James Scruggs

M01255052

6. central limit Theorem

Today we will explore the Central Limit Theorem.

The Central Limit Theorem describes the sampling distribution of the sample mean, so before we talk about it, we must discuss sampling distributions. Recall that a probability distribution describes the likely values of a random number and how probable each value is. A sampling distribution does the same thing, but for a statistic. Let's look at this a little more closely.

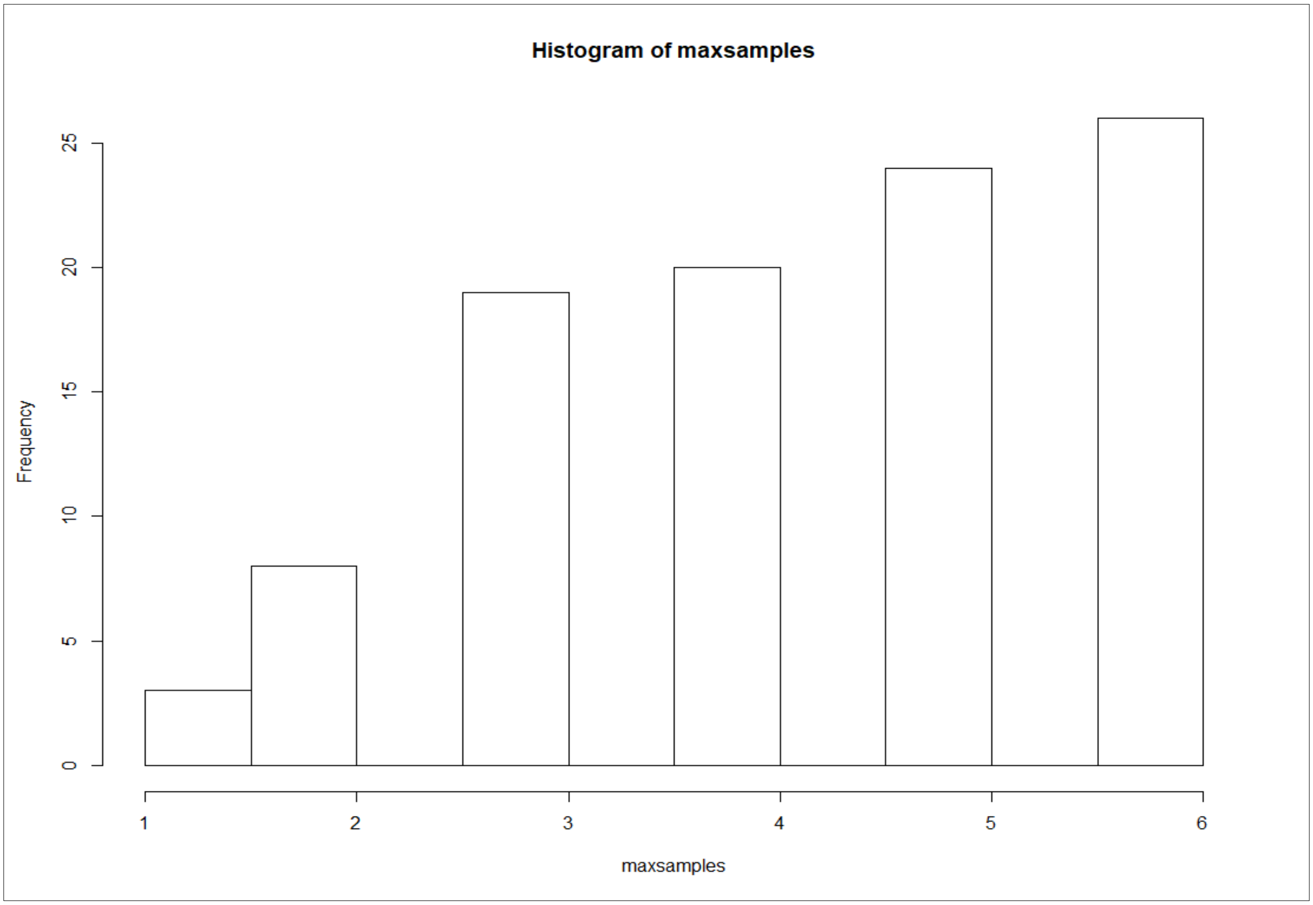
Suppose we roll a fair die twice, 100 times. *samples<-replicate(100,sample(1:6,2,TRUE)))*)

Once we've got the sample, we could do various things with it, depending on what question we are trying to answer. We could average the numbers, find the standard deviation of the numbers, find the minimum number, or find the maximum number. Each of these values, which are derived from a sample, are different statistics. One definition of "statistic" is "a number that describes a sample." For right now, let's focus on the maximum of the three values.

