Unit 22: Sampling Distributions

SUMMARY OF VIDEO

If we know an entire population, then we can compute population parameters such as the population mean or standard deviation. However, we generally don't have access to data from the entire population and must base our information about a population on a sample. From samples, we compute statistics such as sample means or sample standard deviations. However, if we resample, chances are good that we won't get the same results.

This video begins with a population of heights from students in a third grade class at Monica Ros School. A graphic display of the population distribution of heights shows a roughly normal shape with a mean μ = 53.4 inches and standard deviation $\sigma = 1.8$ inches (See Figure 22.1.).

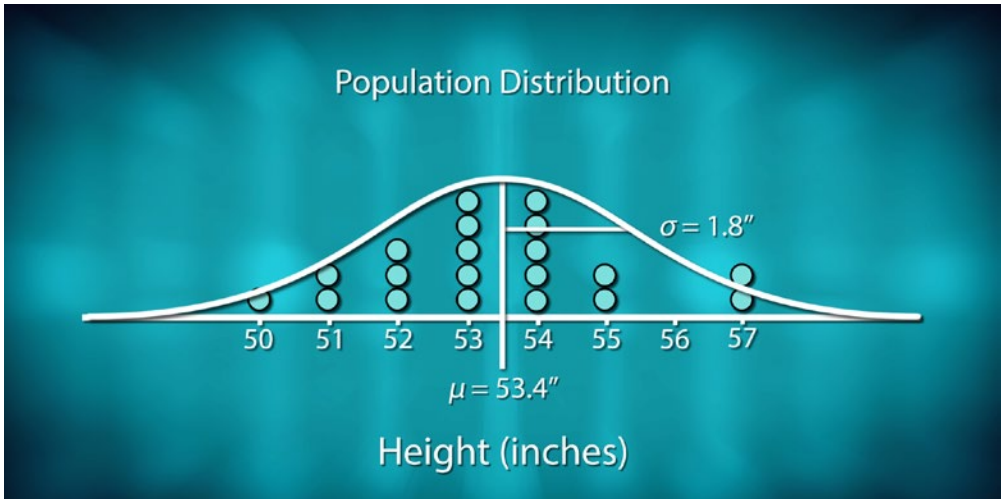


Figure 22.1. Population distribution of heights from third-grade class.

Next, we draw random samples of size four from the class and record the heights. Figure 22.2 shows the results from five samples along with their sample means, which can be found in Table 22.1. Notice that the sample means vary from sample to sample, except for Samples 3 and 4 where the sample means match even though the data values differ.

