Homework 3

Student

```r
library(ggplot2)  # you’ll need this package for the R section

# get ‘play()‘ function for #1
source("https://gist.githubusercontent.com/cwickham/abe3b4c4ba5319e8e1dd5102541f2117/raw")
```

Instructions

Use the R Markdown version of this file to complete and submit your homework. Items in **bold** require an answer. Make sure you change the author in the header to your own name.

Conceptual Questions

1. Describe why we do not usually know the population mean. What statistic do we usually use to estimate the population mean and why?

   We don’t ususally know the population mean because to know it, we would need to know all the values in the population, i.e we would have a census. If we knew all the values in the population, we wouldn’t be trying to estimate its properties by taking a sample, we could calculate them directly.

   We usually estimate the population mean, by taking the average of a sample from the population, i.e. we estimate the population mean with the sample mean. This is a “good” estimate because it is unbiased (i.e. its expected value is the population mean), and it gets better for larger samples (i.e. its variance decreases with increasing sample size). The Central Limit Theorem also suggests that if our sample size is large enough the sample mean will have a sampling distribution that is approximately Normal.

2. Consider two hypothetical histograms:

   i) a histogram of a sample of size n from the population and,
   ii) a histogram of k sample means from samples of size n from the population.

   a) For large values of n, which of the above histograms would give you a good estimate of the population distribution?

      Histogram i)

   b) Which of the above histograms is an estimate of the “sampling distribution of the sample mean for samples of size n”?

      Histogram ii)

   c) Describe ii) in relation to the population distribution. E.g. how do its center, spread and shape compare to the population distribution?

      Histogram ii) will have the same center as the population. As the sample size, n, increases histogram ii) will get narrower and closer in shape to a Normal distribution.

d) Consider these two true statements:

“For large values of n the sampling distribution of the sample mean approaches a Normal distribution”

“For large values of k the histogram in ii) approaches the true sampling distribution of the sample mean.”

One is a consequence of the “Central Limit Theorem” and the other is a consequence of the “Law of Large Numbers”. Which is which?

“For large values of n the sampling distribution of the sample mean approaches a Normal distribution” is a consequence of the **Central Limit Theorem**, which descirbes the behaviour of the sampling distribution of the sample mean.

“For large values of k the histogram in ii) approaches the true sampling distribution of the sample mean.” is a consequence of the **Law of Large Numbers.** *In this case, the Law of Large numbers provides justification that a large simulation (large k) allows us to get a good picture of the sampling distribution of an estimate.*

R questions

1. Consider this game: You roll one die, and lose $50 if you roll a 1, but win $15 if you roll anything else. I’ve written a function for you, `play()` that plays this game (the line starting `source(` in the code chunk at the top of this document gets this function for you).

   You can play by calling the `play()` function

```r
play()
```

```
## You rolled a 3. Your payout is $15.
```

```
## [1] 15
```

The function `play()` returns a numeric value of either `-50` or `15` depending on your roll, and prints out the result. When you simulate many games, the print out will be time consuming, so use `play(silent = TRUE)` to play without printing results.

a) **Use simulation to estimate your expected win/loss value for one roll.** Hint: simulate many plays of the game and take the average of the outcomes.

```r
many_plays <- replicate(5000, play(silent = TRUE))
(estimate <- mean(many_plays))
```

```
## [1] 4.769
```

I estimate the expected win is $4.77.

b) **How many times did you play to find your estimate? How precise do you think your answer is?**

I played 5,000 times. One way to understand how accurate the estimate is to repeat the simulation a few times:

```r
mean(replicate(5000, play(silent = TRUE)))
```

```
## [1] 4.262
```

```
mean(replicate(5000, play(silent = TRUE)))
```

```
## [1] 4.028
```

```
mean(replicate(5000, play(silent = TRUE)))
```

```
## [1] 4.236
```

Looks like the estimates don't change for the first significant figure, so I'd say my estimate was accurate to about \$0.50.

An alternative, but not expected, answer

We could estimate the standard deviation of the game outcome based on our many plays:

```
(sd_play <- sd(many_plays))
```

```
## [1] 23.6739
```

Then use the CLT to approximate the standard deviation of the mean of a sample of size 5000:

```
(se_mean <- sd_play/sqrt(5000))
```

```
## [1] 0.3347996
```

Our estimate should be accurate to about ± \$0.67 (2 standard errors).

c) **How much would you be willing to pay to play this game?**

I'd be willing to play for any price less than the expected win, so accounting for some error, maybe about \$4.

