5.1: Introduction to Sampling Distributions

Learning Objectives

- Define and construct probability distributions of sample statistics with simple random sampling
- Define and construct sampling distributions of sample statistics
- Define and give examples of unbiased estimators
- Explore the impact sample size has on the sampling distributions of a given population
- Represent sampling distributions as continuous random variables

Review and Preview

In the previous chapter, we began our study of random variables by briefly connecting them with inferential and sample statistics. This chapter studies sample statistics as random variables, paying close attention to probability distributions. Recall for each random variable, an underlying random experiment will be conducted. A value is assigned, measured, or computed for each possible outcome of this random experiment. These are the values the random variable is said to take on, and the probability of the values occurring is determined, forming a probability distribution. If the random variable is discrete, each value has a specific probability. If the random variable is continuous, a range of values has a probability based on the area under the probability density function. We can compute the expected value (mean), variance, and standard deviation to help describe the random variable.

Suppose we wanted to know the average height of an American. It is impossible to measure the height of everyone in the country, so we decided to measure the heights of 10,000 randomly selected Americans and found that the average height of these 10,000 people is 68.4 inches. We cannot conclude that the average height of all Americans is 68.4 inches. After all, we may have selected the 10,000 tallest people in the country, or perhaps 8,000 were well below the average height. Even if our sampling method is unbiased, there is still the possibility that we obtained a sample mean far away from the population mean by pure chance. Even if the sample mean is close to the population mean, it likely is not the same. Do we need to measure more heights? Can 10,000 people accurately represent the whole country? To understand the average height of an American, we need to determine the probability that our sample mean is not accurate. What is the probability that our sample mean is off by more than 1 inch? What is the probability that it is off by less than 0.1 inches? To answer these questions, we need to think of sampling as a random variable.

Now that we have looked at the basics of random variables and have an example in mind, we study sample statistics as random variables in depth. We are interested in learning the characteristics of a population (parameters). Studying the entire population may be impossible, too expensive, or time-consuming, so we study a sample and compute a statistic to estimate the parameter. Ideally, the sample is representative of the population; it does not misrepresent the population, and the statistic is close to the parameter. We cannot guarantee such a sample, but by choosing our sample randomly, we can ensure that any bias in the sample is due to random chance. For our initial purposes, we work within the context of simple random sampling. We must decide how large of a sample to use; we denote the sample size as n. The random experiment is defined as randomly selecting n objects from the population. The sample space of our random experiment consists of all the possible samples of size n taken from the population of size N. There are ${}_NC_n$ possible samples. Each sample is assigned a value by computing the sample statistic of interest. These possible values, along with their probabilities, form the probability distribution of the sample statistic under simple random sampling. The questions of interest are: what values can the sample statistic take on, and what are the probabilities?

Constructing Probability Distributions of Sample Statistics: Proportions

We cannot know the values and their associated probabilities without studying the entire population. However, we can build solid intuition by studying small populations, where we can exhaustively study the probability distributions. We will be able to generalize in the following sections.

Consider the family of five that has been with us throughout the text. Adam and Betsy have three children: Cathy, Damon, and Erin. This family will serve as our initial population of interest. We will study several different characteristics of this family. The first step is to decide how many family members we want to sample. Indeed, it is unnecessary to sample in this situation, but this process builds our intuition and understanding of the topic. Sampling 3 family members will be perfect for our example.

Text Exercise 5.1.1

1. Determine the number of samples and then list all possible samples of size $n = 3$ from the population of Adam (A), Betsy (B), Cathy (C), Damon (D), and Erin (E).

Answer

We are selecting 3 family members from a family of 5. The order in which we select them does not matter; therefore, there are ${}_5C_3 = 10$ possible samples.

$$\boxed{\begin{array}{ccccc} ABC & ABD & ABE & ACD & ACE \\ ADE & BCD & BCE & BDE & CDE \end{array}}$$

Notice the pattern of our labelling. Adopting such a method ensures all possible samples are listed easily.

2. Recall that Adam, Betsy, and Cathy all wear glasses. The population proportion p of family members that wear glasses is $p = \frac{3}{5} = 60\%$. For each possible sample of size 3, determine the sample proportion $\hat{p}$.

Answer

We must compute the sample proportion for each of the 10 samples from part one of this text exercise.

Table $5.1.1$: All possible samples and their sample proportions

Sample	$\hat{p}$	Sample	$\hat{p}$
ABC	$\frac{3}{3} = 1$	ADE	$\frac{1}{3}$
ABD	$\frac{2}{3}$	BCD	$\frac{2}{3}$
ABE	$\frac{2}{3}$	BCE	$\frac{2}{3}$
ACD	$\frac{2}{3}$	BDE	$\frac{1}{3}$
ACE	$\frac{2}{3}$	CDE	$\frac{1}{3}$

Note that none of the possible sample proportion values equal the population proportion. Recall from our first text exercise with this family that this is true for all possible sample sizes in this particular context. Still, it is not necessarily true for others (see the last part of the referenced exercise for a refresher).

3. We are considering all the samples of size 3 and their sample proportions and have not conducted a random sampling to produce one of these possibilities. We are developing our understanding of the sample proportion $\hat{p}$ as a random variable. There are three possible values that $\hat{p}$ takes on: $\frac{1}{3}$, $\frac{2}{3}$, and 1. Our task now is to fully understand $\hat{p}$ as a random variable and construct its probability distribution.

Answer

Since we are conducting a simple random sampling of size 3 from the family of 5, each sample is equally probable. We can determine the probabilities of our random variable $\hat{p}$ using the classical approach to probability.

