314 Data Analysis 4

Sampling Distributions

In this data analysis, you will investigate the ways in which the statistics from a random sample of data can serve as point estimates for population parameters. We're interested in formulating a sampling distribution of our estimate in order to learn about the properties of the estimate, such as the estimate's distribution.

If you haven't already done so, work through the tutorial provided on the Data Analysis 4 Canvas page. Once you've worked through the tutorial, write up your responses to the questions listed throughout the tutorial. The same questions are included below to help you format your submissions.

Submit a PDF copy of your responses to Gradescope by the deadline stated on Canvas.

Question 1 (2 points)

Describe the distribution of squeaking vs. no squeaking in this sample. How does it compare to the distribution of the population?

```
# A tibble: 2 × 3
  hip             n sample_proportion
  <chr>       <int>             <dbl>
1 No squeaking   183            0.915
2 Squeaking       17            0.085

# A tibble: 2 × 3
  hip              n        p
  <chr>        <int>    <dbl>
1 No squeaking 232500    0.93
2 Squeaking     17500    0.07
```

Based on the sample data of 200 people who did hip replacement, 183 individuals did not experience squeaking (91.5%) while only 18 reported squeaking (8.5%). This suggests that squeaking is a relatively uncommon issue with the product. When examining the population distribution, out of 232,500 people who received hip replacement, 93% did not report squeaking, which is a higher percentage compared to the sample data. Additionally, only 17,500 individuals reported squeaking, which is a lower percentage compared to the sample data. In conclusion, the distribution of sample data is very close to the population.

Question 2 (2 points)

Would you expect the sample proportion to match the sample proportion of another student's sample? Why, or why not? If the answer is no, would you expect the proportions to be somewhat different or very different?

No, I wouldn't expect the sample proportion to match the sample proportion of another student's sample, because different samples may have different random variations and sampling errors that affect their proportions. However, I would expect the proportions to be somewhat similar.

Question 3 (2 points)

Take a second sample, also of size 200, and call it samp2. How does the sample proportion of samp2 compare with that of samp1?

```
  # A tibble: 2 × 3
 hip               n      sample_proportion
 <chr>           <int>            <dbl>
1 No squeaking   189            0.945
2 Squeaking       11              0.055
```

According to the sample_2, 189 individuals did not experience squeaking (94.5%) which is larger than 91.5% while only 11 reported squeaking (5.5%) is smaller than 8.5%.

Question 4 (1 point)

Suppose we took two more samples, one of size 250 and one of size 1000. Which do you think would provide a more accurate estimate of the population proportion?

Sample size of 1000 is more accurate estimate of the population proportion. Because as the sample size increases, the sampling error decreases, and he sample estimate becomes more precise.

Question 5 (2 points)

To make sure you understand how sampling distributions are built, and exactly what the rep_sample_n function does, try modifying the code to create a sampling distribution of 10 sample proportions from samples of size 100, and put them in a data frame named sample_props_small. Include the output in your submission document. You can print the entire sample_props_small dataset in your R console by simply running the line sample_props_small after you've created it. **How many observations are there in this object called sample_props_small? What does each observation represent?** *Hint: your sample_props_small data set should contain three columns (hip, n, p_hat).*

```
sample_props_small
# A tibble: 10 × 4
# Groups:   replicate [10]
   Replicate  hip          n p_hat
       <int> <chr>     <int>  <dbl>
         1 Squeaking     6  0.06
        2 Squeaking     7  0.07
       3 Squeaking     6  0.06
       4 Squeaking     8  0.08
       5 Squeaking    10  0.1
      6 Squeaking     4  0.04
      7 Squeaking     6  0.06
     8 Squeaking     7  0.07
     9 Squeaking    10  0.1
   10 Squeaking     4  0.04
```

There are 10 observations in this line sample_props_samll, each observation in the sample_props_samll represents a sample of size 100 from the population of patients who has hip replacement surgery. Each observation contains 3 values first one is "hip": which is a categorical variable indicating whether patient experienced squeaking in their replacement surgery. Second "n": the sample size which is fixed at 100 for all observations in this data frame. Third "p_hat": the sample proportion of patient who experienced squeaking in their hip replacement surgery, calculated from each individual sample.

Question 6 (4 points)

A. (2 points) According to the Central Limit Theorem, what distribution does the sample proportion follow if the sample size is 200 and the true population proportion is equal to 0.07? Be sure to include the name of the distribution and the numerical values of its parameters.