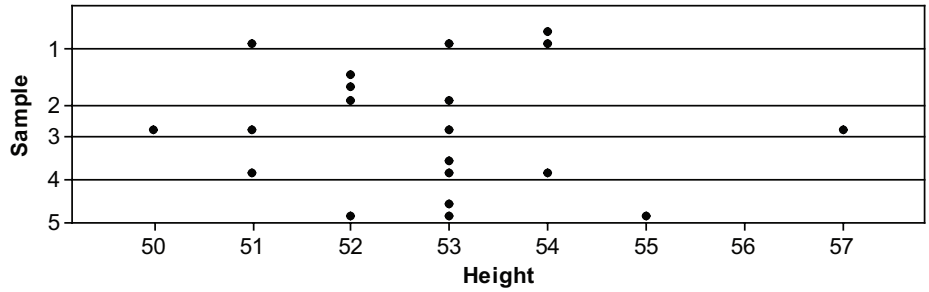
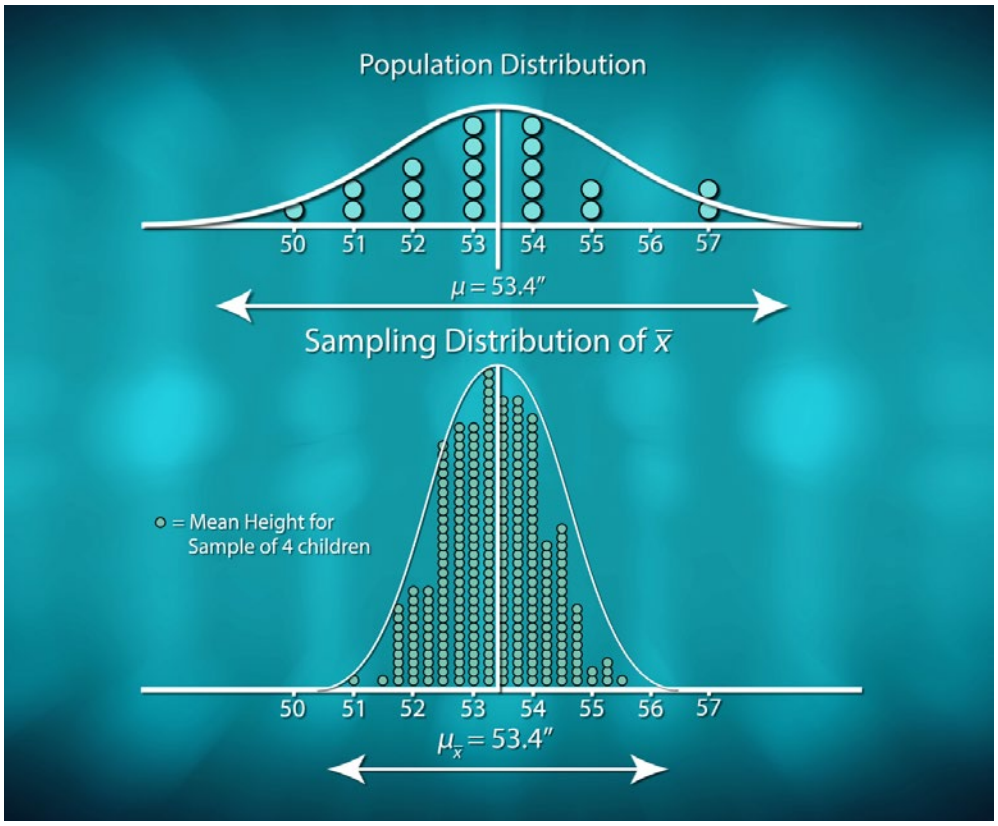


Figure 22.2. Random samples of size four.

We can keep sampling until we've selected all samples of size four from this population of 20 students. If we plot the sample means of all possible samples of size four, we get what is called the sampling distribution of the sample mean (See bottom graph in Figure 22.3.).



Sample	Mean, $\bar{x}$
1	53.00
2	52.25
3	52.75
4	52.75
5	53.25

Table 22.1. Sample means.

Figure 22.3. Sampling distribution of the sample mean.

Now, compare the sampling distribution of $\bar{x}$ to the population distribution. Notice that both distributions are approximately normal with mean 53.4 inches. However, the sampling distribution of $\bar{x}$ is not as spread out as the population distribution.

We can calculate the standard deviation of $\bar{x}$ as follows:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$\sigma_{\bar{x}} = \frac{1.8 \text{ inches}}{\sqrt{4}} \approx 0.9 \text{ inch}$$

Next, we put what we have learned about the sampling distribution of the sample mean to use in the context of manufacturing circuit boards. Although the scene depicted in the video is one that you don't see much anymore in the United States, we can still explore how statistics can be used to help control quality in manufacturing. A key part of the manufacturing process of circuit boards is when the components on the board are connected together by passing it through a bath of molten solder. After boards have passed through the soldering bath, an inspector randomly selects boards for a quality check. A score of 100 is the standard, but there is variation in the scores. The goal of the quality control process is to detect if this variation starts drifting out of the acceptable range, which would suggest that there is a problem with the soldering bath.

Based on historical data collected when the soldering process was in control, the quality scores have a normal distribution with mean 100 and standard deviation 4. The inspector's random sampling of boards consists of samples of size five. Hence, the sampling distribution of $\bar{x}$ is normal with a mean of 100 and standard deviation of $4/\sqrt{5} \approx 1.79$. The inspector uses this information to make an $\bar{x}$ control chart, a plot of the values of $\bar{x}$ against time. A normal curve showing the sampling distribution of $\bar{x}$ has been added to the side of the control chart. Recall from the 68-95-99.7% rule, that we expect 99.7% of the scores to be within three standard deviations of the mean. So, we have added control limits that are three standard deviations (3 × 1.79 or 5.37 units) on either side of the mean (See Figure 22.4.). A point outside either of the control limits is evidence that the process has become more variable, or that its mean has shifted – in other words, that it's gone out of control. As soon as an inspector sees a point such as the one outside the upper control limit in Figure 22.4, it's a signal to ask, what's gone wrong? (For more information on control charts, see Unit 23, Control Charts.)

