Assignment 5

YOUR NAME GOES HERE

Due on 2/14 at 11:59pm

```
library(ggplot2)
library(ggpubr)
```

1

Write what the Central Limit Theorem states. Use LaTeX math mode to write your answer. And, be sure to cite sources from which you obtain your information. The Central Limit Theorem states that the mean of your sample means () should approximate to the actual population mean. In general, what the CLT is saying is that if were were to plot the sample mean repeatedly, those sample means would create a normal distribution centered on the actual population mean. This is beautiful in the sense that even if our population probabilities is not normally distributed, we could make it normally distributed by looking at the sample mean distribution. Furthermore, the CLT also states that the standard deviation of the sample means is gets smaller as the size of each sample increases in size (n). This is also powerful, in the sense that it suggests that increasing sample size closer to population size will lead to less variance and closer to population mean. Here is the CLT formalized:

1. $\mu_{\bar{X}} = \mu$
2. $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

What the first is saying is that the mean of the sample means is equal to the mean of the population What the second is saying is that the standard deviation of our samples is equal to the actual population standard deviation divided by the square root of the size of the sample

sites used: 1. https://www.statisticshowto.com/probability-and-statistics/normal-distributions/central-limit-theorem-definition-examples/ 2. https://courses.lumenlearning.com/odessa-introstats1-1/chapter/the-central-limit-theorem-for-sums/ 3. https://www.khanacademy.org/math/ap-statistics/sampling-distribution-ap/what-is-sampling-distribution/v/central-limit-theorem 4. https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_probability/BS704_Probability12.html

2

Below, you will use R to create some data and plots that will help you demonstrate the core concepts of the CLT below.

2.1

Start by setting the seed at 100.