1. Create a new column that contains the maximum value of the two dice rolls in each row. Then create a histogram of the values in this column, and paste it here.
   In R:

   maxsamples <- apply(samples,2,max)

   hist(maxsamples)

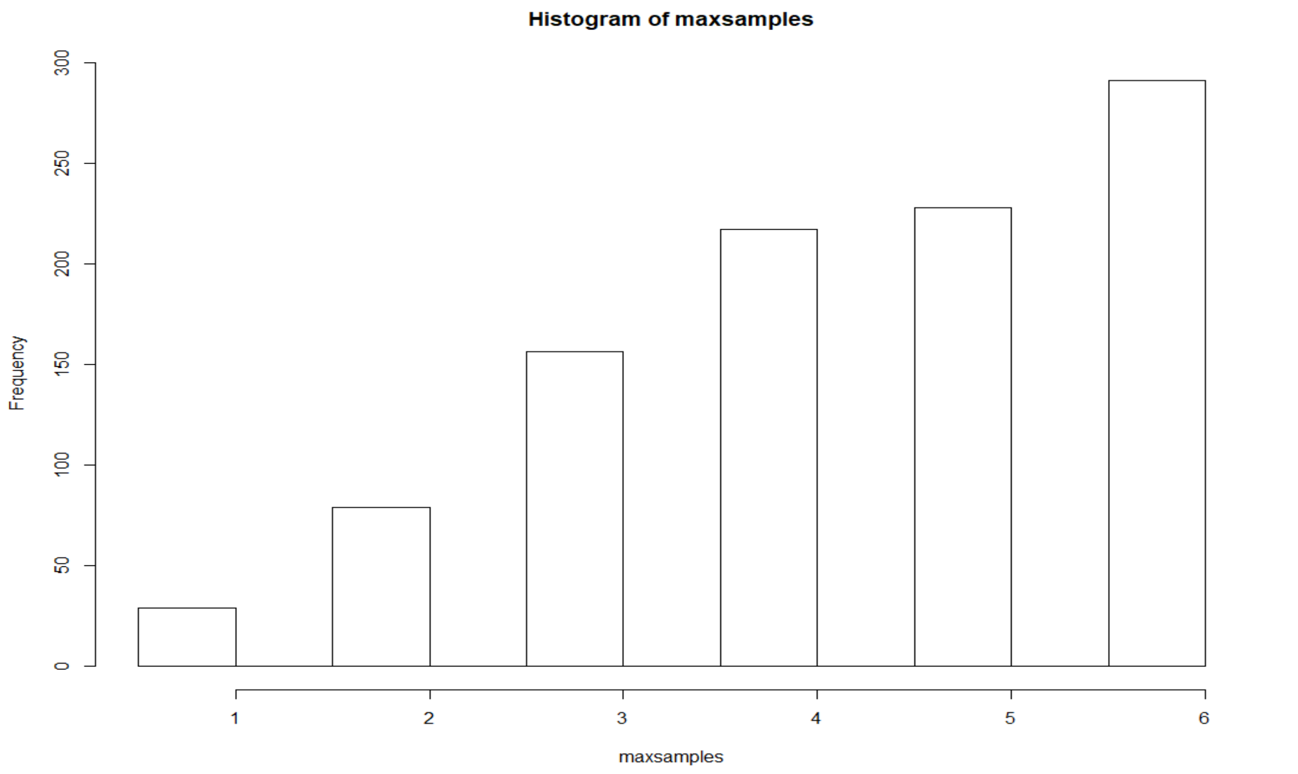


This graph summarizes the possible values that can be the maximum and tells you how frequent each of those values was in the 100 rolls. Of course, the more dice doubles we roll, the more accurate the resulting graph of maximum values will be.