According to the CLT, sample size 200 is sufficient, and the population proportion is not too close to 0 or 1(0.07). Therefore, The Normal distribution follow of the sample size is 200 and the true population proportion is equal to 0.07.
mean= 0.07   SD: sqrt [0.07) (1-0.07)/200] = 0.01804
Thus, the normal distribution can be written as N (0.07,0.0180)

B. (2 points) Using the app above, simulate 10,000 samples of size 200 that come from a population with parameter p=0.07. Locate the values of the mean and standard deviation (std dev) from the simulated sampling distribution. These values are located just under the title of the third plot. How do these values compare to the theoretical mean and standard deviation of the sampling distribution as a result of the Central Limit Theorem? In other words, compare the values in your simulated sampling distribution using the app to the parameter values you reported in part A of this question.

Simulated Sampling distribution mean is 0.0696 and standard deviation is 0.0181. The mean I got in part A is 0.07 which is slightly bigger than Simulated Sampling distribution mean 0.0696, the standard deviation from part A I got is 0.0180 which is slightly smaller than 0.0181. However, the simulated values are very close to the theoretical mean and standard deviation of the sampling distribution as a result of the Central Limit Theorem.

Question 7 (2 points)

Increase the sample size to 500 (leave the number of simulated samples at 10,000). How does the sampling distribution of the sample proportion change when the sample size is increased to 500. Be sure to comment on comparisons between the shape, center, and spreads of the distributions.

Shape: After increased sample size to 500, shape didn't change that much based on increased sample size, both are symmetric and unimodal.
Center: After increased sample size to 500, the value of center is increased to 0.07 from 0.0696
Spreads: After increased sample size to 500, the standard deviation is decreased from 0.018 to 0.0113, and spreads is less spread after increased the size to 500, in other words range become smaller.

Question 8 (2 points)

Decrease the sample size to 50 (leave the number of simulated samples at 10,000). Comment on the shape of the sampling distribution of the sample proportion for this smaller sample size. Are the conditions of the Central Limit Theorem for the sampling distribution of the sample proportion met in this scenario? Why or why not?

Once I decreased size from 500 to 50, the shape of the sampling distribution of the sample proportion become more variable and less normal in shape which looks slightly left skewed.

To see if the conditions of CLT for the sampling distribution of the sample proportion meet this scenario, there are two things to check:

1) n > = 30, here 50 is greater than 30
2) n*p > = 10 & n(1-p) >= 10: 50*0.07 = 3.5 which 3.5 is not equal and greater than 10& 50*0.03 = 1.5 which is not greater and equal than 10.

Thus, the conditions of CLT for the sampling distribution of the sample proportion meet this scenario, because it didn't satisfy the "n*p > = 10 and n(1-p) > = 10"

Question 9 (6 points)

Return to the scenario where the sample size is 200 and the population proportion is p=0.07. Use R to solve the following problems. You are encouraged to include your code used to do the following calculations so that if you answer incorrectly, we can help you understand where you went wrong.

A. (2 points) Using the theoretical sampling distribution of the sample proportion $\hat{p}$ (your answer to 6A), calculate the probability that from a single sample of 200 ceramic hip patients, less than 4% of sampled patients develop squeaking.

   pnorm(0.04, mean = 0.07, SD = 0.01804) = 0.0481591

   Thus, the probability that from a single sample of 200 ceramic hip patients, less than 4% of sampled patients develop squeaking is 0.048.

B. (2 points) Using the theoretical sampling distribution of the sample proportion $\hat{p}$ (your answer to 6A), calculate the probability that from a single sample of 200 ceramic hip patients, more than 10% of sampled patients develop squeaking.

   1-pnorm (0.1, mean = 0.07, SD= 0.1804) = 0.0481591

   Therefore, the probability that form a single sample of 200 ceramic hip patients, more than 10% of sampled patients develop squeaking is 0.048.

C. (2 points) Using the theoretical sampling distribution of the sample proportion $\hat{p}$ (your answer to 6A), calculate the probability that from a single sample of 200 ceramic hip patients, between %5 and 10% of sampled patients develop squeaking.

   pnorm(0.1,0.07,0.0184) - pnorm(0.05,0.07,0.0184) =0.8099672

   Therefore, the probability that form a single sample of 200 ceramic hip patients, between %5 and 10% of sampled patients develop squeaking is 0.81.

Gradescope Page Matching

When you upload your PDF file to Gradescope, you will need to match each question on this assignment to the correct pages. Video instructions for doing this are available in the Start Here module on Canvas on the page “Submitting Assignments in Gradescope”. Failure to follow these instructions will result in a 2-point deduction on your assignment grade. Match this page to outline item “Gradescope Page Matching”.