Homework 3

Nora Quick

2021-10-11

```
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 usually don't know the population mean because we would need to know the answers of the POPULATION. i.e we would need to know the answers of every person within a population and that isn't always plausible. According to the lecures a more practical way to estimate population mean is to use sample means. And by using the Central Limit Theorem we would eventually get the population mean. We do this because a sample mean is much easier to get and with a lot of them we can get the population mean.

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) because Law of Large Numbers

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

      Histogram (ii) because it's a distibution of many samples

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

      The center should be the same center as the population. As n increases the spread and shape of the distribution will become more normal.

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?

The first statement has to do with the Central Limit Theorem because it talks about the sampling distribution in relation to the sample mean resulting in a "Normal distribution". The second statement has to do with The Law of Large Numbers because it relates to a large value resulting in the true sampling distrubution.

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

```
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.

```
x <- replicate(10000, play(silent = TRUE))
estimate <- mean(x)

estimate
```

```
## [1] 4.5155
```

```
s <- sd(x)
m <- s / sqrt(10000)

m
```

```
## [1] 0.2390868
```

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

I played 10,000 (to maintain some consitency with lectures and quizzes). I think that with such a large n I am relatively precise with the answer I'm getting. After running it several times I was able to see that I am pretty precise in my estimations.

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

With a standard deviation of about \$0.5 and a winning of about \$4.25 I would not be willing to pay to play this game. However, if I had to I would spend about half of what I could make so about \$2.

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`

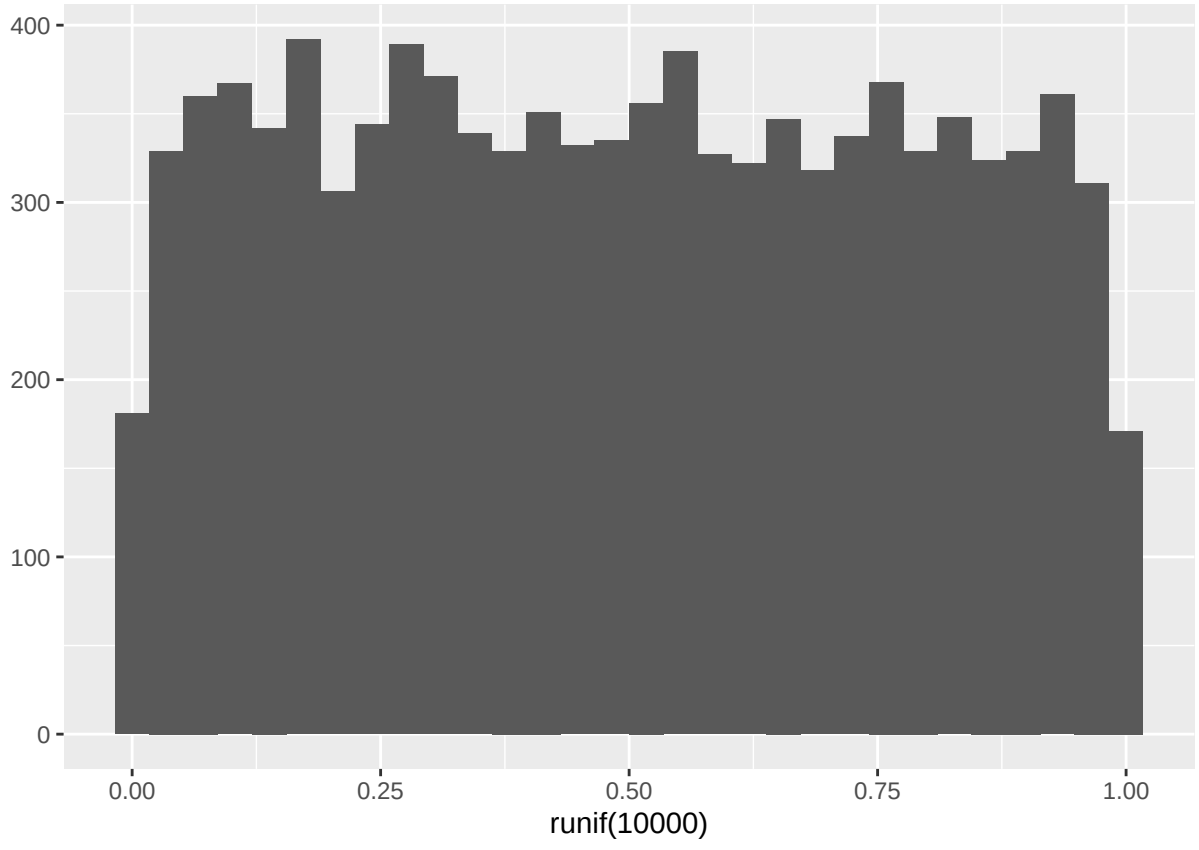
   Write up your exploration in a way that a reader can follow without having to understand your code.