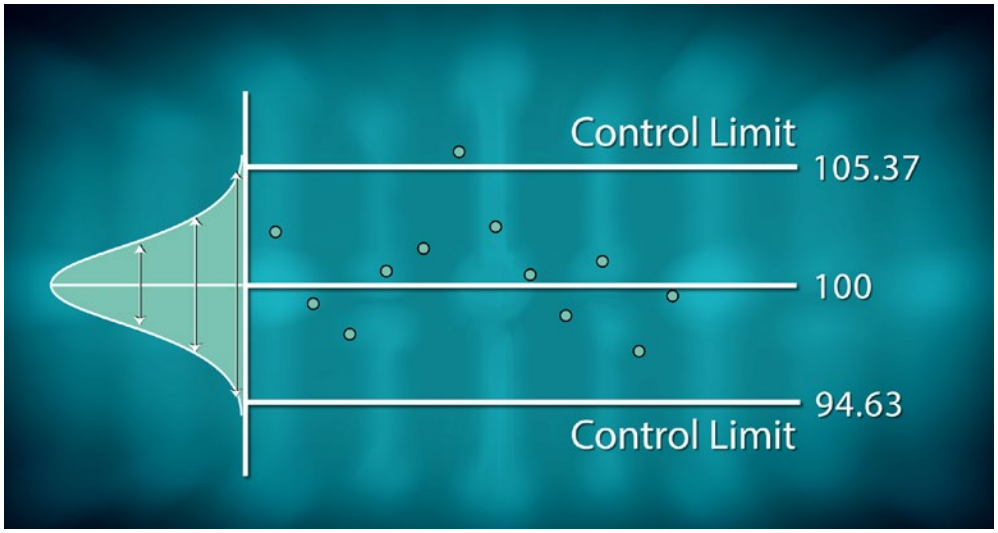


Figure 22.4. Control chart with control limits.

So far we've been looking at population distributions that follow a roughly normal curve. Next, we look at a distribution of lengths of calls coming into the Mayor's 24 Hour Hotline call center in Boston, Massachusetts. Most calls are relatively brief but a few last a very long time. The shape of the call-length distribution is skewed to the right as shown in Figure 22.5.

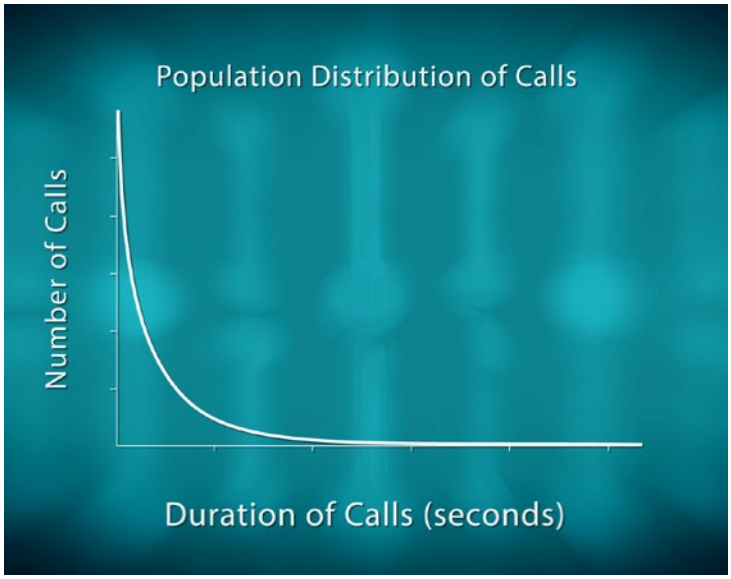


Figure 22.5. Duration of calls to a call center.

To gain insight into the sampling distribution of the sample mean, $\bar{x}$, for samples of size 10, we randomly selected 40 samples of size 10 and made a histogram of the sample means. We repeated this process for samples of size 20 and then again for samples of size 60. The histograms of the sample means appear in Figure 22.6.