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

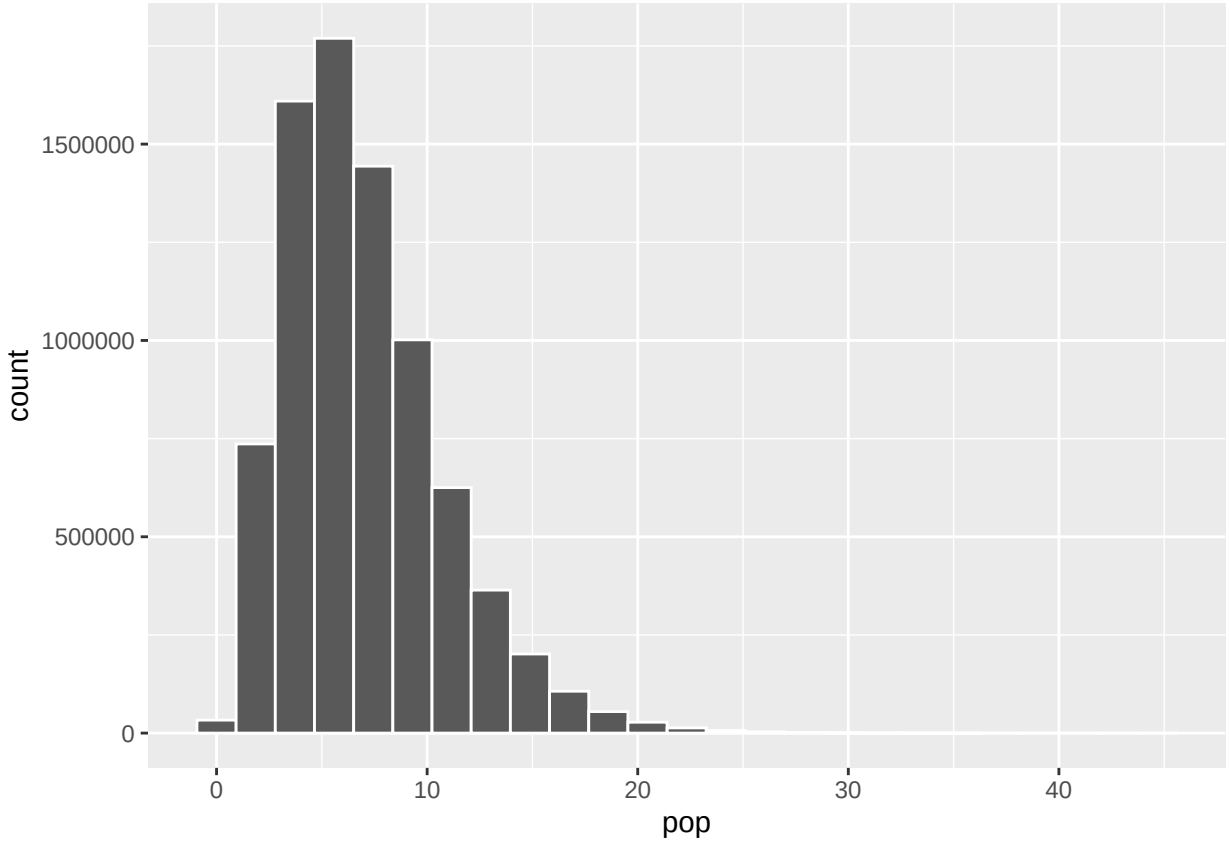
2.2

Generate values for a simulated "population." The population should have a chi-squared distribution with 7 degrees of freedom. HINT: Use the `rchisq()` function. Your N should equal 8,000,000.

Then: Create a histogram of your population and estimate and report the population mean μ and standard deviation σ.

NOTE: You need not worry about degrees of freedom as a concept at this point. But, if you are interested in a very brief statement on the subject, see *Introductory Statistics* by David Lane, p. 664.

```r
# Create population with chi-squared distribution
pop = rchisq(n=8000000,df = 7)
df = data.frame(pop)
# Pop histogram
ggplot(df, aes(pop)) + geom_histogram(bins = 25, col='white')
```




```r
# Pop parameters
(pop_mean <- mean(pop))
```

```
## [1] 6.999939
```

```r
(pop_sd <- sd(pop))
```

```
## [1] 3.740408
```

2.3

Take one sample of size $n = 50$ from your population. Then, calculate the sample mean.

```r
# 1 srs of size 50
sample_50 <- sample(pop, size=50, replace=T)

# mean of srs
(sample_50_mean <- mean(sample_50))
```

```
## [1] 7.19482
```

2.4

You have just created one realization of the sample mean. But to illustrate the CLT, you must lay out the variable's distribution. So, create a vector of size 1,000. Then, repeat the sampling process above 1,000 times, saving the mean of each sample in your newly created vector (i.e., the mean of sample 1 will be stored in the first element of the vector, the mean of the second sample will be stored in the second element of the vector, and so on.)

```r
# Create empty vector of length 1,000
n <- 1000
x_mat <- replicate(n, sample(pop, size=50, replace=T))
# Take 1,000 samples of size
# Calculate the mean for each sample and save in vector
sample_means <- apply(x_mat, 2, mean)
head(sample_means)
```

```
## [1] 6.378152 6.254955 6.956290 6.635847 7.271480 6.591475
```

2.5

The vector you created should contain the means of 1,000 samples. Now estimate its mean and standard deviation and report them

```r
(mean_sample_means <- mean(sample_means))
```

```
## [1] 6.994861
```

```r
(sd_sample_means <- sd(sample_means))
```

```
## [1] 0.5221533
```

```
# (sd_sample_means * sqrt(50))
```