Continuous distribution

For this exploration we need to start with a basic graph.

```
qplot(runif(10000))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



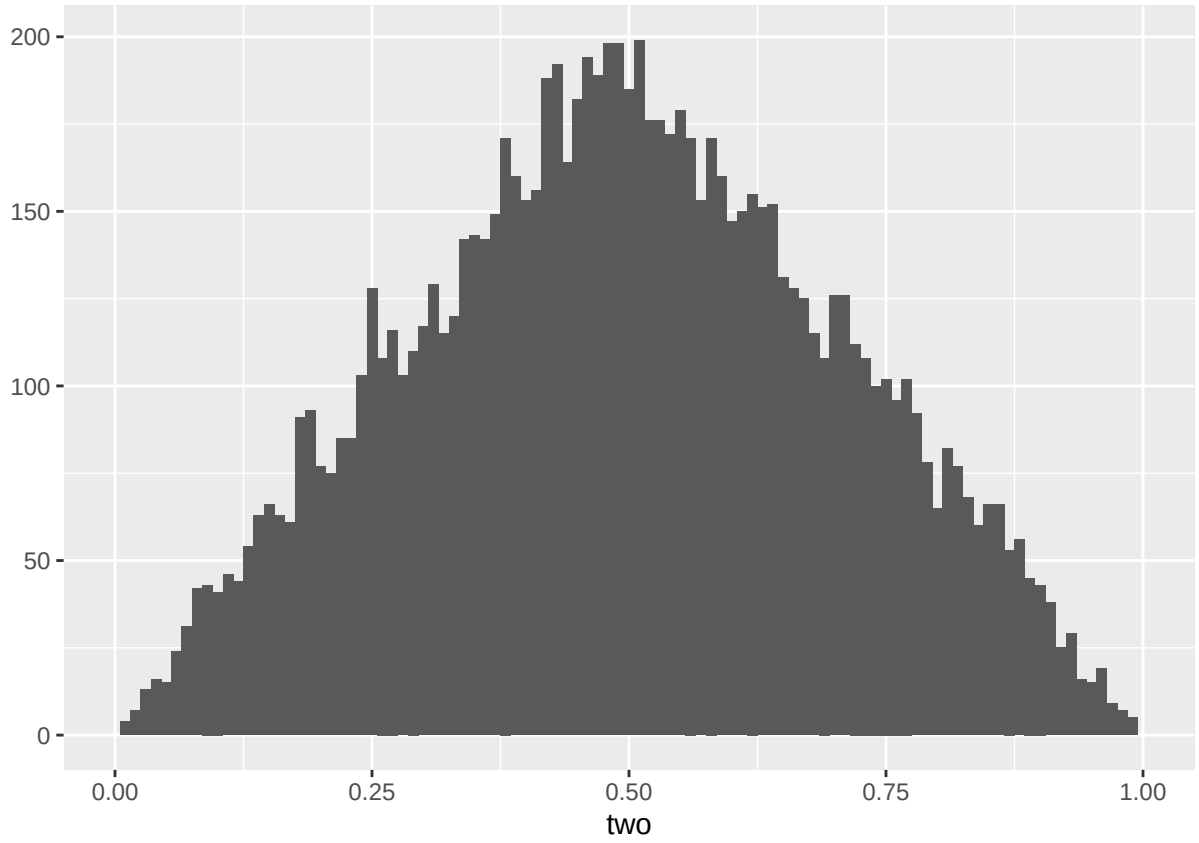
From looking at this graph we can see that there is a relatively consistant distribution throughout the 10,000 runs. This leaves us with a uniform graph which doesn't represent the mean very well. For this explanation of the Central Limit Theorem and how it works we will do the same test with different sample sizes (2, 10, 50, 100) to see how the distribution of the graph will change to show a true mean.