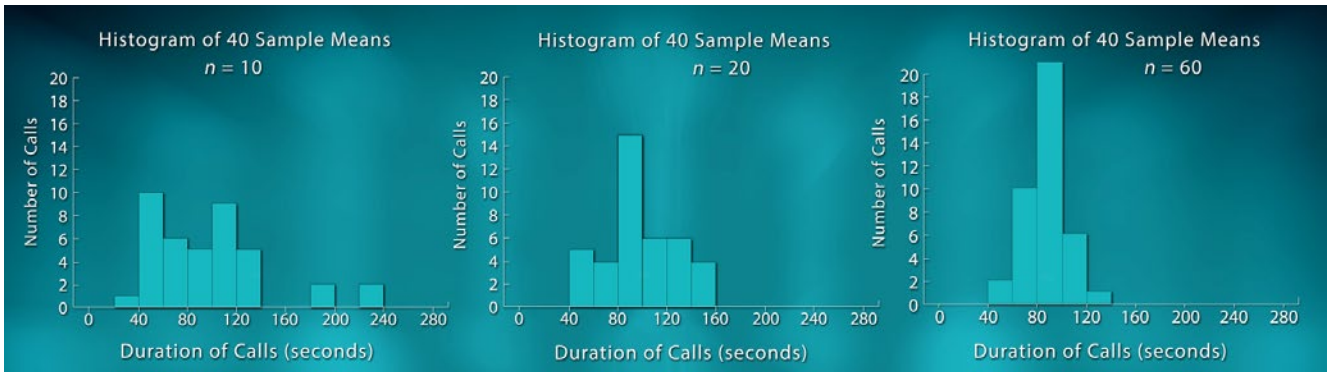


Figure 22.6. Histograms of sample means from samples of size 10, 20, and 60.

Now let's compare our sampling distributions (Figure 22.6) with the population distribution (Figure 22.5). Notice that the spread of all the sampling distributions is smaller than the spread of the population distribution. Furthermore, as the sample size n increases, the spread of the sampling distributions decreases and their shape becomes more symmetric. By the time $n = 60$, the sampling distribution appears approximately normally distributed. What we have uncovered here is one of the most powerful tools statisticians possess, called the Central Limit Theorem. This states that, regardless of the shape of the population, the sampling distribution of the sample mean will be approximately normal if the sample size is sufficiently large. It is because of the Central Limit Theorem that statisticians can generalize from sample data to the larger population. We will be seeing applications of the Central Limit Theorem in later units on confidence intervals and significance tests.

STUDENT LEARNING OBJECTIVES

A. Recognize that there is variability due to sampling. Repeated random samples from the same population will give variable results.

B. Understand the concept of a sampling distribution of a statistic such as a sample mean, sample median, or sample proportion.

C. Know that the sampling distributions of some common statistics are approximately normally distributed; in particular, the sample mean $\bar{x}$ of a simple random sample drawn from a normal population has a normal distribution.

D. Know that the standard deviation of the sampling distribution of $\bar{x}$ depends on both the standard deviation of the population from which the sample was drawn and the sample size n.

E. Grasp a key concept of statistical process control: Monitor the process rather than examine all of the products; all processes have variation; we want to distinguish the natural variation of the process from the added variation that shows that the process has been disturbed.

F. Make an $\bar{x}$ control chart. Use the 68-95-99.7% rule and the sampling distribution of $\bar{x}$ to help identify if a process is out of control.

G. Be familiar with the Central Limit Theorem: the sample mean $\bar{x}$ of a large number of observations has an approximately normal distribution even when the distribution of individual observations is not normal.

CONTENT OVERVIEW

The idea of a sampling distribution, in general, and specifically about the sampling distribution of the sample mean $\bar{x}$, underlies much of introductory statistical inference. The application to $\bar{x}$ charts is important in practice and the discussion of $\bar{x}$ charts, along with other types of control charts, continues in Unit 23, Control Charts.

If repeated random samples are chosen from the same population, the values of sample statistics such as $\bar{x}$ will vary from sample to sample. This variation follows a regular pattern in the long run; the sampling distribution is the distribution of values of the statistic in a very large number of samples. For example, suppose we start with data from the population distribution shown in Figure 22.7. This population is skewed to the right, and clearly not normally distributed.

