very similar to eachother.

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 bet 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)
```

```
## [1] 1390.046 1677.388
```

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 3

For the confidence interval to be valid, the sample mean must be normally distributed and have a standard error $s/\sqrt{(n)}$. What conditions must be met for this to be true?

There are four conditions that must be met. 1. The sample must be collected randomly. 2. The trials must be independent of eachother. 3. The sample must have at least 10 successes and 10 failures. 4. The population must be large enough, or at least 10 times the sample size.

Confidence Intervals

Question 4

What does "95% confidence" mean?

When there is a 95% confidence interval, it means that our sample distribution can accurately predict the average size of a house with 95% confidence. In other words, when taking many samples, the true population mean is incorrectly estimated from a sample distribution only 5% of the time.

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)
```

```
## [1] 1499.69
```

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 a rough outline: 1. Obtain a random sample. 2. Calculate and store the sample's mean and standard deviation. 3. Repeat steps (1) and (2) 50 times. 4. Use these stored statistics to calculate many confidence intervals. 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 frome ach sample. And while we're at it, let's also store the desired sample size as n.

```
sample_mean <- rep(NA, 50)
sample_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){
        sample <- sample(population, n)
        sample_mean[i] <- mean(sample)
        sample_sd[i] <- sd(sample)
}
```

Lastly, we construct the confidence intervals.

```
lower_vector <- sample_mean - 1.96 * sample_sd/ sqrt(n)
upper_vector <- sample_mean + 1.96 * sample_sd/ sqrt(n)
```

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

```
c(lower_vector[1], upper_vector[1])
```

```
## [1] 1309.868 1574.099
```

Question 5

Pick a confidence level of your choosing, provide it is not 95%, calculate 50 confidence intervals as above.

I will choose a 90% confidence interval which has a critical value of 1.64. We follow the same steps as above.

```
for(i in 1:50){
        sample <- sample(population, n)
        sample_mean[i] <- mean(sample)
        sample_sd[i] <- sd(sample)
}
```

Here, we construct the confidence intervals with critical value of 1.64.

```
lower_vector <- sample_mean - 1.64 * sample_sd/ sqrt(n)
upper_vector <- sample_mean + 1.64 * sample_sd/ sqrt(n)
```

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

```
c(lower_vector[1], upper_vector[1])
```

```
## [1] 1430.256 1641.411
```

OpenIntro Questions

Question 6: Problem 4.3

A college counselor is interested in estimating how many credits a student typically enrolls in each semester. The counselor decides to randomly sample 100 students by using the registrar's database of students. The histogram below shows the distribution of the number of credits taken by these students. Sample statistics for this distribution are also provided.

(a) What is the point estimate for the average number of credits taken per semester by students at this college? What about the median?

The point estimate is the mean sample statistic: 13.65 credits. The median is 14 classes.

(b) What is the point estimate for the standard deviation of the number of credits taken per semester by students at this college? What about the IQR?

The sample standard deviation is 1.91. The inter-quartile range is Q3 - Q1, or 15 - 13 = 2.

(c) Is a load of 16 credits unusually high for this college? What about 18 credits? Explain your reasoning.

While 16 is outside of the IQR, we can determine if it is unusual more accurately. We can calculate whether or not 16 credits is outside of a confidence interval of 95% with a critical value of 1.96.

```
u <- 13.65
uhat <- 16
sd <- 1.91
zscore <- (uhat - u)/ sd
zscore
```

```
## [1] 1.230366
```

Here we have that 16 credits has a critical value of 1.230366 which is less than 1.96. And so, 16 credits is within the confidence interval and is not an unusual value.

```
u <- 13.65
uhat <- 18
sd <- 1.91
zscore <- (uhat - u)/ sd
zscore
```

```
## [1] 2.277487
```

On the other hand, 18 credits has a critical value of 2.277487 which is greater than 1.96. This means that 18 credits is outside of the 95% confidence interval which we classify as an unusual value.

(d) The college counselor takes another random sample of 100 students and this time finds a sample mean of 14.02 units. Should she be surprised that this sample statistic is slightly different than the one from the original sample? Explain your reasoning.

Because of the randomness of the sample, we should expect a different mean size. It is very common for the sample mean to be different between different random samples, so she should not be surprised that this sample statistic is slightly different from the original.

(e) The sample means given above are point estimates for the mean number of credits takeb by all students at that college. What measures do we use to quantify the variability of this estimate? Compute this quantity using the data from the original cycle.

We use the standard error to quantify the variability of the sample mean to the true population mean. We can calculate that using the sample standard deviation and the sample size n.

```
sd = 1.91
n = 100

se = sd/sqrt(n)
se
```

```
## [1] 0.191
```

And so, the standard error is 0.191.