If yours is behaving like mine, the bin at one is on the other side of the value “1” than the bin at 6 is. You can fix this by setting where the bins start and stop.

hist(maxsamples, breaks=seq(.5,6,.5))

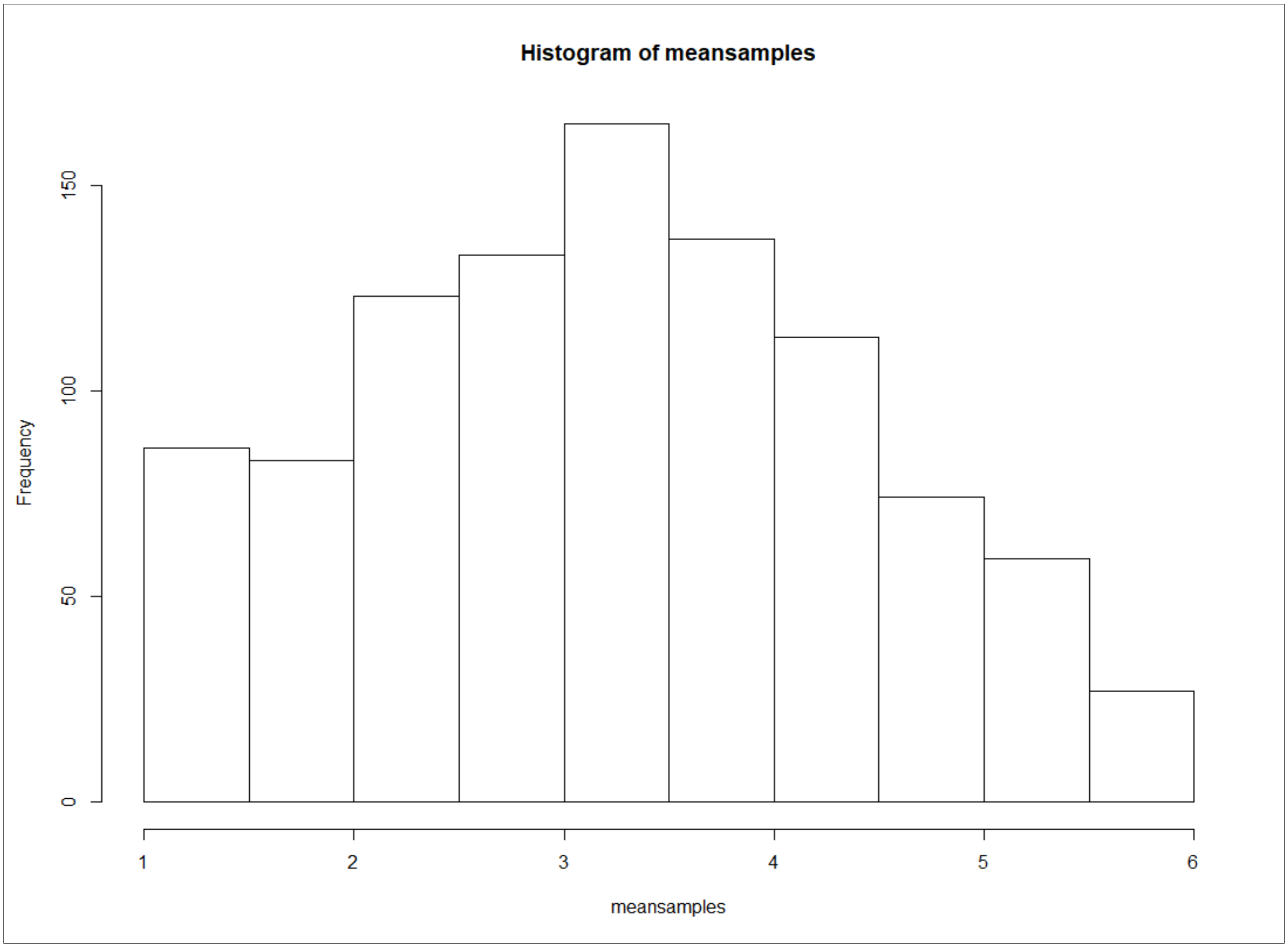
2. Repeat what you did above, but this time use 1000 samples instead of 100. Paste the new graph here, and describe the difference.