2.6

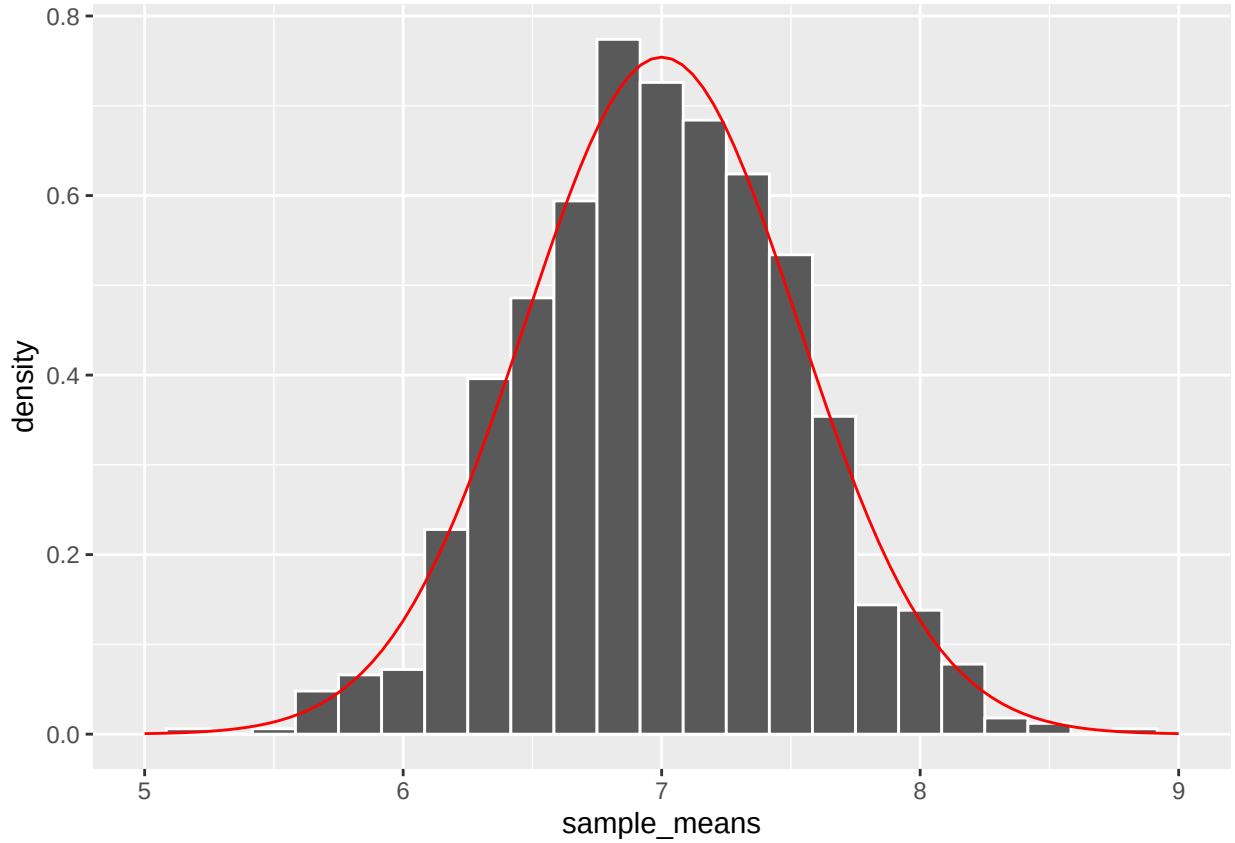
According to the CLT, what is the the theoretical distribution that the variable contained in this vector should follow? What this suggests, according to the Central Limit Theorem, is that the theoretical distribution of the variable should be about the mean sample means, or 6.994861, and the stand deviation should also be about the sd_sample_means times the square root of the sample size, or about 3.692182 # 3

3.1

Make a histogram of the sample means. Use the curve function to add a line that plots the "ideal" theoretical distribution proposed by the CLT. Briefly describe the curve's shape?

```
sample_means_df <- data.frame(sample_means)

p1 <- ggplot(sample_means_df, aes(sample_means)) +
  geom_histogram(aes(y = ..density..), bins=25, col='white') +
  xlim(5, 9) +
  # stat_function(fun = dnorm, args = list(mean = mean_sample_means, sd = sd_sample_means),
  #               col='blue') +
  stat_function(fun = dnorm, args = list(mean = pop_mean, sd = pop_sd / (sqrt(50))),
                col='red')

p1
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```