```
two <- replicate(10000, mean(runif(2, 0, 1)))
ten <- replicate(10000, mean(runif(10, 0, 1)))
fif <- replicate(10000, mean(runif(50, 0, 1)))
one <- replicate(10000, mean(runif(100, 0, 1)))
```

To begin we will look at the distribution of n = 2. This shows a semi-symmetric (unimodal) graph that is pretty spead out. There is a lot of variation in the sample means.

```
qplot(two, xlim = c(0, 1), binwidth = 0.01)
```

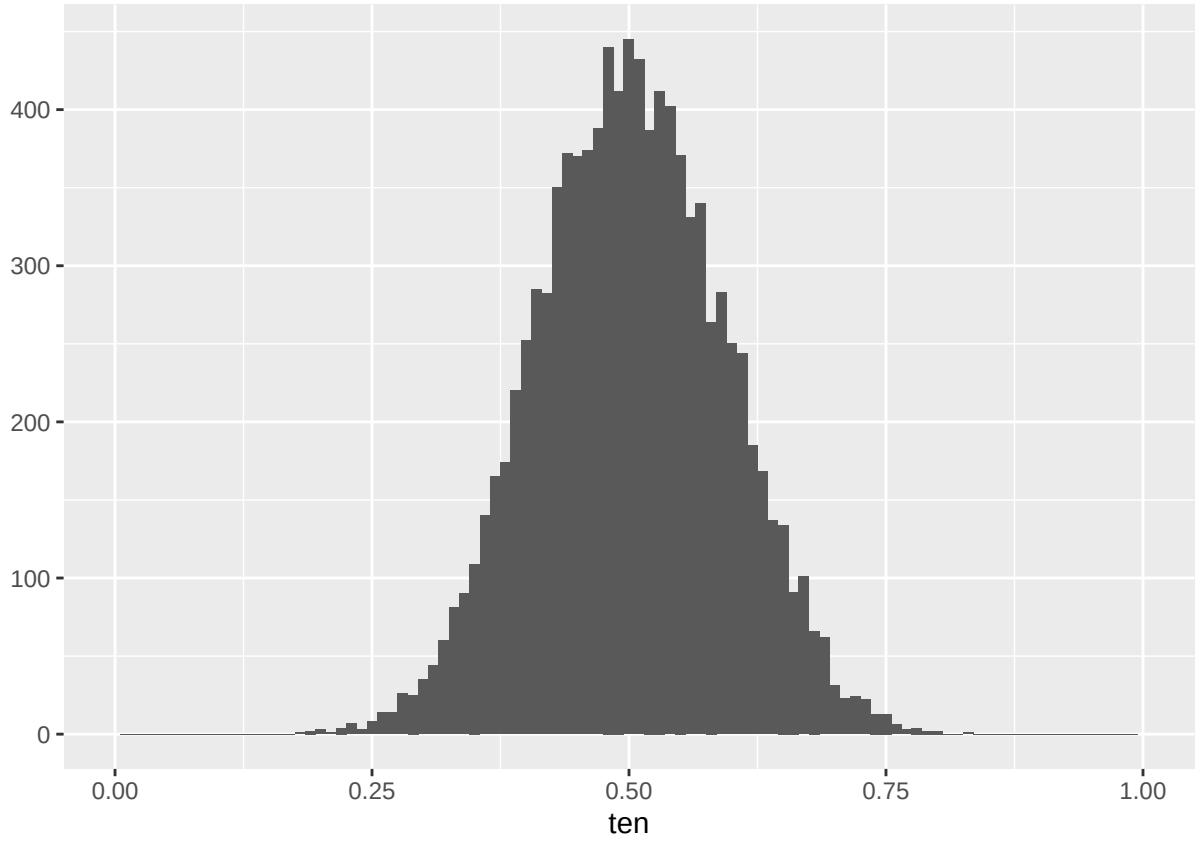
```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