Table $5.1.2$ Probability distribution of sample proportions

$\hat{p}$	$P(\hat{p})$
$\frac{1}{3}$	$\frac{3}{10}$
$\frac{2}{3}$	$\frac{6}{10} = \frac{3}{5}$
1	$\frac{1}{10}$

4. Determine the expected value, variance, and standard deviation of our random variable $\hat{p}$.

Answer

Table 5.1.3 Computation table

$\hat{p}$	$P(\hat{p})$	$\hat{p}\cdot P(\hat{p})$	$\left(\hat{p}-\mu_{\hat{p}}\right)^2\cdot P(\hat{p})$
$\frac{1}{3}$	$\frac{3}{10}$	$\frac{1}{3}\cdot\frac{3}{10}=\frac{1}{10}$	$\left(\frac{1}{3}-\frac{3}{5}\right)^2\cdot\frac{3}{10}=\frac{48}{2250}$
$\frac{2}{3}$	$\frac{6}{10}=\frac{3}{5}$	$\frac{2}{3}\cdot\frac{6}{10}=\frac{4}{10}=\frac{2}{5}$	$\left(\frac{2}{3}-\frac{3}{5}\right))^2\cdot\frac{3}{5}=\frac{3}{1125}$
1	$\frac{1}{10}$	$1\cdot\frac{1}{10}=\frac{1}{10}$	$\left(1-\frac{3}{5}\right))^2\cdot\frac{1}{10}=\frac{2}{125}$
$\mu_{\hat{p}}=\mathrm{E}(\hat{p})=\frac{1}{10}+\frac{4}{10}+\frac{1}{10}=\frac{6}{10}=\frac{3}{5}$			
$\sigma^2_{\hat{p}}=\mathrm{Var}(\hat{p})=\frac{48}{2250}+\frac{6}{2250}+\frac{36}{2250}=\frac{90}{2250}=\frac{1}{25}$			
$\sigma_{\hat{p}}=\sqrt{\mathrm{Var}(\hat{p})}=\sqrt{\frac{1}{25}}=\frac{1}{5}$			

Note that $\mu_{\hat{p}}=\frac{3}{5}=p.$ even though none of the possible sample proportions are equal to the population proportion, the expected value of the probability distribution of sample proportions is the population proportion.

When considering probability distributions of sample statistics (for example, sample proportions in our previous exercise), we use subscripts to indicate the sample statistic. This visual reminder provides a simple method of improving clarity and reducing the risk of errors. Reading symbolic expressions with meaning is an important skill to maintain and develop, especially at this point in the course.

Constructing Probability Distributions of Sample Statistics: Means

To continue our exploration, we consider additional data regarding our family of five (presented below): the number of states visited by each person. When considering the characteristic of needing eyeglasses, a qualitative variable, each member of our population either possessed the characteristic or did not. Proportions and modes would be natural sample statistics to consider. On the other hand, the number of states visited is a quantitative variable, so we have many more options to consider for sample statistics. We shall consider some in the following text exercises.

Table 5.1.4: Number of States Visited

Family Member	**Number of States Visited**
Adam	20
Betsy	30
Cathy	15
Damon	12
Erin	5

? Text Exercise 5.1.2

1. Construct the probability distribution of sample means using a sample size of 3.

Answer

For each of the 10 samples, we must compute the sample mean. We do so by filling out a table.

Table 5.1.5 All possible samples and their sample means

Sample	**Sample Data**	$\bar{x}$	**Sample**	**Sample Data**	$\bar{x}$
ABC	20, 30, 15	$\frac{20+30+15}{3}=\frac{65}{3}$	ADE	20, 12, 5	$\frac{20+12+5}{3}=\frac{37}{3}$

Sample	Sample Data	$\bar{x}$	Sample	Sample Data	$\bar{x}$
ABD	20, 30, 12	$\frac{20+30+12}{3}=\frac{62}{3}$	BCD	30, 15, 12	$\frac{30+15+12}{3}=\frac{57}{3}=1$
ABE	20, 30, 5	$\frac{20+30+5}{3}=\frac{55}{3}$	BCE	30, 15, 5	$\frac{30+15+5}{3}=\frac{50}{3}$
ACD	20, 15, 12	$\frac{20+15+12}{3}=\frac{47}{3}$	BDE	30, 12, 5	$\frac{30+12+5}{3}=\frac{47}{3}$
ACE	20, 15, 5	$\frac{20+15+5}{3}=\frac{40}{3}$	CDE	15, 12, 5	$\frac{15+12+5}{3}=\frac{32}{3}$

Table $5.1.6$: Probability distribution of sample means

$\bar{x}$	$P(\bar{x})$
$\frac{32}{3}$	$\frac{1}{10}$
$\frac{37}{3}$	$\frac{1}{10}$
$\frac{40}{3}$	$\frac{1}{10}$
$\frac{47}{3}$	$\frac{2}{10}=\frac{1}{5}$
$\frac{50}{3}$	$\frac{1}{10}$
$\frac{55}{3}$	$\frac{1}{10}$
$\frac{57}{3}$	$\frac{1}{10}$
$\frac{62}{3}$	$\frac{1}{10}$
$\frac{65}{3}$	$\frac{1}{10}$

2. Determine the expected value, variance, and standard deviation of our random variable $\bar{x}$. Compare the expected value of the probability distribution $\mu_{\bar{x}}$ with population mean μ.

Answer

Table $5.1.7$: Table of computations

$\bar{x}$	$P(\bar{x})$	$\bar{x}\cdot P(\bar{x})$	$(\bar{x}-\mu_{\bar{x}})^2\cdot P(\bar{x})$
$\frac{32}{3}$	$\frac{1}{10}$	$\frac{32}{3}\cdot\frac{1}{10}=\frac{32}{30}$	$\left(\frac{32}{3}-\frac{82}{5}\right)^2\cdot\frac{1}{10}=\frac{7396}{2250}$
$\frac{37}{3}$	$\frac{1}{10}$	$\frac{37}{3}\cdot\frac{1}{10}=\frac{37}{30}$	$\left(\frac{37}{3}-\frac{82}{5}\right)^2\cdot\frac{1}{10}=\frac{3721}{2250}$
$\frac{40}{3}$	$\frac{1}{10}$	$\frac{40}{3}\cdot\frac{1}{10}=\frac{40}{30}$	$\left(\frac{40}{3}-\frac{82}{5}\right)^2\cdot\frac{1}{10}=\frac{2116}{2250}$
$\frac{47}{3}$	$\frac{2}{10}=\frac{1}{5}$	$\frac{47}{3}\cdot\frac{2}{10}=\frac{94}{30}$	$\left(\frac{47}{3}-\frac{82}{5}\right)^2\cdot\frac{2}{10}=\frac{242}{2250}$
$\frac{50}{3}$	$\frac{1}{10}$	$\frac{50}{3}\cdot\frac{1}{10}=\frac{50}{30}$	$\left(\frac{50}{3}-\frac{82}{5}\right)^2\cdot\frac{1}{10}=\frac{16}{2250}$