```
# note the blue represents the distribution of our own sample data,
# and the red represents the distribtuion expected using CLT
```

3.2

Above, each sample drawn from the population had 50 observations (i.e., $n = 50$). Take another 1,000 random samples from the population, but this time, increase the number of observations in each sample to 1,000 (i.e., $n = 1000$).

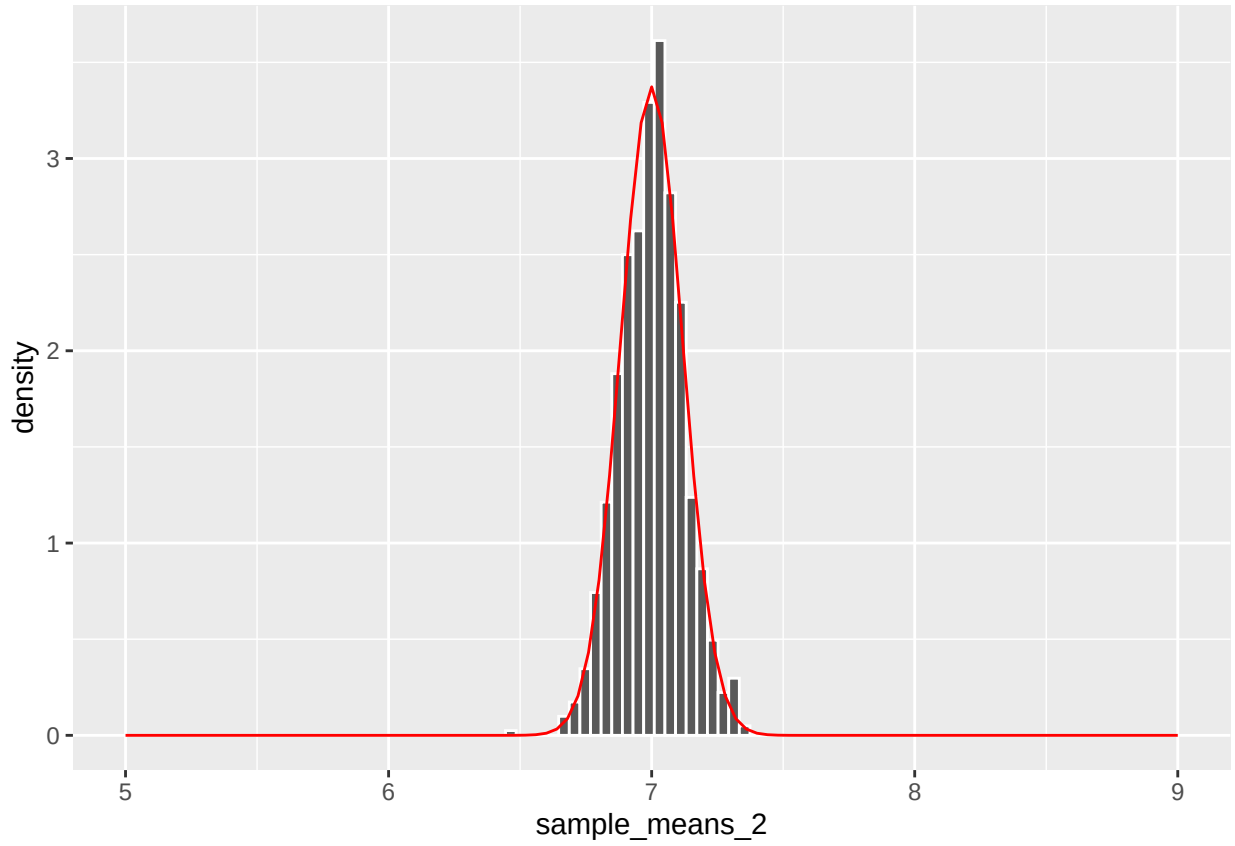
Then, create a new histogram and compare it to your earlier one. (You may need to format the x axes to make the comparison crystal clear.) a) How do the distributions differ? b) Which one better approximates the theoretical distribution proposed by the CLT? c) Why is that? d) What does your comparison suggest about the relationship between sample size (i.e., number of observations in each sample) and the variability in our estimates?