Next, we have n = 10 which will begin to have less variance in its sample mean as it begins the process of showing 0.5 as the true mean.

```
qplot(ten, xlim = c(0, 1), binwidth = 0.01)
```

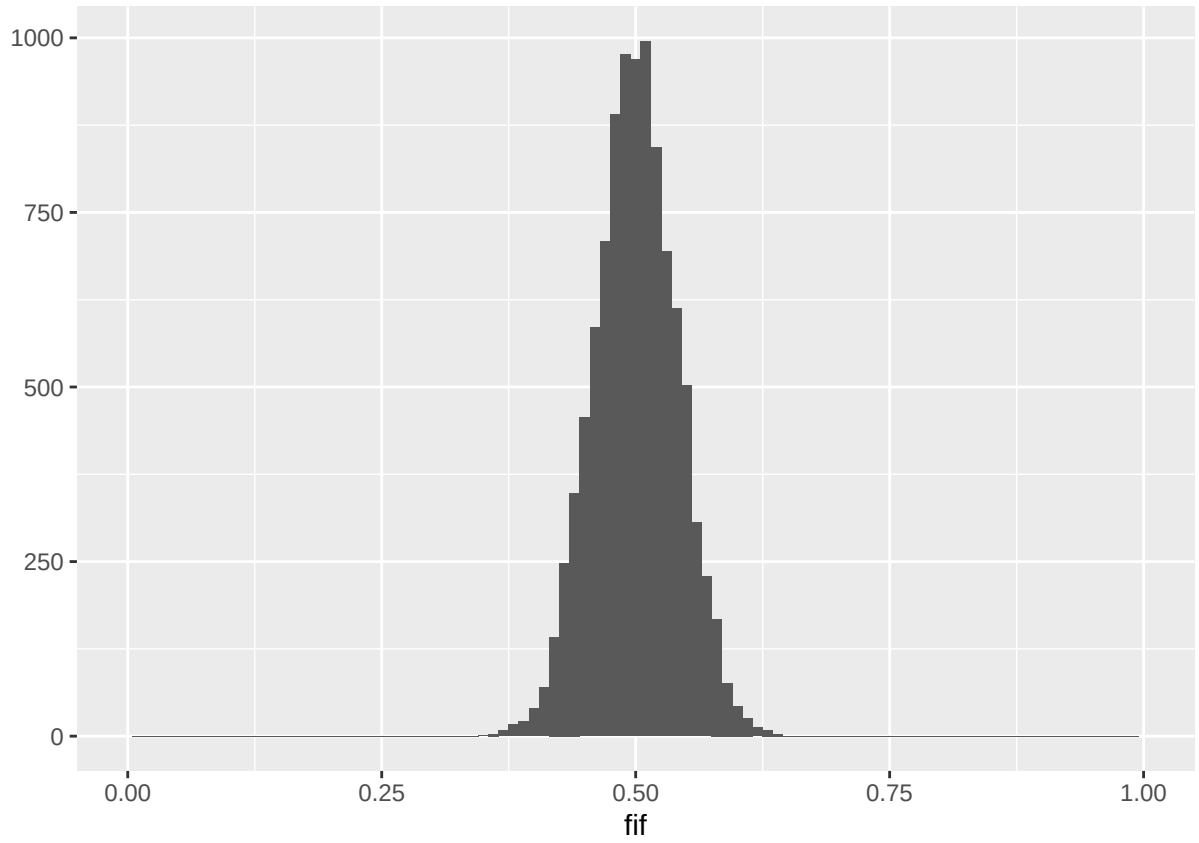
```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



At n = 50 we start to see a very clear graph to show the sample mean at 0.5. The data is less distributed due to the frequency in which a number closer to or at 0.5 is found.