$\bar{x}$	$P(\bar{x})$	$\bar{x} \cdot P(\bar{x})$	$(\bar{x} - \mu_{\bar{x}})^2 \cdot P(\bar{x})$
$\frac{55}{3}$	$\frac{1}{10}$	$\frac{55}{3} \cdot \frac{1}{10} = \frac{55}{30}$	$\left(\frac{55}{3} - \frac{82}{5}\right)^2 \cdot \frac{1}{10} = \frac{841}{2250}$
$\frac{57}{3}$	$\frac{1}{10}$	$\frac{57}{3} \cdot \frac{1}{10} = \frac{57}{30}$	$\left(\frac{57}{3} - \frac{82}{5}\right)^2 \cdot \frac{1}{10} = \frac{1521}{2250}$
$\frac{62}{3}$	$\frac{1}{10}$	$\frac{62}{3} \cdot \frac{1}{10} = \frac{62}{30}$	$\left(\frac{62}{3} - \frac{82}{5}\right)^2 \cdot \frac{1}{10} = \frac{4096}{2250}$
$\frac{65}{3}$	$\frac{1}{10}$	$\frac{65}{3} \cdot \frac{1}{10} = \frac{65}{30}$	$\left(\frac{65}{3} - \frac{82}{5}\right)^2 \cdot \frac{1}{10} = \frac{6241}{2250}$
$\mu_{\bar{x}} = \mathrm{E}(\bar{x}) = \frac{32}{30} + \frac{37}{30} + \ldots + \frac{65}{30} = \frac{492}{30} = \frac{82}{5} = 16.4$			
$\sigma^2_{\bar{x}} = \mathrm{Var}(\bar{x}) = \frac{7396}{2250} + \frac{3721}{2250} + \ldots + \frac{6241}{2250} = \frac{26,190}{2250} = \frac{291}{25} = 11.64$			
$\sigma_{\bar{x}} = \sqrt{\mathrm{Var}(\bar{x})} = \sqrt{\frac{291}{25}} \approx 3.4117$			

To compute the population mean μ, we have $\frac{20+30+15+12+5}{5} = \frac{82}{5} = 16.4$ meaning $\mu_{\bar{x}} = \mu$. The average of all the sample means is the same as the population mean.

Constructing Probability Distributions of Sample Statistics: Range

Text Exercise 5.1.3

1. Construct the probability distribution of sample ranges using a sample size of 3.

Answer

For each of the 10 samples, we must compute the sample range range. We again do so by filling out a table.

Table 5.1.8 All possible samples and their sample ranges

Sample	Sample Data	Sample Range (range)	Sample	Sample Data	Sample Range (range)
ABC	20, 30, 15	15	ADE	20, 12, 5	15
ABD	20, 30, 12	18	BCD	30, 15, 12	18
ABE	20, 30, 5	25	BCE	30, 15, 5	25
ACD	20, 15, 12	8	BDE	30, 12, 5	25
\(ACE)	20, 15, 5	15	CDE	15, 12, 5	10

Table 5.1.9 Probability distribution of sample ranges

range	$P(\text{range})$
8	$\frac{1}{10}$
10	$\frac{1}{10}$
15	$\frac{3}{10}$
18	$\frac{2}{10} = \frac{1}{5}$

range	$P(\text{range})$
25	$\frac{3}{10}$

2. Determine our random variable's expected value, variance, and standard deviation range. Compare the expected value of the probability distribution $\mu_{\bar{x}}$ with the population range.

Answer

Table 5.1.10 Table of computations

range	$P(\text{range})$	$\text{range} \cdot P(\text{range})$	$(\text{range} - \mu_{\text{range}})^2 \cdot P(\text{range})$
8	$\frac{1}{10}$	$8 \cdot \frac{1}{10} = \frac{8}{10} = \frac{4}{5}$	$\left(8 - \frac{87}{5}\right)^2 \cdot \frac{1}{10} = \frac{2209}{250}$
10	$\frac{1}{10}$	$10 \cdot \frac{1}{10} = \frac{10}{10} = 1$	$\left(10 - \frac{87}{5}\right)^2 \cdot \frac{1}{10} = \frac{1369}{250}$
15	$\frac{3}{10}$	$15 \cdot \frac{3}{10} = \frac{45}{10} = \frac{9}{2}$	$\left(15 - \frac{87}{5}\right)^2 \cdot \frac{3}{10} = \frac{432}{250}$
18	$\frac{2}{10} = \frac{1}{5}$	$18 \cdot \frac{2}{10} = \frac{36}{10} = \frac{18}{5}$	$\left(18 - \frac{87}{5}\right)^2 \cdot \frac{2}{10} = \frac{18}{250}$
25	$\frac{3}{10}$	$25 \cdot \frac{3}{10} = \frac{75}{10} = \frac{15}{2}$	$\left(25 - \frac{87}{5}\right)^2 \cdot \frac{3}{10} = \frac{4332}{250}$
$\mu_{\text{range}} = \text{E}(\text{range}) = \frac{4}{5} + 1 + \ldots + \frac{15}{2} = \frac{174}{10} = \frac{87}{5} = 17.4$			
$\sigma^2_{\text{range}} = \text{Var}(\text{range}) = \frac{2209}{250} + \frac{1369}{250} + \ldots + \frac{4332}{250} = \frac{8360}{250} = \frac{836}{25} = 33.44$			
$\sigma_{\text{range}} = \sqrt{\text{Var}(\text{range})} = \sqrt{\frac{207}{10}} \approx 5.7827$			

To compute the population range, we have $30 - 5 = 25$ meaning $\mu_{\text{range}} \neq \text{Population Range}$. The average of the sample ranges is not the same as the population range. We might have expected this. A sample range could equal the population range if it includes the largest and smallest values. However, for most samples, the sample range will be smaller than the population range because the largest and smallest data values will not be included in the sample. Moreover, the sample range can never exceed the population range. Therefore, the average of the sample ranges is an average of numbers that are never larger than the population range, making the average smaller. This reasoning implies that for any sufficiently varied population, the average of the sample ranges will be less than the population range. Contrast this with the concept of a sample mean.

Probability Distributions of Sample Statistics and Sample Size

In each of our previous examples, we used the sample size $n = 3$. We could have sampled a single person $(n = 1)$ or any number less than 5. If n were to be 5 in our context, $n = N$, we would be studying the entire population, not sampling from it, and then computing the parameter rather than estimating it. We now explore what happens to the probability distributions as we change the sample size while remaining in the same population. Again, we will consider the proportion of family members who wear glasses. The probability distribution of sample proportions for $n = 1, 2, 3,$ and 4 and the expected value, variance, and standard deviation are reported in the table below.

