Lab 3 Key

Anwesha Guha & Janette Avelar

1/25/2022

For Lab 3, we're going to be exploring the concepts of normal distribution and sampling distribution by creating our own custom population and samples using R. We have provided the necessary code for you to create the population and pull samples, but you are expected to draw from previous labs and class materials to answer the following questions.

The first thing we'll do is read in the ECLS-K data. Refer to Lab 2 on directions on how to import data files into R.

For the purpose of this assignment, we will be looking at the continuous SES measure for Kindergarten students in this dataset, which is the column X12SESL.

```
require(rio)
```

```
## Loading required package: rio
```

```
require(here)
```

```
## Loading required package: here
```

```
## here() starts at /Users/janetteavelar/EMPL/ge_winter22_educ614/EDUC614_Labs/lab3
```

```
ecls <- import(here("data", "ecls-k-sub.csv"))
```

Even though X12SESL already has quite a few NAs, there are still some values in the dataset for -9 that you need to recode as `NA` before getting started.

```
ecls$X12SESL[ecls$X12SESL == -9] <- NA
```

Question 1: Find the descriptive statistics for your population data, `ecls-k-sub$X12SESL`. Remember to require the `psych` package before using the `describe` function.

```
require(psych)
```

```
## Loading required package: psych
```

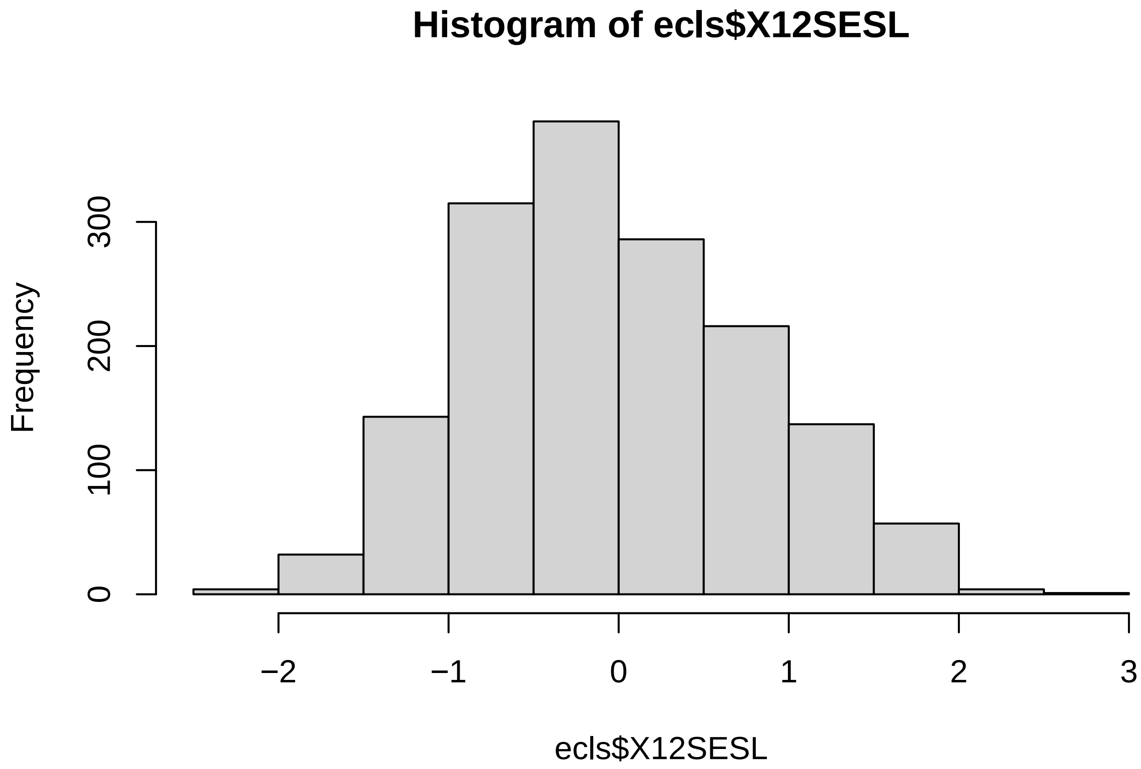
```
describe(ecls$X12SESL)
```

```
##    vars    n  mean   sd median trimmed  mad   min max range skew kurtosis   se
## X1    1 1576 -0.04 0.82  -0.14   -0.06 0.88 -2.33 2.6  4.93 0.28    -0.41 0.02
```

