Math 275 October 10, 2019
K.Tosun

Simulations for the Sampling Distribution and Central Limit Theorem

Suppose that we draw all possible samples (of the same size) from a population. Suppose further that we compute a statistic (e.g., a mean, proportion, standard deviation) for each sample. These values vary from sample to sample. Different random samples taken from the same population will give different statistics. But there is a predictable pattern in the long run. The distribution of a statistic calculated from all possible samples of the population is called the **sampling distribution**.

The number of possible samples is much larger than the population of individual values. There are $\binom{N}{n} = \frac{N!}{n!(N-n)!}$ different ways we can choose n individuals from the population of size N (no repetition, order doesn't matter). Consider all Siena students as a population. There are 3,174 enrolled students. The table below shows the number of unique samples which can be drawn for various sample sizes:

Sample Size	2	3	4	10
Number of Unique Samples	5,035,551	5,324,255,924	2.67599e+15	2.819309e+28

The sampling distribution of the sample means is itself a very large population. When a population has a Normal distribution, the distribution of all sample means (of the same size samples) has also a Normal distribution. Furthermore, if the population is $N(\mu, \sigma)$ then the distribution of sample means is $N(\mu, \sigma/\sqrt{n})$ for the same sample size n. That is, the distribution of sample means are less variable than individual observations. Furthermore, means are more Normal (bell-shaped and symmetric) than individual observations **making the use of sample means a frequent choice for use in statistical inference**. In the next example, we will illustrate the idea of the sampling distribution.

Example: The table below shows heights (inches) of the entire population of Siena Math majors. In this example we have access to the entire population, but this is rarely the case in real life. Gathering information on an entire population is often extremely costly or impossible. Because of this, we often take a sample of the population and use that to understand the properties of the population.

65	62	62	65	70	70	65	72	74	65	69	75	70	66	74	64	66.5
62	73	73	66	66	71	68.5	61	68	68	68	70	72	67	67	71	66

The data file "SienaMath-height.csv" can be found at the course Canvas page.

(a) Find the mean and standard deviation of the height of Siena Math majors. This is the mean μ and the standard deviation σ of the population.

```
> hgt=read.csv("~/Dropbox/Teaching Siena/MATH 275 Fall19/SienaMath-height.csv")
> mean(hgt$height)
[1] 68
> sd(hgt$height)
[1] 3.763863
```

Hence, $\mu = 68$ inches and $\sigma = 3.763863$ inches.

(b) Select a random sample of size 3 from this population and compute its mean.

```
> s=sample(hgt$height,3)
> s
[1] 73 65 70
> mean(s)
[1] 69.33333
```

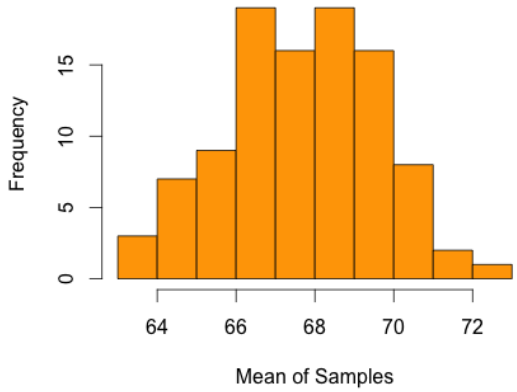
(c) Now randomly select 10 samples of size 3 from the population and compute the mean of each one of these samples. What is the mean value of these means.

```
> MeanXbar=numeric(10)
> for (i in 1:10){smp=sample(hgt$height,3);MeanXbar[i]=mean(smp)}
> MeanXbar
 [1] 69.33333 68.33333 67.66667 69.66667 66.16667 67.00000 66.50000
 [8] 67.00000 66.00000 69.16667
> mean(MeanXbar)
[1] 67.68333
```

The **R codes** can be explained as follows:
The command **MeanXbar=numeric(10)** initiates the vector variable we name MeanXbar with zero in each of its 10 entries.The command **for (i in 1:10)** returns a for loop over a list of 10 numbers. The body of the for loop, **smp=sample(hgt$height,3); MeanXbar[i]=mean(smp)**, creates a random sample of size 3 drawn from the population of $hgt\$height$, in the i^{th} iteration. The mean of this sample is calculated and stored in **MeanXbar**. This process is repeated 10 times. By typing MeanXbar we can observe the mean value of each of the 10 samples stored in the vector MeanXbar.