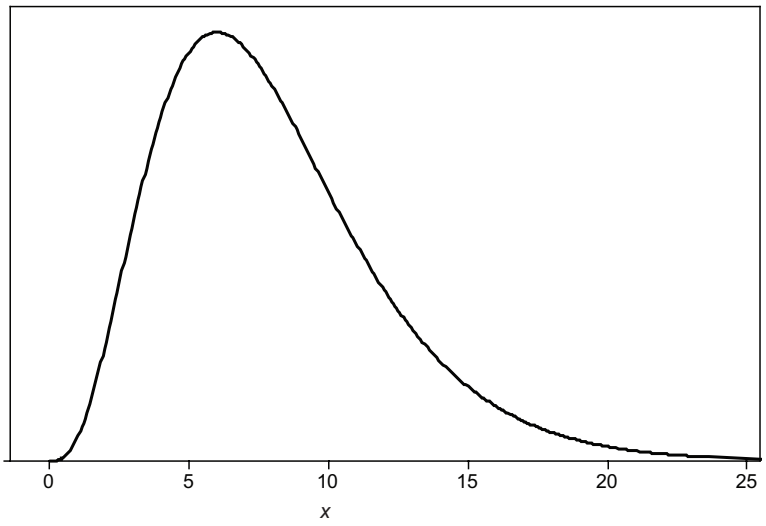


Figure 22.7. Population distribution.

Now, we draw a random sample of size 50 from this population and compute two statistics, the mean and the median, and get 20.7 and 19.8, respectively. Next we take another sample of size 50 and compute the mean and median for that sample. We keep resampling until we have a total of 1000 samples. Histograms of the 1000 means and 1000 medians from those samples appear in Figures 22.8 and 22.9, respectively. In both cases, the sampling distribution of the statistic appears approximately normally distributed. The sampling distribution of the sample mean, $\bar{x}$, is centered around 24 and the sampling distribution of the sample median at around 22.

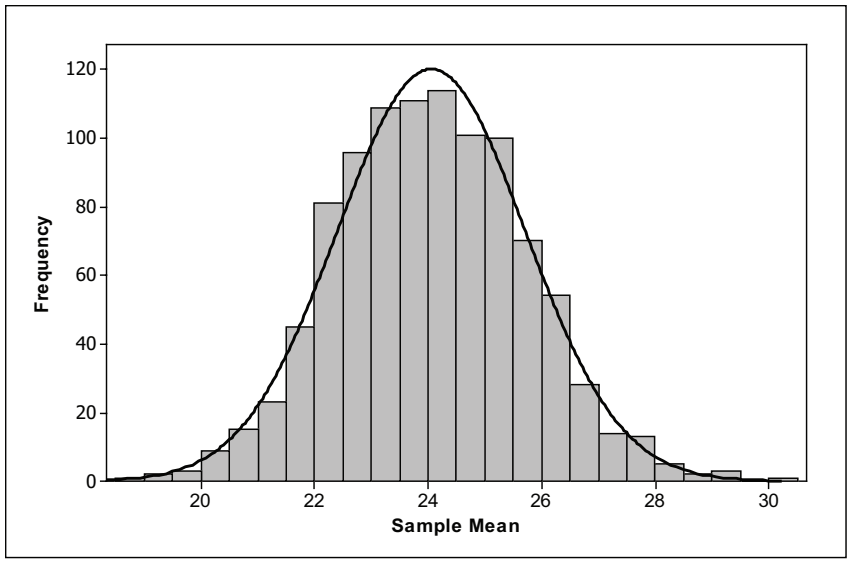


Figure 22.8. Distribution of the sample mean from 1000 samples of size 50.