Table 5.1.11: Probability Distribution of Sample Proportions for $n = 1, 2, 3,$ and 4

	$n = 1$	$n = 2$	$n = 3$	$n = 4$
$\mu_{\hat{p}}$	$\frac{3}{5}$	$\frac{3}{5}$	$\frac{3}{5}$	$\frac{3}{5}$
$\sigma^2_{\hat{p}}$	0.24	0.09	0.04	0.015

	$n=1$	$n=2$	$n=3$	$n=4$
$\sigma_{\hat{p}}$	≈ 0.4899	0.3	0.2	≈ 0.1225

Students struggling to understand these probability distributions and their construction are encouraged to verify the results above by first building each probability distribution and then computing each measure . Click to check your work.

Probability distribution of sample proportions $n=1$

Table 5.1.12 Probability Distribution of Sample Proportions $n=1$ (5 Samples)

$n=1$ **(5 Samples)**	
$\hat{p}$	$P(\hat{p})$
0	$\frac{2}{5}$
1	$\frac{3}{5}$
$\mu_{\hat{p}}=\frac{3}{5}$	
$\sigma^2_{\hat{p}}=0.24$	
$\sigma_{\hat{p}}\approx 0.4899$	

Probability distribution of sample proportions $n=2$

Table 5.1.13 Probability Distribution of Sample Proportions $n=2$ (10 Samples)

$n=2$ **(10 Samples)**	
$\hat{p}$	$P(\hat{p})$
0	$\frac{1}{10}$
$\frac{1}{2}$	$\frac{6}{10}$
1	$\frac{3}{10}$
$\mu_{\hat{p}}=\frac{3}{5}$	
$\sigma^2_{\hat{p}}=0.09$	
$\sigma_{\hat{p}}=0.3$	

Probability distribution of sample proportions $n=3$

Table 5.1.14 Probability Distribution of Sample Proportions $n=3$ (10 Samples)

$n=3$ **(10 Samples)**	
$\hat{p}$	$P(\hat{p})$
$\frac{1}{3}$	$\frac{3}{10}$
$\frac{2}{3}$	$\frac{6}{10}=\frac{3}{5}$
1	$\frac{1}{10}$

$n = 3$ (10 Samples)	
$\hat{p}$	$P(\hat{p})$
$\mu_{\hat{p}} = \frac{3}{5}$	
$\sigma^2_{\hat{p}} = 0.04$	
$\sigma_{\hat{p}} = 0.2$	

Probability distribution of sample proportions $n = 4$

Table 5.1.15: Probability Distribution of Sample Proportions $n = 4$ (5 Samples)

$n = 4$ (5 Samples)	
$\hat{p}$	$P(\hat{p})$
$\frac{1}{2}$	$\frac{6}{10}$
$\frac{3}{4}$	$\frac{4}{10}$
$\mu_{\hat{p}} = \frac{3}{5}$	
$\sigma^2_{\hat{p}} = 0.015$	
$\sigma_{\hat{p}} \approx 0.1225$	

The expected value of each probability distribution of sample proportions is the same as the population proportion, regardless of the sample size; however, the variance and standard deviation values change with the sample size. The variance decreases as n increases, indicating that as n increases, the probability distribution is packed more closely around the population proportion. The spread decreases while remaining centered on the population proportion. Let us construct the probability distributions of sample means and ranges to see if similar patterns emerge.

Table 5.1.16: Probability Distribution of Sample Means $n = 1, 2, 3$, and 4

	$n = 1$	$n = 2$	$n = 3$	$n = 4$
$\mu_{\bar{x}}$	$\frac{3}{5}$	$\frac{3}{5}$	$\frac{3}{5}$	$\frac{3}{5}$
$\sigma^2_{\bar{x}}$	0.24	0.09	0.04	0.015
$\sigma_{\bar{x}}$	≈ 0.4899	0.3	0.2	≈ 0.1225

Again, we encourage students struggling to understand these probability distributions and their construction to verify the results above by first building each probability distribution and then computing each measure. Click to check your work.

Probability distribution of sample means $n = 1$

Table 5.1.17: Probability Distribution of Sample Means $n = 1$ (5 Samples)

$n = 1$ (5 Samples)	
$\bar{x}$	$P(\bar{x})$
5	$\frac{1}{5}$
12	$\frac{1}{5}$

$n=1$ (5 **Samples**)	
$\bar{x}$	$P(\bar{x})$
15	$\frac{1}{5}$
20	$\frac{1}{5}$
30	$\frac{1}{5}$
$\mu_{\bar{x}}=16.4$	
$\sigma_{\bar{x}}^{2}=69.84$	
$\sigma_{\bar{x}} \approx 8.3570$	

Probability distribution of sample means $n=2$

Table 5.1.18 Probability Distribution of Sample Means $n=2$ (10 Samples)

$n=2$ (10 **Samples**)	
$\bar{x}$	$P(\bar{x})$
8.5	$\frac{1}{10}$
10	$\frac{1}{10}$
12.5	$\frac{1}{10}$
13.5	$\frac{1}{10}$
16	$\frac{1}{10}$
17.5	$\frac{2}{10}$,
21	$\frac{1}{10}$
22.5	$\frac{1}{10}$
25	$\frac{1}{10}$
$\mu_{\bar{x}}=16.4$	
$\sigma_{\bar{x}}^{2}=26.19$	
$\sigma_{\bar{x}} \approx 5.1176$	

Probability distribution of sample means $n=3$

Table 5.1.19 Probability Distribution of Sample Means $n=3$)(\\(10 Samples)

$n=3$ (10 **Samples**)	
$\bar{x}$	$P(\bar{x})$
$10.\bar{6}$	$\frac{1}{10}$
$12.\bar{3}$	$\frac{1}{10}$
$13.\bar{3}$	$\frac{1}{10}$

$n = 3$ (10 **Samples)**	
$\bar{x}$	$P(\bar{x})$
$15.\bar{6}$	$\frac{2}{10}$
$16.\bar{6}$	$\frac{1}{10}$
$18.\bar{3}$	$\frac{1}{10}$
19	$\frac{1}{10}$
$20.\bar{6}$	$\frac{1}{10}$
$21.\bar{6}$	$\frac{1}{10}$
$\mu_{\bar{x}} = 16.4$	
$\sigma^2_{\bar{x}} = 11.64$	
$\sigma_{\bar{x}} \approx 3.4117$	