FYI, the true expected value of this game is:

$$\frac{1}{6}(-50) + \frac{1}{6}(15 + 15 + 15 + 15 + 15) = 4.167$$

2. In lab you explored the Central Limit Theorem when the population distribution was a Gamma(5, 1). The amazing thing about the Central Limit Theorem is that it applies no matter the shape of the distribution (as long as the distribution has an expected value, and a finite variance). For this question, **choose one of the following distributions, and replicate the exploration from the lab with sample sizes of 2, 10, 50 and 100**:

   - Continuous uniform on (0, 1), see `?runif`
   - Discrete uniform on $1, \ldots, 10$, use `sample()`
   - A poisson distribution with your choice of parameter, see `?rpois`
   - Beta distribution with both parameters set to 0.5, see `?rbeta`

   Write up your exploration in a way that a reader can follow without having to understand your code.

A good answer should:

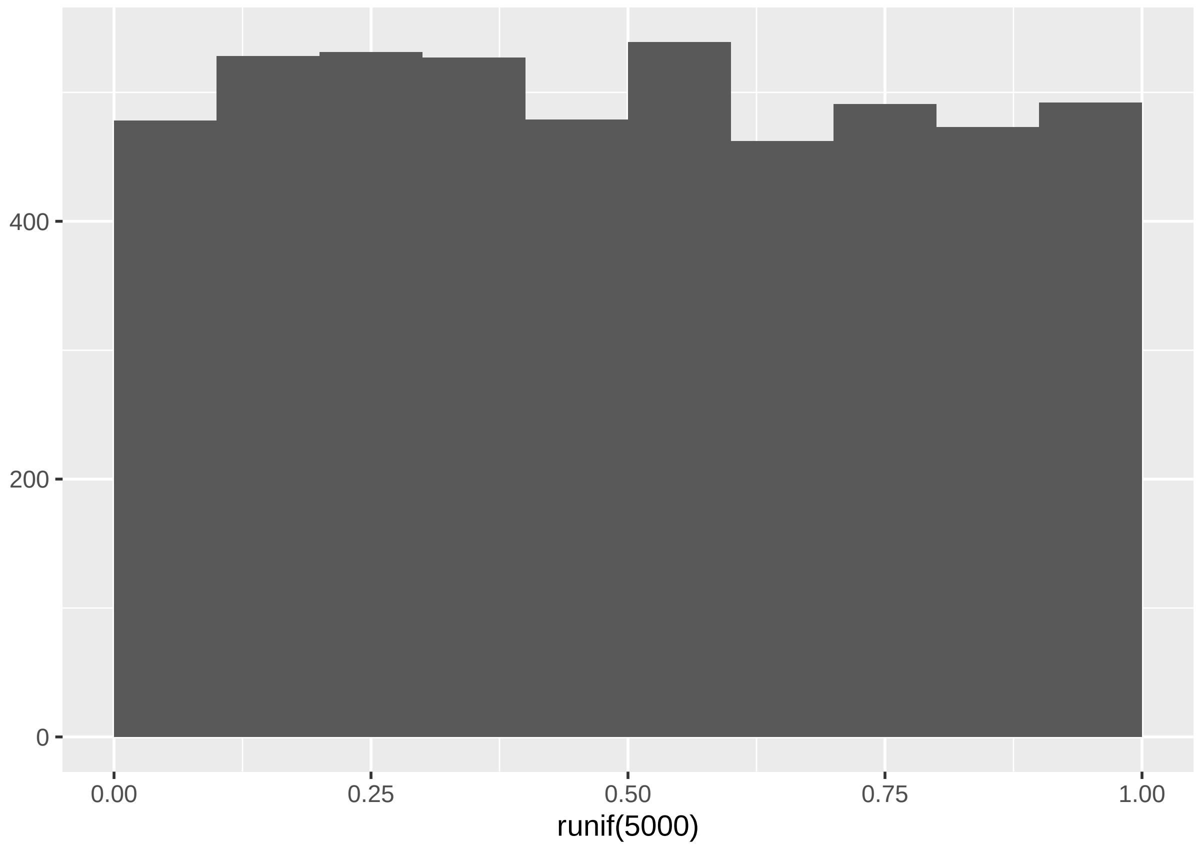
- show histograms of the sampling distribution of the sample mean for all sample sizes
- explicitly describe the features of these histograms that illustrate the CLT at work

An example answer follows

Continuous uniform

To get a feel for the population distribution's shape we can simulate a large sample from the population and examine a histogram:

```
qplot(runif(5000), boundary = 0, binwidth = 0.1)
```



The continuous uniform on (0, 1) has rectangular shape and is centered on 0.5.