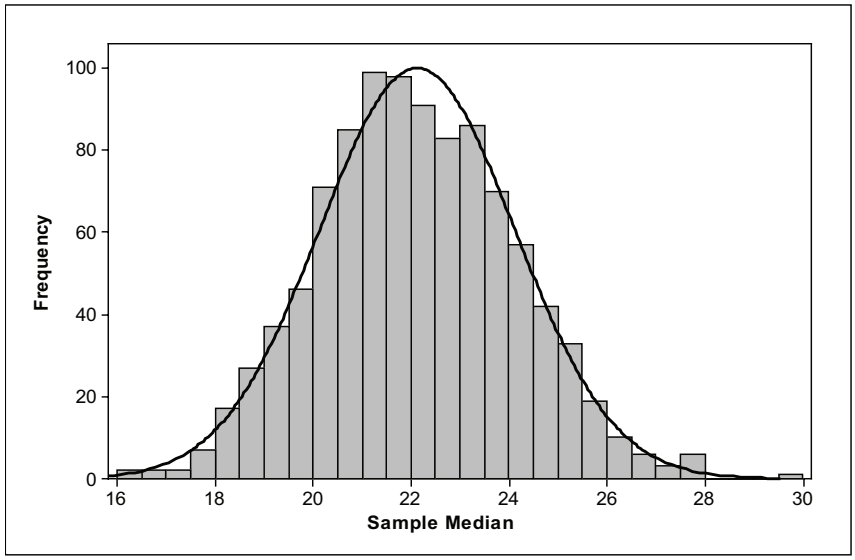


Figure 22.9. Distribution of the sample median from 1000 samples of size 50.

Although basic statistics such as the sample mean, sample median and sample standard deviation all have sampling distributions, the remainder of this unit will focus on the sampling distribution of the sample mean, $\bar{x}$. If $\bar{x}$ is the mean of a simple random sample of size *n* from a population with mean μ and standard deviation σ, then the mean and standard deviation of the sampling distribution of $\bar{x}$ are:

$$\mu_{\bar{x}} = \mu$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

If a population has the normal distribution with mean μ and standard deviation σ, then the sample mean $\bar{x}$ of *n* independent observations has a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$. In our example above, the population distribution was not normal (see Figure 22.7). In such cases, the Central Limit Theorem comes to the rescue – if the sample size is large (say $n > 30$), the sampling distribution of $\bar{x}$ is approximately normal for any population with finite standard deviation.

Control charts for the sample mean $\bar{x}$ provide an immediate application for the sampling distribution of $\bar{x}$. In the 1920's Walter Shewhart of Bell Laboratories noticed that production workers were readjusting their machines in response to every variation in the product. If the diameter of a shaft, for example, was a bit small, the machine was adjusted to cut a larger diameter. When the next shaft was a bit large, the machine was adjusted to cut smaller. Any process has some variation, so this constant adjustment did nothing except increase variation. Shewhart wanted to give workers a way to distinguish between the natural variation in the process and the extraordinary variation that shows that the process has been disturbed and hence, actually requires adjustment.

The result was the Shewhart $\bar{x}$ control chart. The basic idea is that the distribution of sample mean $\bar{x}$ is close to normal if either the sample size is large or individual measurements are normally distributed. So, almost all the $\bar{x}$-values lie within ±3 standard deviations of the mean. The correct standard deviation here is the standard deviation of $\bar{x}$, which is $\sigma/\sqrt{n}$ (where σ is the standard deviation of individual measurements). So, the control limits $\mu \pm 3\sigma/\sqrt{n}$ contain the range in which sample means can be expected to vary if the process remains stable. The control limits distinguish natural variation from excessive variation.

KEY TERMS

If repeated random samples are chosen from the same population, the values of sample statistics such as $\bar{x}$ will vary from sample to sample. This variation follows a regular pattern in the long run; the **sampling distribution** is the distribution of values of the statistic in a very large number of samples.

If $\bar{x}$ is the mean of a simple random sample (SRS) of size *n* from a population having mean μ and standard deviation σ, then the **mean and standard deviation of** $\bar{x}$ are:

$$\mu_{\bar{x}} = \mu$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

If a population has a normal distribution with mean μ and standard deviation σ, then the **sampling distribution of the sample mean**, $\bar{x}$, of *n* independent observations has a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$.

If the population is not normal but *n* is large (say $n > 30$), then the **Central Limit Theorem** tells us that the **sampling distribution of the sample mean,** $\bar{x}$, of *n* independent observations has an *approximate* normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$.

THE VIDEO

Take out a piece of paper and be ready to write down answers to these questions as you watch the video.

1. What is the difference between parameters and statistics?

2. Does statistical process control inspect all the items produced after they are finished?

3. The inspector samples five circuit boards at regular intervals and finds the mean solder quality score $\bar{x}$ for these five boards. Do we expect $\bar{x}$ to be exactly 100 if the soldering process is functioning properly?

4. If the quality of individual boards varies according to a normal distribution with mean μ = 100 and standard deviation $\sigma = 4$, what will be the distribution of the sample averages, $\bar{x}$? (Recall the sample size is n = 5.)

