Chapter 5

Sampling Distributions and the Central Limit Theorem

In this chapter, we will discuss the following topics:

- We will show simulations of a sampling distribution.
- We will show simulations of the Central limit theorem to demonstrate that the distribution of the sample mean is approximately normal for large enough sample size.
- We will look at examples using the Central limit theorem.

Sampling Distributions

- Suppose we have a large population and draw all possible samples of size n from the population.
- Suppose for each sample, we compute a statistics (for example the sample mean, $\bar{x}$).
- The sampling distribution is the probability distribution of this statistics considered as a random variable.
- We measure the variability of the sampling distribution by its variance or its standard deviation.
- We denote the sample mean of a simple random sample, X_1, X_2,...,X_n, by $\overline{X}$, where

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

  The sample mean $\overline{X}$ is a random variable. We denote the value of the sample mean by $\bar{x}$.

If $\overline{X}$ is the sample mean of a simple random sample of size n drawn from a large population with mean μ and standard deviation σ, then the sampling distribution of $\overline{X}$ has mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. In our next example, we will illustrate the idea of a sampling distribution.

Problem. The table below shows the population of 10 students and their scores on an exam. (Idea from [1])

Student	1	2	3	4	5	6	7	8	9	10
Score	85	61	85	67	74	72	70	75	59	66

(a) Find the mean and standard deviation of the 10 scores in the population of students. This is the population mean μ and the standard deviation σ of the population.
(b) Select a sample of size 4 from this population and compute its mean $\bar{x}$. Do this process 10 times such that there are 10 samples from the population of size 4. Construct a histogram of the 10 values of $\bar{x}$. This is an approximation to the sampling distribution of $\overline{X}$. Find the approximate mean -and the approximate standard deviation of the sample mean $\overline{X}$. Compare these values to μ and $\frac{\sigma}{\sqrt{n}}$ respectively.

Solution to part (a). We obtain:

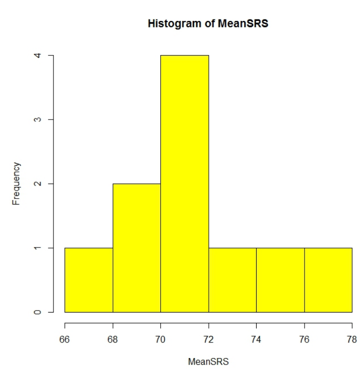
```
> Score=c(85,61,85,67,74,72,70,75,59,66)
> mean(Score)
[1] 71.4
> sd(Score)
[1] 8.834277
```

Hence, $\mu = 71.4$ and $\sigma = 8.8343$.

Solution to part (b). We obtain:

```
> MeanSRS=numeric(10)
> for (i in 1:10){SRS=sample(Score,4);MeanSRS[i]=mean(SRS)}
> hist(MeanSRS,col="yellow")
> MeanSRS
[1] 70.75 66.00 74.00 76.25 70.00 70.25 71.75 75.25 68.75 71.50
> mean(MeanSRS)
[1] 71.45
> sd(MeanSRS)
[1] 3.074989
> sd(Score)/sqrt(10)
[1] 2.793644
```

The approximate mean of $\overline{X}$ is 71.45 and the approximate standard deviation of $\overline{X}$ is 2.7936. The true mean and standard deviation of the sample mean $\overline{X}$ is $\mu = 71.4$ and $\frac{\sigma}{\sqrt{10}} = 2.7937$, respectively.



Explanation. The code can be explained as follows:

- The command **MeanSRS=numeric(10)** initiates the vector variable we name *MeanSRS* 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*, **{SRS=sample(Score,4);MeanSRS[i]=mean(SRS)}**, creates a simple random sample of size 4 drawn from the population of **Score** in the i^{th} iteration. The mean of this sample is calculated and stored in **MeanSRS**. This process is repeated 10 times.

- By typing **MeanSRS** we can observe the mean value of each of the 10 samples stored in the vector **MeanSRS**.

- The remaining steps calculate the mean and standard deviation of the data in **MeanSRS**, and draws its histogram.

The Central Limit Theorem

Let $\overline{X}$ be the mean of a simple random sample, X_1, X_2,..., X_n, of size n from a population with mean μ and standard deviation σ. The Central limit theorem says that the sampling distribution of the sample mean $\overline{X}$ is approximately Normal with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$ for large enough sample size n.
When can we use the Central limit theorem? [2]

- The Normal distribution is a good approximation to the sampling distribution of the mean when n is greater than 25 or 30.

- If the underlying distribution is continuous, symmetric and with only one peak, the Normal approximation can be good for n as small as 4 or 5.

- If the underlying distribution is approximately normal, then the distribution of $\overline{X}$ will be approximately normal for n as small as 2 or 3.