```
# Create empty vector of length 1,000
x_mat_2 <- replicate(1000, sample(pop, size=1000, replace=T))

# Take 1,000 samples of size
# Calculate the mean for each sample and save in vector
sample_means_2 <- apply(x_mat_2, 2, mean)
head(sample_means_2)
```

```
## [1] 7.103869 7.117004 6.867080 7.007188 6.908158 6.866711
```

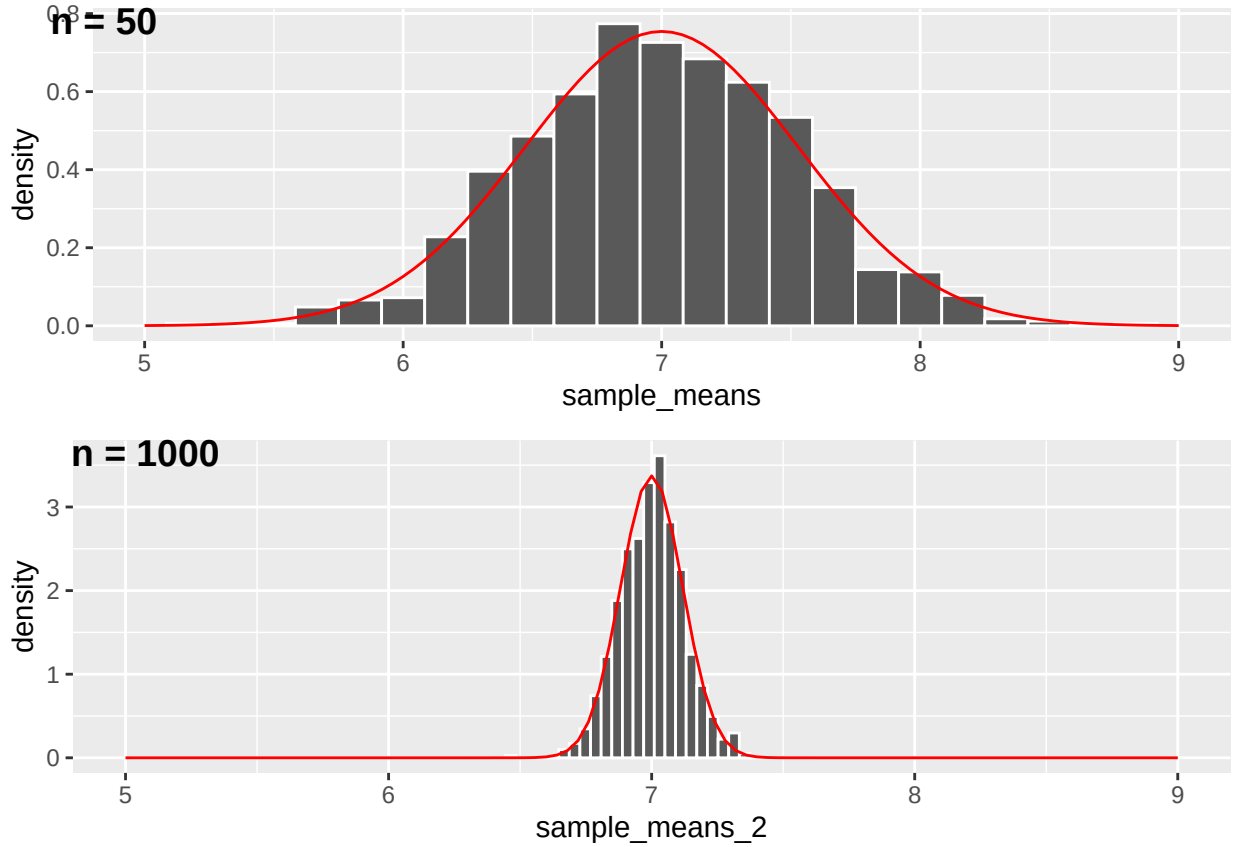
```
sample_means_df_2 <- data.frame(sample_means_2)