5. In general, is the mean of several observations more or less variable than single observations from a population? Explain.

6. The distribution of call lengths to a call center is strongly skewed. What does the Central Limit Theorem say about the distribution of the mean call length $\bar{x}$ from large samples of calls?

UNIT ACTIVITY:
SAMPLING DISTRIBUTIONS OF THE SAMPLE MEAN

Write each of these numbers	On this many slips
50	10
49, 51	9
48, 52	9
47, 53	8
46, 54	6
45, 55	5
44, 56	3
43, 57	2
42, 58	1
41, 59	1
40, 60	1

Table 22.2. Numbered slips for the population distribution.

1. Your instructor has a container filled with numbered strips as shown in Table 22.2. Make a histogram of this distribution. Describe its shape.

2. You will need 100 samples of size 9. Your instructor will provide instructions for gathering these samples. After the data have been collected, you will need a copy of the table of results before you can answer parts (a) and (b).

a. Find the sample mean for each of the samples. Record the sample means in the results table. (Save your results table. You will need this table again for the activity in Unit 24, Confidence Intervals.)

b. To get an idea of the characteristics of the sampling distribution for the sample mean, make a histogram of the sample means. (Use the same scaling on the horizontal axis that you used in question 1.) Compare the shape, center and spread of the sampling distribution to that of the original distribution (question 1).

Extension

3. A population has a uniform distribution with density curve as shown in Figure 22.10.

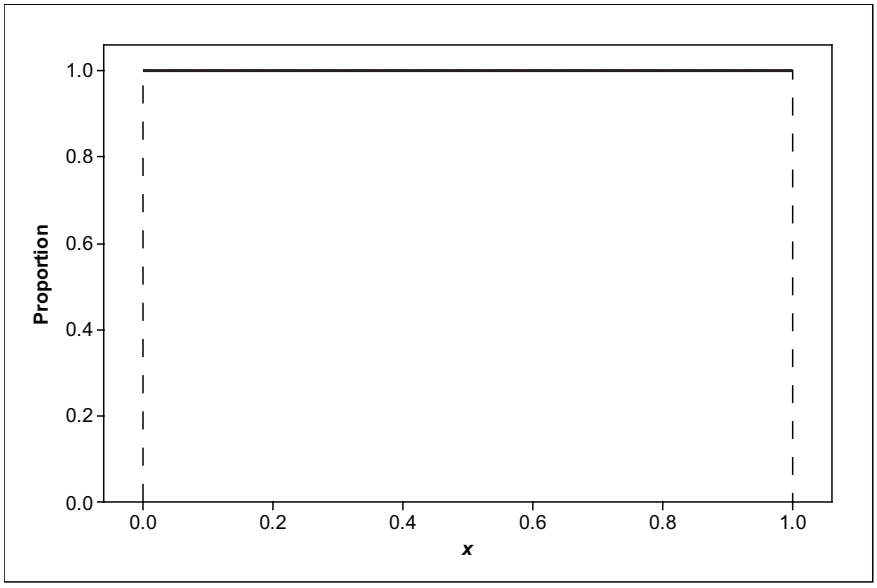


Figure 22.10. Density curve for uniform distribution.

a. Your instructor will give you directions for using technology to generate 100 samples of size 9 from this distribution.

b. Once you have your 100 samples, find the sample means.

c. Make a histogram of the 100 sample means. Describe the shape of your histogram. Compare the center of this sampling distribution with the center of the population distribution from Figure 22.10.

EXERCISES

1. The law requires coal mine operators to test the amount of dust in the atmosphere of the mine. A laboratory carries out the test by weighing filters that have been exposed to the air in the mine. The test has a standard deviation of $\sigma = 0.08$ milligram in repeated weighings of the same filter. The laboratory weighs each filter three times and reports the mean result.

a. What is the standard deviation of the reported result?

b. Why do you think the laboratory reported a result based on the mean of three weighings?

2. The scores of students on the ACT college entrance examination in a recent year had the normal distribution with mean μ = 18.6 and standard deviation $\sigma = 5.9$.

a. What fraction of all individual students who take the test have scores 21 or higher?

b. Suppose we choose 55 students at random from all who took the test nationally. What is the distribution of average scores, $\bar{x}$, in a sample of size 55? In what fraction of such samples will the average score be 21 or higher?