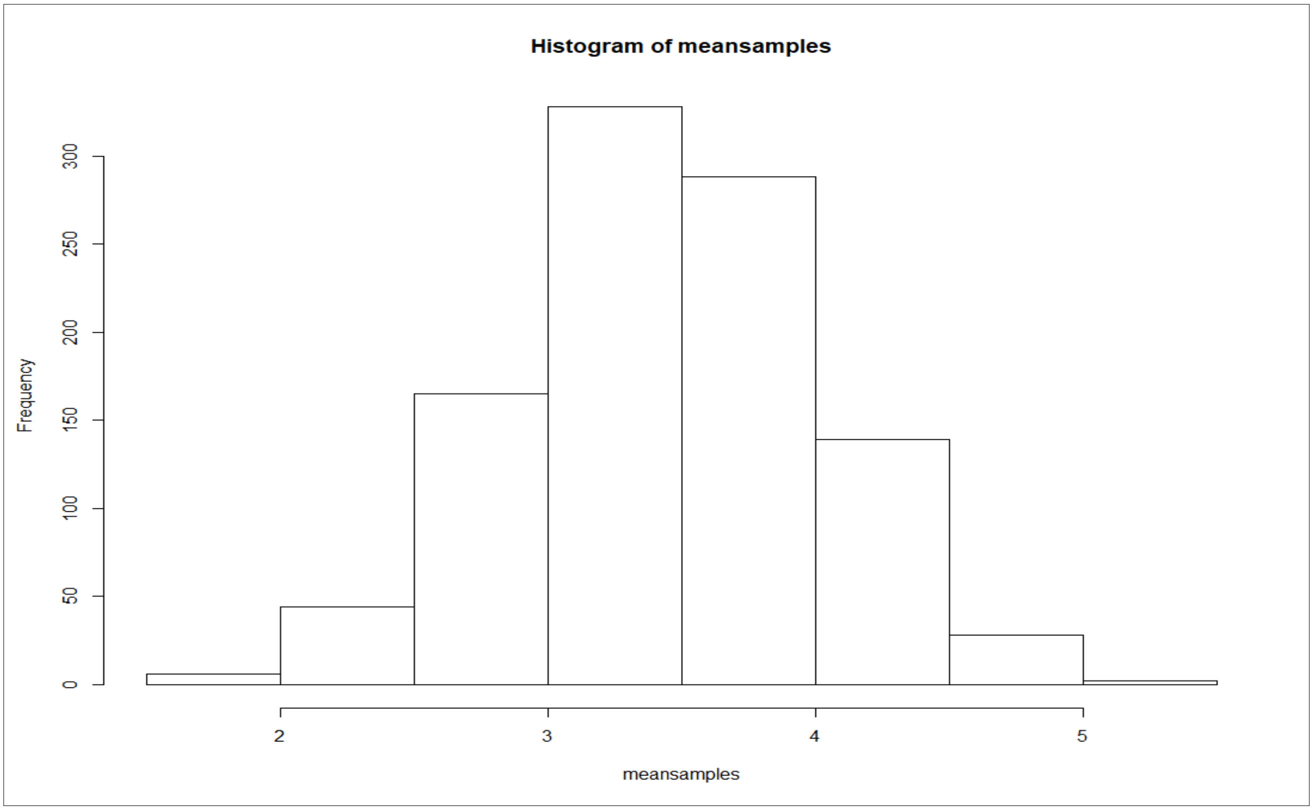
The data appears to be show nearly the same thing as the previous graph. The more samples taken, the higher the maximum values appears.

The Central Limit Theorem describes the shape of the distribution of sample means. Before we talk about the theorem itself, let's look at some graphs of sample means.

4. Starting with your dice roll data, create a vector that contains the sample means of the dice rolls. (*meansamples<-apply(samples,2,mean)*) Then create a histogram from the new vector. Paste your histogram here.



5. What if we had larger sample sizes? Recall that the sample size in our data was 2, and we had 1000 samples. Create some more columns of dice rolls, so that you have a sample size of 10, with 1000 samples. Then calculate the mean of these samples, and create a histogram of the means. Paste it here.



Notice that the bell shape of the normal distribution is showing up here. This turns out to be true in general: if you graph the distribution of sample means, you get a bell shape. This is not true of other statistics, in general. The Central Limit Theorem applies to sample means:

Central Limit Theorem:

If the sample size n is large, then the distribution of the sample means is

- normal (bell-shaped)
- with mean μ
- with standard deviation $\frac{\sigma}{\sqrt{n}}$

OK, so we already discussed the bell-shape. What about the mean and standard deviation? Look back above at the graph you made of the means of samples of size 2 (number 4). Compare it to the graph you made of the means of samples of size 10 (number 5). Notice that both graphs are centered at about the same place, but the graph for the larger sample size is less spread out than the graph for the smaller sample size. This is because both of them will be centered at the average value for the population, which for a dice roll is 3.5, and the standard deviation will be the standard deviation of the population (1.71 for a dice roll) divided by the square root of the sample size (which is 2 for number 4, and 10 for number 5). If we calculate the expected standard deviation for the sample in number 4, we get 1.21 and if we calculate the expected standard deviation for the sample in number 5, we get 0.54.