p2 <- ggplot(sample_means_df_2, aes(sample_means_2)) +
  geom_histogram(aes(y = ..density..), bins=100, col='white') +
  xlim(5, 9) +
  # stat_function(fun = dnorm, args = list(mean = mean_sample_means, sd = sd_sample_means),
  #               col='blue') +
  stat_function(fun = dnorm, args = list(mean = pop_mean, sd = pop_sd / (sqrt(1000))),
                col='red')

p2
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```




```
ggarrange(p1, p2,
  labels = c("n = 50", "n = 1000"),
  ncol = 1, nrow = 2)
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
## Removed 2 rows containing missing values (geom_bar).
```



a). The distributions differ in mainly their variance/standard deviation, in that since for n = 50, the sample means are more likely to vary more than if we were to take the mean of a sample of size 1000, the overal distribtuion of the means will be far more sparse for n = 50 compared to the distribution when the size of the sample is 1000.

b). note that looking at these two graphs, its not entire clear which one would better represetn the theoretical distribution proposed by the CLT. One things to better visualize this however, is what if we chose a sample size of 1? Then it is clear that the distributions of this small sample size would not give us a nice distribution if we replciate the sampling a few times. What this suggests is that the greater the sample size, the more we can rely on its distribution as being normally distribution, and thus the 2nd graph (n = 1000) is a better representation of the theoretical distribution proposed by the CLT.

c). The reason for this is as sort of suggested in the previous question. CLT tells us that in fact, as n approaches infinity, the distribution of the sample mean approximates the normal distribution (meaning it is a better representation of the distribution proposed by CLT).

d). This comparison suggests that as sample size increases, the variability in our estimates decreases (which is why we also see in n = 1000 how much "tighter" the distribution is).

4

**Work from the plots that you just produced to write a short (300 words maximum) article that explains the significance of the Central Limit Theorem to readers who are unfamiliar with statistics. Use simply language; avoid jargon. Focus on the "meaning" of the CLT. Why should readers care about it?

Why should the people care about CLT? The reason why we should care is that it can take any population data, and tell us something about its distribution as long as we supply it enough samples. The normal

distribution is such a fundamental principle in statistics: in a sense, the normal distribtuion is very *natural*: it exists in our world everywhere. When we roll a few dice, and simulate it the sum of the dices we get each time, we get a normal distribtuion. When we look at the distribution of heights across a region, the grades of scores, the quanitity of proteins in our body, and so on, the normal distribution is truly something that is a fundamental property of *randomness*.

Then, it is beautiful that no matter what data we supply it (which does not necessarily have to follow the normal distribution). Take the example of rolling an unfair dice for example that can never be 2 or 5, for example, and have a high probability of being 1 and 6, and a low probability of being 3 and 4. This clearly not a normal distribution, and is more like two spikes at the end a small hill in the middle in terms of distribution. Even if this is the case, we can simulate *samples* many times, and plot this to get a normal distribution that still has the mean of the population sample. In essense, taking samples of the population and reiterating it can reveal a fundamental property of the population as a normal distribution.

There is the added benefit that sometimes, we cannot direction emasure the distribution of the population, for that is cumbersome and more often than not impossible. What we can do, is take a sample. And the CLT tells us that if we take enough samples, we can get a normal distirbution that tells us something about the actual popuklation, even if we never take the whole population!