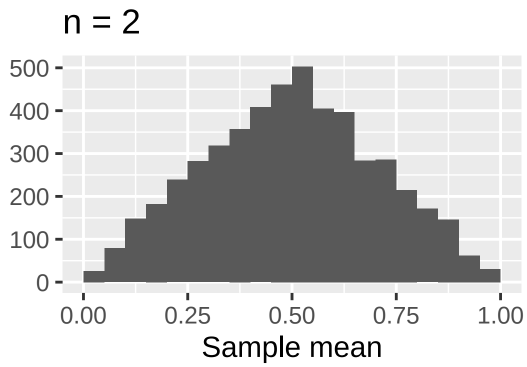
For each sample size, $n = 2, 10, 50, 100$, we repeat 5,000 times:

1. Draw a sample of size n, from a continuous Uniform distribution on (0, 1)
2. Take the mean of the sample in 1.

```
n_sim <- 5000
cunif_2 <- replicate(n_sim, mean(runif(n = 2, min = 0, max = 1)))
cunif_10 <- replicate(n_sim, mean(runif(n = 10, min = 0, max = 1)))
cunif_50 <- replicate(n_sim, mean(runif(n = 50, min = 0, max = 1)))
cunif_100 <- replicate(n_sim, mean(runif(n = 100, min = 0, max = 1)))
```

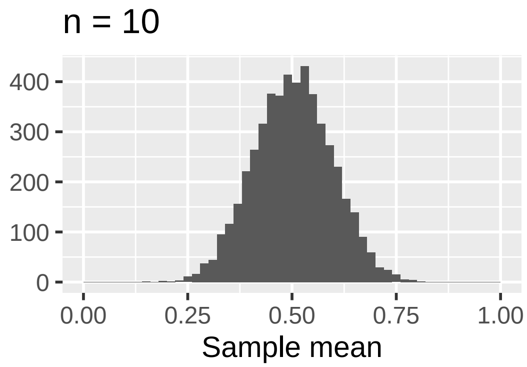
For a sample of size 2, the sampling distribution of the sample mean, is still centered on 5, covers most of the (0, 1) range, but is much more triangular in shape:

```
qplot(cunif_2, xlim = c(0, 1), boundary = 0, binwidth = 0.05) +
  labs(title = "n = 2", x = "Sample mean")
```



For a sample of size 10, the sampling distribution of the sample mean, is still centered on 5, is noticeably narrower and is starting to look more bell shaped.

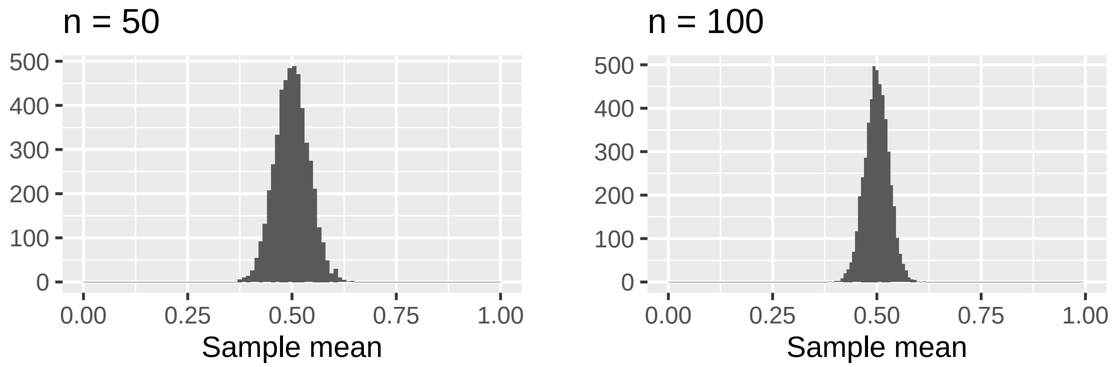
```
qplot(cunif_10, xlim = c(0, 1), boundary = 0, binwidth = 0.02) +
  labs(title = "n = 10", x = "Sample mean")
```



For a samples of size 50 and 100, we see the behavior we expect from the CLT, the sampling distribution of the sample mean, is still centered on 5, continues to get narrower for larger samples and is closer in shape to a Normal distribution:

```
qplot(cunif_50, xlim = c(0, 1), boundary = 0, binwidth = 0.01) +
  labs(title = "n = 50", x = "Sample mean")
qplot(cunif_100, xlim = c(0, 1), boundary = 0, binwidth = 0.007) +
  labs(title = "n = 100", x = "Sample mean")
```

```
## Warning: Removed 1 rows containing missing values (geom_bar).
```


n = 50
Sample mean
n = 100
Sample mean