Probability distribution of sample means $n = 4$

Table 5.1.20 Probability Distribution of Sample Means $n = 4$ (5 Samples)

$n = 4$ (5 **Samples)**	
$\bar{x}$	$P(\bar{x})$
13	$\frac{1}{5}$
13.5	$\frac{1}{5}$
16.75	$\frac{1}{5}$
17.5	$\frac{1}{5}$
19.25	$\frac{1}{5}$
$\mu_{\bar{x}} = 16.4$	
$\sigma^2_{\bar{x}} = 4.365$	
$\sigma_{\bar{x}} \approx 2.0893$	

We notice a similar trend with the probability distributions of sample means. Regardless of our sample size n, the expected value is the population mean. As n increases, the spread of our distributions decreases. We now look at a final example: sample ranges. Range, indeed all measures of spread, are not very informative (nor well-defined) if there is only one observation. Think about why that is. We will only consider samples with at least 2 observations.

Table 5.1.21: Probability Distribution of Sample Ranges $n = 1, 2, 3$, and 4

	$n = 2$	$n = 3$	$n = 4$
μ_{range}	11.6	17.4	21.6
σ^2_{range}	40.04	33.44	18.24
σ_{range}	≈ 6.3277	≈ 5.7827	≈ 4.2708

Click to verify probability distribution construction as needed.

Probability distribution of sample ranges $n = 2$

Table $5.1.22$ Probability Distribution of Sample Ranges $n = 2$ $(10$ Samples)

$n = 2$ (10 **Samples**)	
range	$P(\text{range})$
3	$\frac{1}{10}$
5	$\frac{1}{10}$
7	$\frac{1}{10}$
8	$\frac{1}{10}$
10	$\frac{2}{10}$
15	$\frac{2}{10}$
18	$\frac{1}{10}$
25	$\frac{1}{10}$
$\mu_{\text{range}} = 11.6$	
$\sigma^2_{\text{range}} = 40.04$	
$\sigma_{\text{range}} \approx 6.3277$	

Probability distribution of sample ranges $n = 3$

Table $5.1.23$ Probability Distribution of Sample Ranges $n = 3$ $(10$ Samples)

$n = 3$ (10 **Samples**)	
range	$P(\text{range})$
8	$\frac{1}{10}$
10	$\frac{1}{10}$
15	$\frac{1}{10}$
18	$\frac{2}{10}$
25	$\frac{1}{10}$
$\mu_{\text{range}} = 17.4$	
$\sigma^2_{\text{range}} = 33.44$	
$\sigma_{\text{range}} \approx 5.7827$	

Probability distribution of sample ranges $n = 4$

Table $5.1.24$ Probability Distribution of Sample Ranges $n = 4$ (5 Samples)

$n = 4$ (5 **Samples**)	
range	$P(\text{range})$

$n = 4$ (5 **Samples**)	
range	$P(\text{range})$
15	$\frac{1}{5}$
18	$\frac{1}{5}$
25	$\frac{3}{5}$
$\mu_{\text{range}} = 21.6$	
$\sigma^2_{\text{range}} = 18.24$	
$\sigma_{\text{range}} \approx 4.2708$	

Having seen that the expected value of the probability distribution of sample ranges was not the population range in the case of $n = 3$, it is not surprising to see a similar situation in the other two sample sizes. The expected value of the probability distribution equaling the population parameter is not typical among the various statistics we consider. However, as n increases, we see a decrease in the spread of our probability distributions.

Sampling Distributions of Sample Statistics

Let us review what we know so far. We have constructed probability distributions of sample statistics under simple random sampling by computing a particular sample statistic for every possible sample of a specific size, n, and then determining the probability that these values occur. For some sample statistics, the probability distribution is centered around the population parameter; that is, for some statistics, the expected value of the probability distribution is the associated population parameter. Finally, we noticed that the spread of the probability distribution decreases as the sample size increases for each of the statistics studied so far.

We now introduce the probability distribution that much of inferential statistics is built upon, the **sampling distribution of sample statistics**. A sampling distribution is similar in nature to the probability distributions that we have been building in this section, but with one fundamental difference: rather than sampling using simple random sampling, the sampling method is to select randomly n objects, one at a time, from the population with replacement. Note that the order of selection will matter. As such, when considering a population of size N, there are N possibilities for each random selection, indicating that there are N^n samples to consider instead of the ${}_NC_n$ as with simple random sampling.

One may question why a distribution constructed from sampling with replacement takes priority in inferential statistics when the probability distributions above seem much more intuitive and easy to construct (having less samples to consider). As it so happens, when populations are large enough compared to the sample size (we will discuss this more later), the probability distributions of sample statistics constructed from simple random sampling are approximated well using the sampling distribution of sample statistics.

What intuition we have built from our previous constructions transfer directly to sampling distributions. As the sample size increases, the spread of the sampling distribution decreases. This should appeal to our basic intuition that larger samples better represent the population than smaller samples. For certain statistics, the expected value of the sampling distribution is the population parameter. We call such sample statistics **unbiased estimators.** Most introductory statistics books examine three unbiased estimators: sample means, proportions, and variances. The sample range is a biased estimator of the population range for the same reasons discussed above.

? Text Exercise 5.1.4

1. Within the context of the family of five from above and using a sample size of 3, construct the sampling distribution of sample proportions for the proportion of family members who wear glasses. Recall that the sampling method used in producing sampling distributions is selecting a member at random n times with replacement which means that order matters and the same member could be in the sample multiple times.

Answer

We begin our solution by counting the number of samples that we need to consider. Since the population size is $N = 5$ and our sample size is $n = 3$, we have $5^3 = 125$ different possible samples. That is a lot!