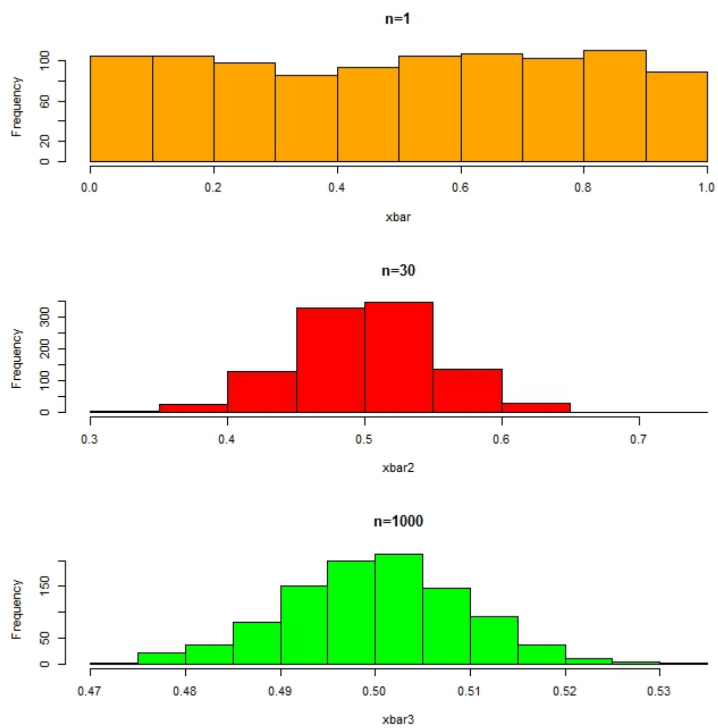
Problem. Use R to simulate 1000 times the sampling distribution of the mean, $\overline{X}$, of 1, 30, and 1000 observations from the uniform distribution. Create a histogram and determine the mean and standard deviation of these simulations.

Solution. We simulate for sample sizes 1, 30, and 1000, respectively:

```
> xbar=numeric(1000)
> for (i in 1:1000){x=runif(1);xbar[i]=mean(x)}
> hist(xbar,col="orange",main="n=1")
> mean(xbar)
[1] 0.4996082
> sd(xbar)
[1] 0.2887753

> xbar2=numeric(1000)
> for (i in 1:1000){x=runif(30);xbar2[i]=mean(x)}
> hist(xbar2,col="red",main="n=30")
> mean(xbar2)
[1] 0.5007057
> sd(xbar2)
[1] 0.05266257
```

```
> xbar3=numeric(1000)
> for (i in 1:1000){x=runif(1000);xbar3[i]=mean(x)}
> hist(xbar3,col="green",main="n=1000")
> mean(xbar3)
[1] 0.5000741
> sd(xbar3)
[1] 0.009484849
```



The uniform distribution on the interval from 0 to 1 has population mean $\mu = 0.5$ and standard deviation $\sigma = \sqrt{\frac{1}{12}}$. When $n = 1000$, the distributional standard deviation of the mean is $\frac{\sigma}{\sqrt{1000}} = \frac{1}{\sqrt{12000}} = 0.009128709$. We see that the sample mean of the simulations approaches 0.5. The standard deviation of the mean from the simulations gets smaller as n gets larger and is close to the distributional value when $n = 1000$. We can also see that as n gets larger, the distribution of $\overline{X}$ approaches the normal distribution.

Explanation. The code can be explained as follows:

- The commands **runif(1)**, **runif(30)**, and **runif(1000)** draw random samples from the uniform distribution of sizes 1, 30, and 1000, respectively. (See the previous chapter). These processes are repeated 1000 times.

Problem. Suppose a certain manufacturer produces steel shafts for which the diameter has mean 0.3 inches and standard deviation 0.05 inches.
(a) Determine the mean and standard deviation of the average diameter of 30 shafts.
(b) Use the Central limit theorem to determine the probability that the average diameter of 30 shafts is greater than 0.31.

Solution to part (a). The mean of the average diameter of 30 shafts is 0.3 and the standard deviation is $\frac{\sigma}{\sqrt{n}} = \frac{0.05}{\sqrt{30}} = 0.0091287$.

Solution to part (b). Let X_1,..., X_{30} be the diameter of the 30 shafts. We want to find $P(\overline{X} > 0.31)$, where $\overline{X}$ is approximately normally distributed with mean 0.3 and standard deviation $= 0.0091287$ by the Central limit theorem. Using R:

```
> pnorm(0.31,0.3, 0.0091287,lower.tail=FALSE)
[1] 0.1366606
```

Thus, $P(\overline{X} > 0.31) = 0.137$.

References

[1] D. S. Moore, W. I. Notz, M. A. Fligner, R. Scoot Linder. *The Basic Practice of Statistics.* W. F. Freeman and Company, New York, 2013.

[2] E. A. Tanis, R. V. Hogg. *A Brief Course in Mathematical Statistics.* Pearson Prentice Hall, Upper Saddle River, NJ, 2008.

00