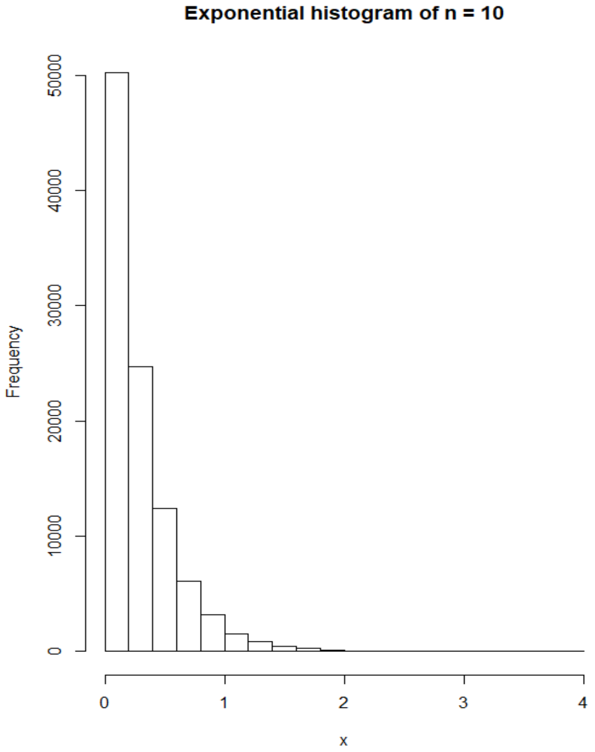
6. Calculate the standard deviations for the vectors you used in numbers 4 and 5, and compare them to the expected standard deviations. (*sd(meansamples)*)

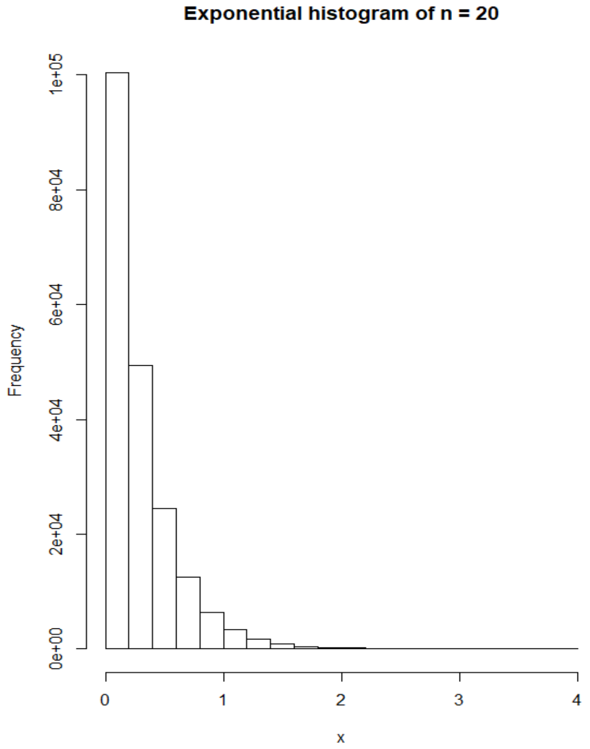
For a sample size of 2, I got a standard deviation of 1.197 which is close to the expected standard deviation.

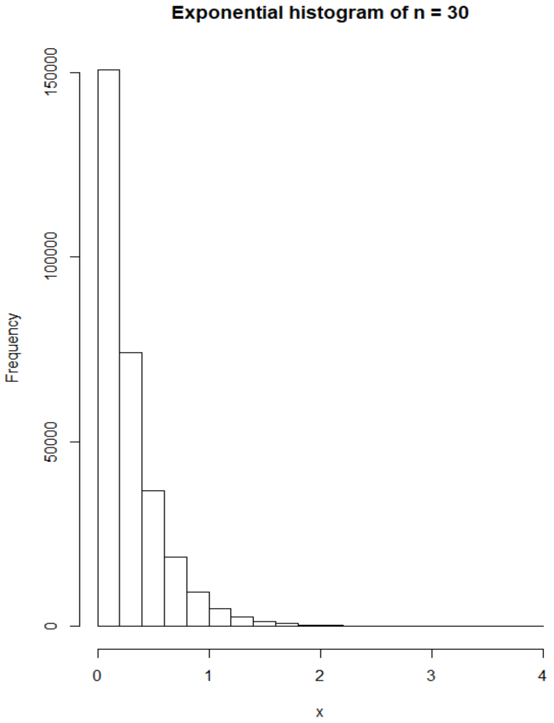
For a sample size of 10, I got a standard deviation of 0.538 which is also close the expected standard deviation.