AAA	BAA	CAA	DAA	EAA
AAB	BAB	CAB	DAB	EAB
AAC	BAC	CAC	DAC	EAC
AAD	BAD	CAD	DAD	EAD
AAE	BAE	CAE	DAE	EAE
ABA	BBA	CBA	DBA	EBA
ABB	BBB	CBB	DBB	EBB
ABC	BBC	CBC	DBC	EBC
ABD	BBD	CBD	DBD	EBD
ABE	BBE	CBE	DBE	EBE
ACA	BCA	CCA	DCA	ECA
ACB	BCB	CCB	DCB	ECB
ACC	BCC	CCC	DCC	ECC
ACD	BCD	CCD	DCD	ECD
ACE	BCE	CCE	DCE	ECE
ADA	BDA	CDA	DDA	EDA
ADB	BDB	CDB	DDB	EDB
ADC	BDC	CDC	DDC	EDC
ADD	BDD	CDD	DDD	EDD
ADE	BDE	CDE	DDE	EDE
AEA	BEA	CEA	DEA	EEA
AEB	BEB	CEB	DEB	EEB
AEC	BEC	CEC	DEC	EEC
AED	BED	CED	DED	EED
AEE	BEE	CEE	DEE	EEE

We leave the computation of each sample proportion to the reader and tabulate the results in the following table.

Table $5.1.25$ Sampling distribution of sample proportions

$\hat{p}$	$P(\hat{p})$
0	$\frac{8}{125}$
$\frac{1}{3}$	$\frac{36}{125}$
$\frac{2}{3}$	$\frac{54}{125}$
1	$\frac{27}{125}$

2. Compute the expected value, variance, and standard deviation of the sampling distribution of sample proportions found in the previous portion of this text exercise.

Answer

Table $5.1.26$ Table of computations

$\hat{p}$	$P(\hat{p})$	$\hat{p} \cdot P(\hat{p})$	$\left(\hat{p}-\mu_{\hat{p}}\right)^2 \cdot P(\hat{p})$
0	$\frac{8}{125}$	0	$\left(0-\frac{3}{5}\right)^2 \cdot \frac{8}{125}=\frac{72}{3125}$
$\frac{1}{3}$	$\frac{36}{125}$	$\frac{1}{3} \cdot \frac{36}{125}=\frac{12}{125}$	$\left(\frac{1}{3}-\frac{3}{5}\right)^2 \cdot \frac{36}{125}=\frac{576}{28125}=\frac{64}{312}$

$\hat{p}$	$P(\hat{p})$	$\hat{p} \cdot P(\hat{p})$	$\left(\hat{p}-\mu_{\hat{p}}\right)^{2} \cdot P(\hat{p})$
$\frac{2}{3}$	$\frac{54}{125}$	$\frac{2}{3} \cdot \frac{54}{125} = \frac{36}{125}$	$\left(\frac{2}{3}-\frac{3}{5}\right))^{2} \cdot \frac{54}{125} = \frac{54}{28125} = \frac{6}{312}$
1	$\frac{27}{125}$	$1 \cdot \frac{27}{125} = \frac{27}{125}$	$\left(1-\frac{3}{5}\right))^{2} \cdot \frac{27}{125} = \frac{108}{3125}$
$\mu_{\hat{p}} = \mathrm{E}(\hat{p}) = \frac{12}{125} + \frac{36}{125} + \frac{27}{125} = \frac{75}{125} = \frac{3}{5}$			
$\sigma_{\hat{p}}^{2} = \operatorname{Var}(\hat{p}) = \frac{72}{3125} + \frac{64}{3125} + \frac{6}{3125} + \frac{108}{3125} = \frac{250}{3125} = \frac{2}{25}$			
$\sigma_{\hat{p}} = \sqrt{\operatorname{Var}(\hat{p})} = \sqrt{\frac{2}{25}} \approx 0.2828$			

Sampling Distributions and Large Populations

The number of possible samples is quite large when we study most populations. As an example, FHSU enrolls about 3600 on-campus students in a typical semester. We might consider randomly sampling 30 students. In which case, the number of possible samples is about $4.8874 \cdot 10^{106}$. We naturally expect many possible sample statistic values with so many different possible samples. As such, we expect that sampling distributions take on large numbers of values and can be reasonably represented as continuous random variables, which can be understood using histograms and probability density functions.

With larger populations, we can no longer expect to construct all possible samples of a given size, n, from a population to develop an understanding of sampling distributions. However, that does not mean we cannot build a reasonably accurate representation. We can approximate sampling distributions by randomly sampling from all the possible samples and then constructing histograms to visualize the shape of the distribution. We build relative frequency histograms to estimate the probability distribution of the sampling distribution. This process is tedious but can be easily implemented with a computer.

? Text Exercise 5.1.5

The Online StatBook Project provides a program that operates with frequency counts rather than relative frequency counts. Since both preserve the general shape of a distribution, we can build an intuition about sampling distributions. Open the link and read the instructions page carefully. Then click on the "Begin" button at the top left part of your screen. Note: within this program, N stands for sample size, not population size.

1. Use the following settings for this program. **Parent Population:** Normal. **First Sampling Distribution:** Mean $N=5$. **Second Sampling Distribution:** None $N=5$. Click the "Animated" button and watch the animation.
   a. What do the boxes in the second distribution represent?
   b. What does the singular box in the third distribution mean?
   c. Construct a sampling distribution using 125 random samples (Reps=125), then $10,125$, and $50,125$. What is happening as we randomly sample more and more?

Answer

The five boxes in the second distribution represent the five observations randomly selected to form our single sample of size 5. The singular box in the third distribution is the sample mean of the sample in the second distribution. The distributions should look different since we are running a simulation randomly selecting samples, but they should be reasonably close to the figure provided below.