Mean: -0.04 Median: -0.14 Standard deviation: 0.82 Skew: 0.28 Kurtosis: -0.41

Question 2: Create a histogram for your population data.

```
hist(ecls$X12SESL)
```



Question 3: Now we're going to generate a random sample of 10 from our population.

```
set.seed(100)

sample_10 <- sample(ecls$X12SESL, 10, replace = TRUE)

describe(sample_10)
```

```
##    vars  n mean   sd median trimmed  mad   min  max range skew kurtosis   se
## X1    1 10 0.12 0.74   0.27    0.14 0.71 -1.05 1.08  2.13 -0.3    -1.63 0.24
```

Report the descriptive statistics for your sample.

Mean: 0.12 Median: 0.27 Standard deviation: 0.74 Skew: -0.3 Kurtosis: -1.63

Question 4: Run the code again, replacing 10 with 30, to generate a random sample of 30 from our population.

```
set.seed(100)
```

```
sample_30 <- sample(ecls$X12SESL, 30, replace = TRUE)

describe(sample_30)
```

```
##    vars  n mean   sd median trimmed  mad   min  max range skew kurtosis   se
## X1    1 28 0.24 0.74   0.28    0.22 0.71 -1.05 1.58  2.63 0.09    -1.12 0.14
```

Report the descriptive statistics for that sample below. What changes?

Mean: 0.24 Median: 0.28 Standard deviation: 0.74 Skew: 0.09 Kurtosis: -1.12

Ideally, increasing the sample size would make the sample metrics look closer to those for the parent population. However, remember these samples are still small and random (which also means your descriptive statistics above may differ slightly). You can try using different numbers for the `set.seed()` and you will get a variety of different results. So, we increase the number of times we sample the data to get a better estimate of the population below.

Question 4: We're now going to create a sampling distribution by generating 30 random samples of 10.

```
set.seed(100)

samples_10_30 <- replicate(30, sample(ecls$X12SESL, 10, replace = TRUE))
```

Report the mean, median, and standard deviation for your new sampling distribution. Note that your new sampling distribution `samples_10_30` is a dataset, and the `describe()` function will run descriptive statistics for each individual sample of 10 within it, rather than a summary of the totals. That's why we need to run the functions `mean()`, `median()`, and `sd()` to extract the values. *Be sure to include `na.rm = TRUE` in your argument!*

```
mean(samples_10_30, na.rm = TRUE)
```

```
## [1] -0.06564921
```

```
median(samples_10_30, na.rm = TRUE)
```

```
## [1] -0.1185
```

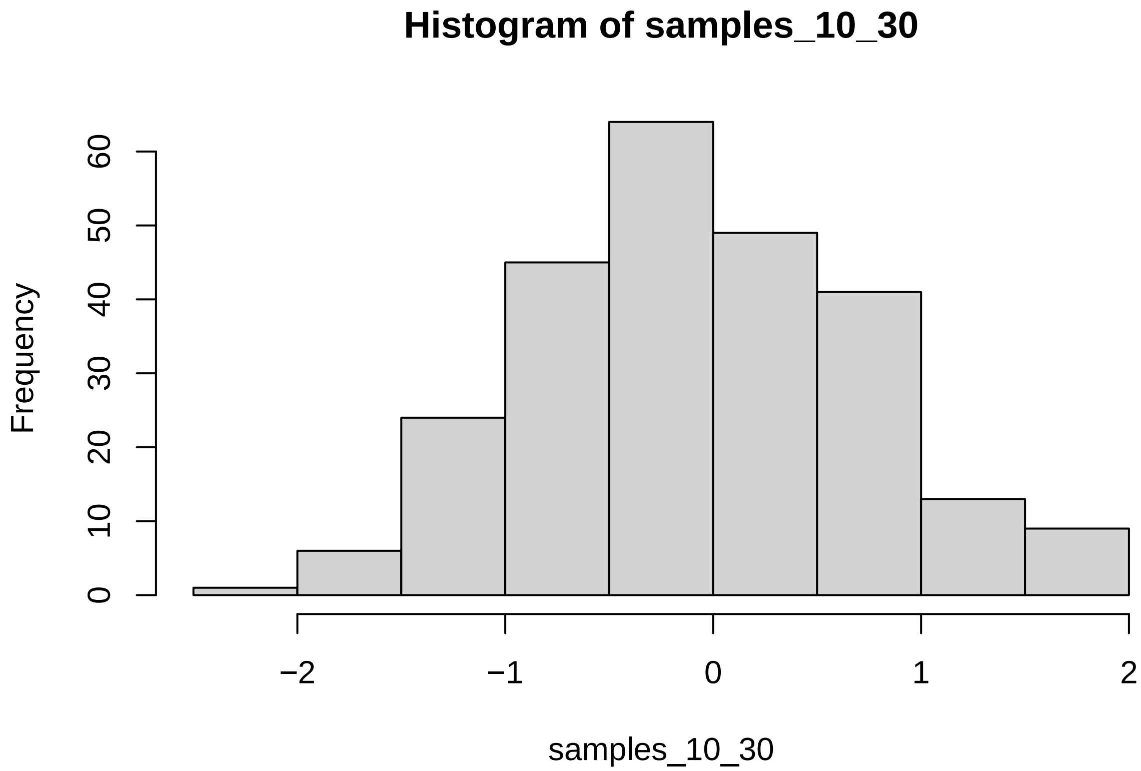
```
sd(samples_10_30, na.rm = TRUE)
```

```
## [1] 0.797901
```