```
    qplot(fif, xlim = c(0, 1), binwidth = 0.01)
```

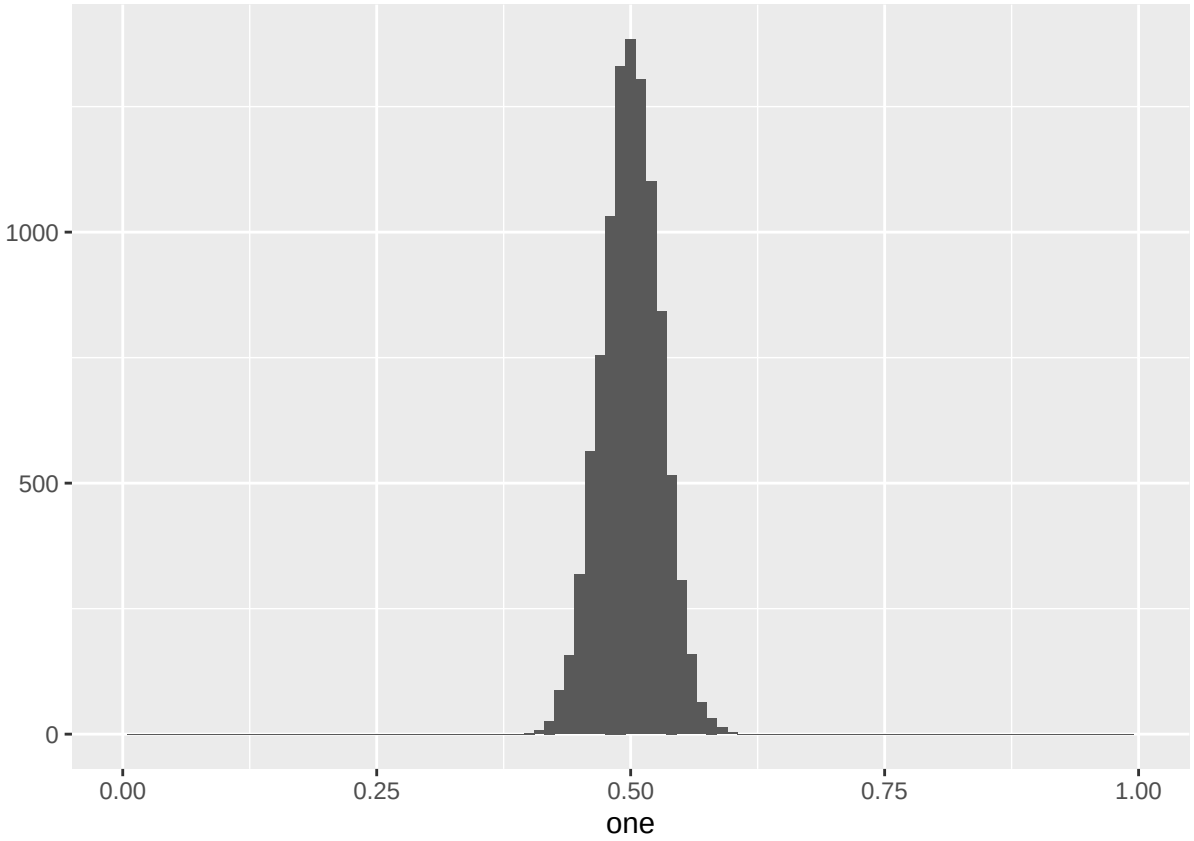
```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



Finally when we get to n = 100 we can see a much clearer picture of the sample means. We can see what the true mean is for the range 0-1.

```
    qplot(one, xlim = c(0, 1), binwidth = 0.01)
```

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



Throughout these images we can see the process of the Central Limit Theorem. As we get more and more means with larger values of n we are able to visually see the data get closer to the true mean. The shape becomes more normal in appearance and the peak at 0.5 becomes more pronounced.