Question 7: Problem 4.6

Elijah and Tyler, two high school juniors, conducted a survey on 15 students at their school, asking the students whether they would like the school to offer an afterschool art program, counted the number of "yes" answers, and recorded the sample proportion. 14 out of 15 students responded "yes". They repeated this 100 times and built a distribution of sample means.

(a) What is this distribution called?

Because this distribution represents the distribution of the point estimates based on samples of a fixed size (15) from a certain population (high school students), we have a sampling distribution.

(b) Would you expect the shape of this distribution to be symmetric, right skewed, or left skewed? Explain your reasoning.

While the one of the sample proportions is very high at 14/15 or, 93%, the distribution will show 99 other sample proportions that have similar proportions. Because we have a large n of 100 and the Central Limit Theorem, I would expect the shape of this distribution to be symmetric.

(c) Calculate the variability of this distribution and state the appropriate term used to refer to this value.

The standard deviation of the sample proportion can be calculated as follows:

```
p = 14/15
n = 100
ssd = sqrt((p*(1-p))/n)
ssd
```

```
## [1] 0.02494438
```

This can be referred to as the sampling variability.

(d) Suppose that the students were able to recruit a few more friends to hel them with sampling and are now able to collect data from random samples of 25 students. Once again, they record the number of "yes" answers, and record the sample proportion. How will the variability of this new distribution compare to the variability of the original?

When the sample size is larger, the variance gets smaller. In other words, the error gets smaller and it is closer to the point estimate. So in this case, the variability of this new distribution will be less than the original.

Question 8: Problem 4.14

The 2009 holiday retail season, which kicked off on November 27, 2009 (the day after Thanksgiving), had been marked by somewhat lower self-reported consumer spending than was seen during the comparable period in 2008. To get an estimate of consumer spending, 436 randomly sampled American adults were surveyed. Daily consumer spending for the six-day period after Thanksgiving, spanning the Black Friday weekend and Cyber Monday, averaged $84.71. A 95% confidence interval based on this sample is ($80.31, $89.11). Determine whether the following statements are true or false, and explain your reasoning.

(a) We are 95% confident that the average spending of these 436 American aduts is between $80.31 and $89.11.

False, the average spending of these 436 American adults can be referred to as the sample mean which is always in the confidence interval. Instead, we are 95% confident that the population mean is within the confidence interval.

(b) This confidence interval is not valid since the distribution of spending in the sample is right skewed.

False, a skewed distribution does not affect the validity of the confidence interval.

(c) 95% of random samples have a sample mean between $80.31 and $89.11.

False, 95% of this specific random sample has a sample mean between $80.31 and $89.11. However, random samples of different sizes will have different confidence intervals due to variability.

(d) We are 95% confident that the average spending of all American adults is between $80.31 and $89.11.

True, this refers to the population mean we referenced in (a).

(e) A 90% confidence interval would be narrower than the 95% confidence interval since we don't need to be as sure about our estimate.

True, because we are 5% less sure about our estimate.

(f) In order to decrease the margin of error of a 95% confidence interval to a third of what it is now, we would need to use a sample 3 times larger.

False, we have that the margin of error is z * SE, where $SE = \sigma/sqrt(n)$. And so, if we would like the make margin error 1/3 of what it is now, we would need the sample size to be equal to 9, and not 3.

(g) The margin of error is 4.4.

True, the margin of error measures the distance from the mean to one confidence interval limit. Here, the margin of error is 4.4.

```
89.11-84.71
```

```
## [1] 4.4
```

Question 9: Problem 4.16

The National Survey of Family Growth conducted by the Centers for Disease Control gathers information on family life, marriage and divorce, pregnancy, infertility, use of contraception, and men's and women's health. One of the variables collected on this survey is the age at first marriage. The histogram below shows the distribution of ages at first marriage of 5,534 randomly sampled women between 2006 and 2010. The average age at first marriage among these women is 23.44 with a standard deviation of 4.72. Estimate the average age at first marriage of women using a 95% confidence interval, and interpret this interval in context. Discuss any relevant assumptions.

```
n = 5534
u = 23.44
sd = 4.72
z = 1.96
se = sd/sqrt(n)
lower = u - z*se
upper = u + z*se
c(lower, upper)
```

```
## [1] 23.31564 23.56436
```

We are 95% confident that the average age at first marriage of women (population) is between 23.32 and 23.56. In order for this to be true, we have to assume the Central Limit Theorem. Also, while we know that the women were randomly sampled, we must assume they were also sampled independently. We know that there were at least 10 successes and 10 failures, but we have to assume that the population is at least 10 times the sample size of 5534.