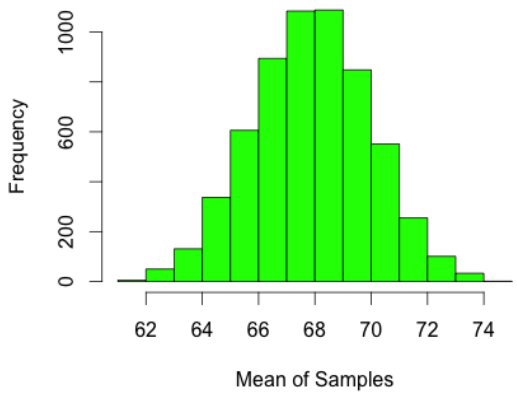
(d) Randomly select 100 samples of size 3 from the population. Construct a histogram for the mean height values of these 100 samples. Find the approximate mean and standard deviation of the sample means. Compare these two values with μ and $\frac{\sigma}{\sqrt{n}}$.

```
> MeanXbar=numeric(100)
> for (i in 1:100){smp=sample(hgt$height,3);MeanXbar[i]=mean(smp)}
> hist(MeanXbar,col="orange", xlab="Mean of Samples",main="")
> mean(MeanXbar)
[1] 67.81333
> mean(hgt$height)
[1] 68
> sd(MeanXbar)
[1] 1.94138
> sd(hgt$height)/sqrt(3)
[1] 2.173067
```



(e) Note that, there are $\binom{34}{3} = 5984$ possible unique samples. Now, select 5984 samples of size 3 from the population and repeat the previous part of the example.

```
> MeanXbar=numeric(5984)
> for (i in 1:5984){smp=sample(hgt$height,3);MeanXbar[i]=mean(smp)}
> hist(MeanXbar,col="green", xlab="Mean of Samples",main="")
> mean(MeanXbar)
[1] 68.00905
> mean(hgt$height)
[1] 68
> sd(MeanXbar)
[1] 2.081801
> sd(hgt$height)/sqrt(3)
[1] 2.173067
```



The approximate mean of sample means is 68.00905 and standard deviation of sample means is 2.081801. The true mean and standard deviation of the population are $\mu = 68$ and $\sigma = 3.763863$. Although the sample size is small ($n = 3$), the results are quite satisfactory. Large samples give a better estimation. That is called the Central Limit Theorem. If you want to see this works, repeat the example with a larger sample size (for instance $n = 13$).

Central Limit Theorem

When randomly sampling from **any population** with mean μ and standard deviation σ, when $\boldsymbol{n}$ **is large enough**, the distribution of sample means (sampling distribution) is approximately normal: $N(\mu, \sigma/\sqrt{n})$. We will depend on the Central Limit Theorem (CLT) again and again in order to do normal probability calculations when we use sample means to draw conclusions about a population mean. We now know that we can do this even if the population distribution is not normal.

How large is "large enough" ?

i. If the underlying distribution is Normal, then the sampling distribution of means will be Normal for any n.

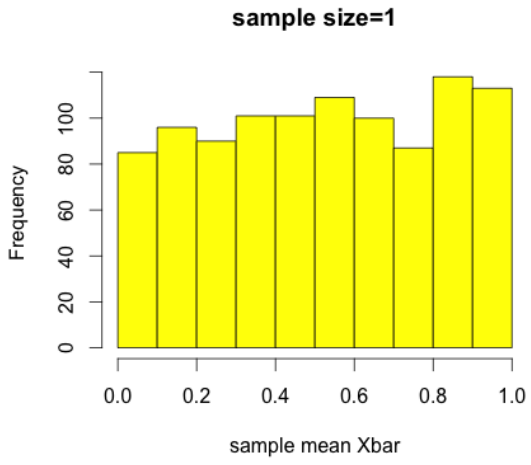
ii. If the underlying distribution is roughly symmetric and unimodel (with only one peak), the normal approximation can be good for n as small as 4.

iii. If there is no outlier and the distribution is not highly skewed, then the Normal distribution is a good approximation to the sampling distribution of means when n is greater than 30.

Example: Use R to simulate 1000 times the sampling distribution of the means of 1, 30, and 1000 observations from the $Uniform(0, 1)$ distribution. Create a histogram and determine the mean and standard deviation of these simulations.

The command **runif(n = k, min = a, max = b)** draw random samples from the $Uniform(a, b)$ distribution of size k.

```
> xbar=numeric(1000)
> for (i in 1:1000){x=runif(n=1,min=0,max=1);xbar[i]=mean(x)}
> hist(xbar,col="yellow",main="sample size=1",xlab="sample mean Xbar")
> mean(xbar)
[1] 0.5198617
> sdr(xbar)
[1] 0.2884736
```




```
> xbar2=numeric(1000)
> for (i in 1:1000){x=runif(n=30,min=0,max=1);xbar2[i]=mean(x)}
> hist(xbar2,col="red",main="sample size=30",xlab="sample mean Xbar")
> mean(xbar2)
```

```
[1] 0.5005839
> sd(xbar2)
[1] 0.05463668

> xbar3=numeric(1000)
> for (i in 1:1000){x=runif(n=1000,min=0,max=1);xbar3[i]=mean(x)}
> hist(xbar3,col="green",main="sample size=1000",xlab="sample mean Xbar")
> mean(xbar3)
[1] 0.499538
> sd(xbar3)
[1] 0.009006156
```