Mean: -0.0656 Median: -0.1185 SD: 0.7979

Question 5: Create a histogram for your sampling distribution.

```
hist(samples_10_30)
```



Question 6: Suppose that you randomly sample from your population with a size of 10 and compute the mean for each sample. You repeat this an infinite number of times. What would you expect the mean and standard deviation of your sampling distribution to be? Compute the mean and standard deviation of the hypothetical sampling distribution and explain your reasoning.

The mean of all means would remain the same, at -0.04. The standard deviation would follow the following formula $sd = \frac{\sigma}{\sqrt(n)}$, so it would equal $\frac{0.82}{\sqrt(10)} = 0.259$.

Question 7: Let's check our work by drawing 10,000 random samples of 10 from our population.

```
set.seed(100)
samples_more <- replicate(10000, sample(ecls$X12SESL, 10, replace = TRUE))
```

Calculate the descriptive statistics for your 10,000 samples. Compare them to the descriptive statistics of your sampling distribution in questions 1 and 2. Do they match or not? Explain why you think that is.

```
mean(samples_more, na.rm = TRUE)
```

```
## [1] -0.03570287
```

```
sd(samples_more, na.rm = TRUE)
```

```
## [1] 0.8175495
```

```r
median(samples_more, na.rm = TRUE)
```

```
## [1] -0.13
```

Mean: -0.357 Median: -0.13 Standard deviation: 0.8175

Each of these statistics are approaching the population values (Mean: -0.04; Median: -0.14; Standard deviation: 0.82). In an infinite sample, the descriptive statistics would totally match.

Theoretical Questions:

Question 1: A recently admitted class of graduate students at a large state university has a mean of GRE verbal score of 650 with a standard deviation of 50. The scores are reasonably normally distributed. One student, whose mother just happens to be on the board of trustees, was admitted with a GRE score of 490. Should the local newspaper editor, who loves scandals, write a scathing editorial about favoritism?

A GRE score of 490 falls more than 3 standard deviations away from the mean. This student has a 0.0001% of getting in using GRE alone (z = 490 – 650/50 = -160/50 = -3.2 -> 0.0001%). While the local newspaper editor could write an article about favoritism if they were assuming the GRE was a significant determinant of admittance, they should also be careful (correlation ≠ causation) and consider that a variety of factors go into admittance, especially if the student had any other outstanding qualities (other than being related to a board of trustees member).

Question 2: The amount of money college students spend each semester on textbooks is normally distributed with a mean of $195 and a standard deviation of $20. Suppose you take a random sample of 100 college students from this population. There would be a 68% chance that the sample mean for the amount spent on textbooks would be between **$175** and **$215**.

According to the CLT, the mean of the sample would be the same as the population at $195, and the standard deviation will be $2, according to the following calculations:

$$sd = \frac{\sigma}{\sqrt{(n)}} = \frac{20}{\sqrt{(100)}} = 2$$

Since 68% encompasses approximately one standard deviation, the amount spent on textbooks will be between $193 and $197, according to the following calculations:

$$mean \pm 1 * sd = 195 \pm 2$$