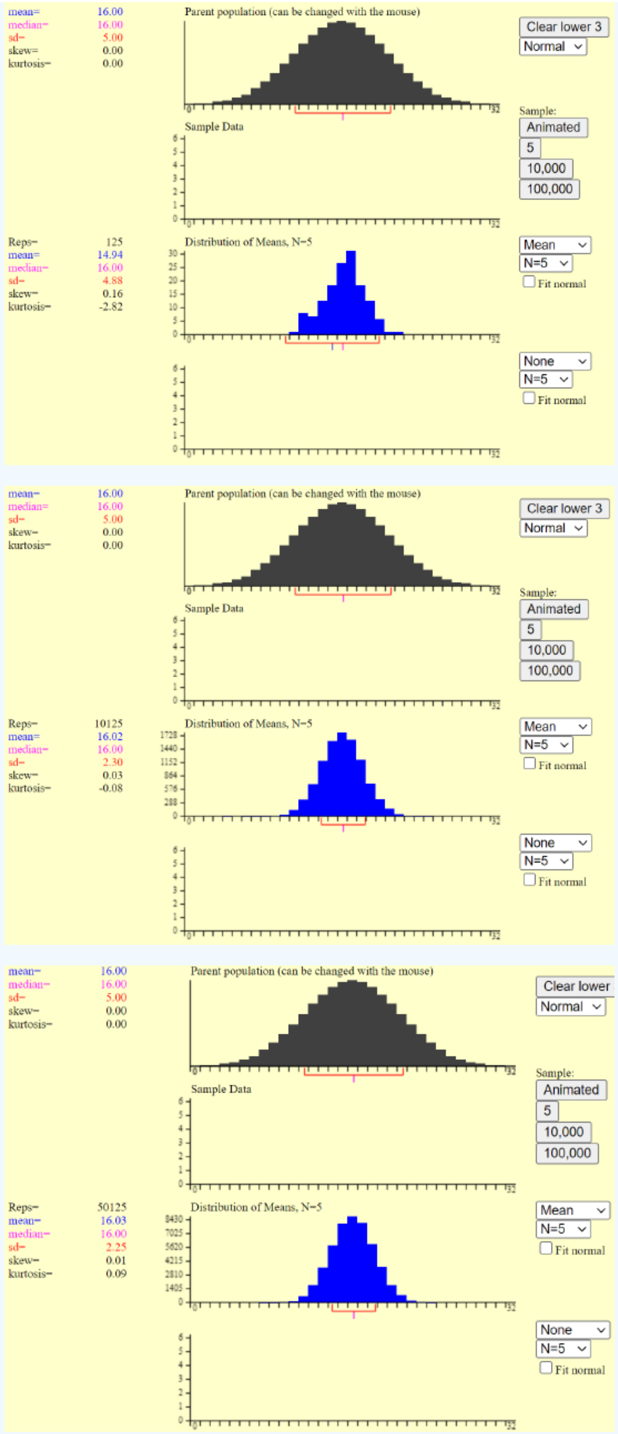

Figure $5.1.1$: Sampling distribution simulation

As we take more and more samples of size 5 from the population, we eventually converge to a consistent shape. Once the distribution does not change much, we have a decent idea of the sampling distribution.

2. Use the following settings for this program. **Parent Population:** Normal. **First Sampling Distribution:** Mean $N = 2$ Fit Normal Checked. **Second Sampling Distribution:** Mean $N = 5$ Fit Normal Checked. Run the simulation using $100,000$ random samples to estimate the sampling distributions of sample means. Compare the two sampling distributions.

Answer

Once again, our distributions will not be the same as those produced by the simulation for you, but they should be quite similar.

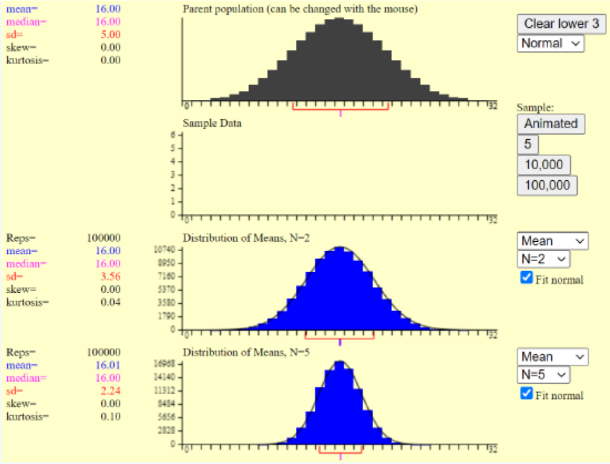

Figure 5.1.2 Sampling distribution simulation

Both sampling distributions are centered reasonably close to the population mean. When the sample size is 5, the sampling distribution is less spread out compared to the sampling distribution of sample size 2. Both sampling distributions have less spread than the parent population and fit the normal curve well.

3. Use the following settings for this program. **Parent Population:** Skewed. **First Sampling Distribution:** Mean $N = 2$ Fit Normal Checked. **Second Sampling Distribution:** Mean $N = 20$ Fit Normal Checked. Run the simulation using $100,000$ random samples to estimate the sampling distributions of sample means. Compare the two sampling distributions.

Answer

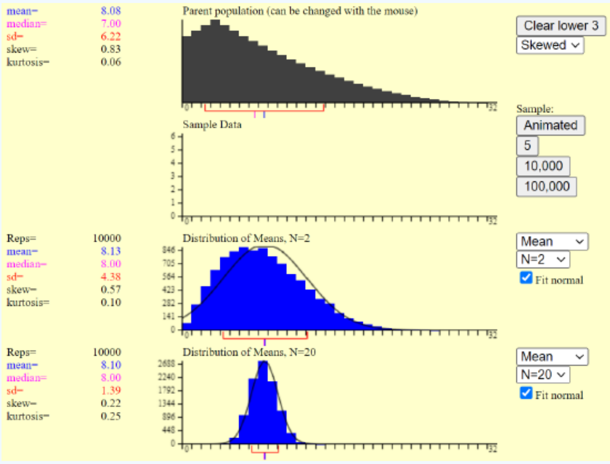

Figure 5.1.3 Sampling distribution simulation

Both sampling distributions are centered reasonably close to the population mean. The spread again decreases as the sample size increases. The first sampling distribution appears skewed to the right, just like the parent population, and is not normal. With a greater sample size, the second sampling distribution fits a normal curve much better.

4. Use the following settings for this program. **Parent Population:** Custom. **First Sampling Distribution:** Mean $N = 10$ Fit Normal Checked. **Second Sampling Distribution:** Mean $N = 25$ Fit Normal Checked. Using your mouse, construct a parent population so that when you run the simulation using $100,000$random samples to estimate the sampling distributions of sample means, it does not appear to be normal.

Answer

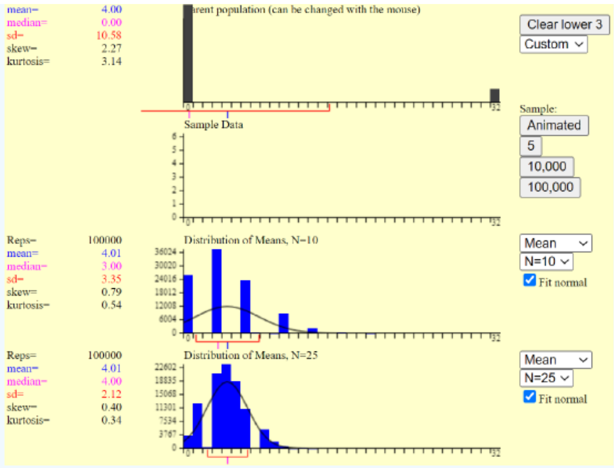

Figure 5.1.4 Sampling distribution simulation

It takes an extreme parent population for the sampling distributions not to fit a normal curve with a sample size of 25.