Now, the hard part is the first phrase in the statement of the theorem: “If the sample size n is large...” How large is large enough? It turns out that the answer is “it depends.” We saw that with dice rolls, a sample size of 10 was enough, but in general you need more. We'll stick with 30 as a general rule: if n>=30, we can assume we know the distribution of the sample mean, if n<30, we don't.

7. Select either an exponential distribution (*rexp(samplesize, mean)*) or a binomial distribution with a p of .1 (*rbinom(samplesize, n, p=.1)*). Select 10000 samples of size 10, calculate the mean of each sample, and create a histogram. Then select 10000 samples of size 20, calculate the mean of each and create a histogram of means. Then select 10000 samples is size 30, calculate the mean of each sample and create a histogram. Below, paste your code and your three histograms. Comment on what happens as the sample size increases.


Exponential histogram of n = 10
Frequency
0
10000
20000
30000
40000
50000
0
1
2
3
4
x


Exponential histogram of n = 20
Frequency
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
0
1
2
3
4
x


Exponential histogram of n = 30
Frequency
0
50000
100000
150000
0
1
2
3
4
x

```
# n = 10

samplesize<-replicate(10000,sample(1:6,10,TRUE))

meansamples<-apply(samples,2,mean)

x <- rexp(samplesize, meansamples)

hist(x,main="Exponential histogram of n = 10")


# n = 20

samplesize<-replicate(10000,sample(1:6,20,TRUE))

meansamples<-apply(samples,2,mean)

x <- rexp(samplesize, meansamples)

hist(x,main="Exponential histogram of n = 20")


# n = 30

samplesize<-replicate(10000,sample(1:6,30,TRUE))

meansamples<-apply(samples,2,mean)

x <- rexp(samplesize, meansamples)

hist(x,main="Exponential histogram of n = 30")
```

It appears that all histogram distribution of data end up looking exactly alike.

8. Assume we interview 150 people and ask them the mileage on their car. If we assume that the population mean mileage is 100000 miles, with a population standard deviation of 30000 miles, what does the Central Limit Theorem tell us about the distribution of the sample mean?

Because our sample size (150) is >= 30 we can use the central limit theorem. The sample will be normal with the mean of 100,000 and the SD = (30,000 / sqrt(150)) = 2450.

9. In the situation in number 7, what is the probability that we get a sample mean between 99000 and 101000?

```
NormSamples<- replicate(1000,rnorm(150,mean=100000,sd=30000))

meansamples<-apply(NormSamples,2,mean)

count  <-  sum(meansamples<101000&meansamples>99000)

prob  <-  count/1000  #gives  an  empirical  estimate  of  the  probability
```

#To get the theoretical probability, use the command pnorm:

#Calculate the mean and standard deviation of the sample mean from the Central Limit Theorem. Suppose we call them meanofmeans and sdofmeans.

#then use pnorm(101000, meanofmeans, sdofmeans)-pnorm(99000, meanofmeans, sdofmeans)

Empirical Probability = 33.6%

Theoretical Probability = 33.3%

10. In the previous problem, you were given code for finding an empirical probability and for finding a theoretical probability. What is the difference between empirical and theoretical probability?

Empirical probability was given by the actual data from our tests. In the empirical probability, we took the total number of times the event occurred (number of times mean was in between the 101,000 and 99,000) and divided it by total number of incidents (1,000). This is what ”actually” happened.

Theoretical probability was given by all the number of way the event could happen divided by the all the possible outcomes. This is what we “expect” to happen.

11. What you can say about, normal (bell-shaped), mean μ , and standard deviation $\frac{\sigma}{\sqrt{n}}$ for large samples.

The distribution of the sample mean is equivalent to all three, and it is normally distributed.