5. Use the following settings for this program. **Parent Population:** Skewed. **First Sampling Distribution:** Var (U) $N = 2$ Fit Normal Checked. **Second Sampling Distribution:** Var (U) $N = 25$ Fit Normal Checked. Run the simulation using $100,000$ random samples to estimate the sampling distributions of sample variances. Compare the two sampling distributions.

Answer

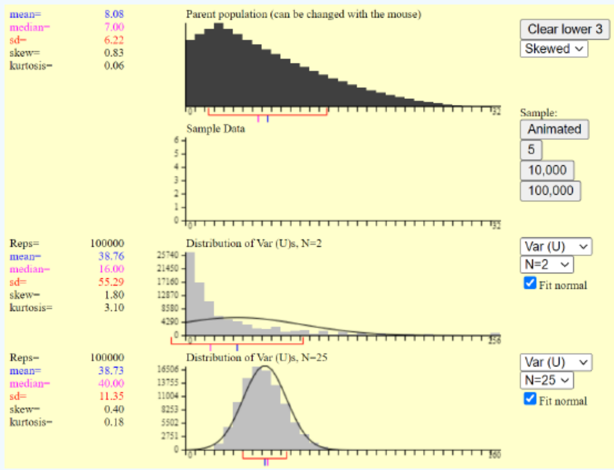

Figure 5.1.5 Sampling distribution simulation

Notice that the variance of the parent population $6.22^2 = 38.6884$is very close to the expected values (mean) of the sampling distributions. The sampling distributions are roughly centered on the population parameter. Recall that variance was the third unbiased estimator. Again, the spread of the sampling distributions decreases as the sample size increases. These sampling distributions, however, do not appear to be normally distributed. It seems closer when the sample size is 25, but it is still not a great fit.

6. In the previous parts of this text exercise, the mean of the sampling distributions has been very close to the population parameter. We do want to provide an intuition about unbiased estimators. Use the following settings for this program. **Parent Population:** Skewed. **First Sampling Distribution:** Range $N = 2, 5, 10, 16, 20, 25$ **Second Sampling Distribution:** None $N = 5$. For each sample size, run the simulation using $100,000$random samples to estimate the sampling distributions of sample ranges. Describe what happens to the expected value of the sampling distribution of sample ranges (the mean of the second distribution) as the sample size increases. How is this different from all other sampling distributions within this text exercise?

Answer

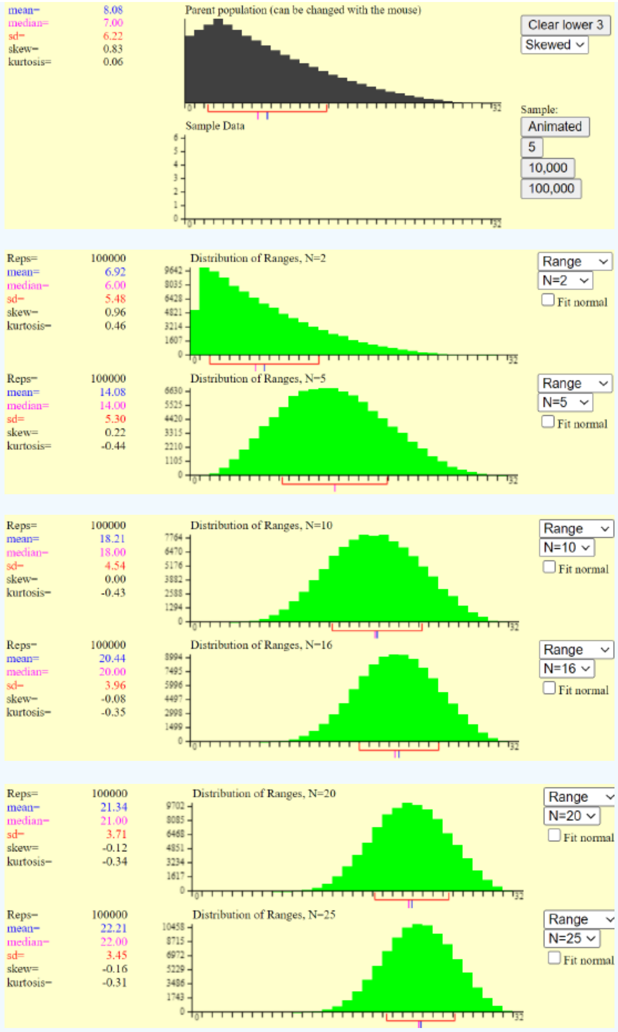

Figure $5.1.6$ Sampling distribution simulation

The population range is about 28. If we look at the expected values of the sampling distributions of sample ranges as the sample size increases, we see that the expected values increase each time. This is in contrast to what happened with the sampling distributions of sample means and variances. Regardless of the sample size, the sampling distributions of sample means and variances had expected values close to the population parameter. The distributions were centered on the population parameter. Some sample means were less than the population mean; some were more. There was no inherent bias in the estimation using sample means or sample variances. This is not the case for sample ranges. We are guaranteed that the sample range is less than or equal to the population range. Why do you think this is true? Because of this, the sampling distribution cannot be centered on the population range unless we have a trivial population, making the sample range a biased estimator.

Bridging Theory and Application

While simple random sampling and random selection with replacement are two fundamentally different approaches to sampling, when populations are large enough and the sample size is not too large relative to the population size, we consider the two methods approximately interchangable in regards to the probability distributions of sample statistics that are produced. That is, when the size conditions are met (discussed in future sections), we utilize the sampling distribution of sample statistics for a given sample size n to understand the probability of sample statistic values from simple random samples of that same size n. It is important to note that statistical analyses have been developed for cases where the size conditions are not met, but that these considerations are beyond the scope of this text. Interested readers are encouraged to continue taking more advanced statistics courses.

- **9.1: Introduction to Sampling Distributions** by David Lane is licensed Public Domain. Original source: https://onlinestatbook.com.