3. The number of accidents per week at a hazardous intersection varies with mean 2.2 and standard deviation 1.4. This number, *x*, takes only whole-number values, and so is certainly not normally distributed.

a. Let $\bar{x}$ be the mean number of accidents per week at the intersection during a year (52 weeks). What is the approximate distribution of $\bar{x}$ according to the Central Limit Theorem?

b. What is the approximate probability that, on average, there are fewer than two accidents per week over a year?

c. What is the approximate probability that there are fewer than 100 accidents at the intersection in a year? (Hint: Restate this event in terms of $\bar{x}$.)

4. A company produces a liquid that can vary in its pH levels unless the production process is carefully controlled. Quality control technicians routinely monitor the pH of the liquid. When the process is in control, the pH of the liquid varies according to a normal distribution with mean μ = 6.0 and standard deviation σ = 0.9.

a. The quality control plan calls for collecting samples of size three from batches produced each hour. Using $n = 3$, calculate the lower control limit (LCL) and upper control limit (UCL).

b. Samples collected over a 24-hour time period appear in Table 22.3.

Sample		pH level		Sample mean
1	5.8	6.2	6.0	
2	6.4	6.9	5.3	
3	5.8	5.2	5.5	
4	5.7	6.4	5.0	
5	6.5	5.7	6.7	
6	5.2	5.2	5.8	
7	5.1	5.2	5.6	
8	5.8	6.0	6.2	
9	4.9	5.7	5.6	
10	6.4	6.3	4.4	
11	6.9	5.2	6.2	
12	7.2	6.2	6.7	
13	6.9	7.4	6.1	
14	5.3	6.8	6.2	
15	6.5	6.6	4.9	
16	6.4	6.1	7.0	
17	6.5	6.7	5.4	
18	6.9	6.8	6.7	
19	6.2	7.1	4.7	
20	5.5	6.7	6.7	
21	6.6	5.2	6.8	
22	6.4	6.0	5.9	
23	6.4	4.6	6.7	
24	7.0	6.3	7.4	

Table 22.3. pH of samples.

c. Make an $\bar{x}$ chart by plotting the sample means versus the sample number. Draw horizontal reference lines at the mean and lower and upper control limits.

d. Do any of the sample means fall below the lower control limit or above the upper control limit? This is one indication that a process is “out of control.”

e. Apart from sample means falling outside the lower and upper control limits, is there any other reason why you might be suspicious that this process is either out of control or going out of control? Explain.

REVIEW QUESTIONS

1. Suppose a chemical manufacturer produces a product that is marketed in plastic bottles. The material is toxic, so the bottles must be tightly sealed. The manufacturer of the bottles must produce the bottles and caps within very tight specification limits. Suppose the caps will be acceptable to the chemical manufacturer only if their diameters are between 0.497 and 0.503 inch. When the manufacturing process for the caps is in control, cap diameter can be described by a normal distribution with $\mu = 0.500$ inch and $\sigma = 0.0015$ inch.

a. If the process is in control, what percentage of the bottle caps would have diameters outside the chemical manufacturer's specification limits?

b. The manufacturer of the bottle caps has instituted a quality control program to prevent the production of defective caps. As part of its quality control program, the manufacturer measures the diameters of a random sample of $n = 9$ bottle caps each hour and calculates the sample mean diameter. If the process is in control, what is the distribution of the sample mean $\bar{x}$? Be sure to specify both the mean and standard deviation of $\bar{x}$'s distribution.

c. The cap manufacturer has a rule that the process will be stopped and inspected any time the sample mean falls below 0.499 inch or above 0.501 inch. If the process is in control, find the proportion of times it will be stopped for inspection.

2. A study of rush-hour traffic in San Francisco records the number of people in each car entering a freeway at a suburban interchange. Suppose that this number, x, has mean 1.5 and standard deviation 0.75 in the population of all cars that enter at this interchange during rush hours.

a. Could the exact distribution of x be normal? Why or why not?

b. Traffic engineers estimate that the capacity of the interchange is 700 cars per hour. According to the Central Limit Theorem, what is the approximate distribution of the mean number of persons, $\bar{x}$, per car in 700 randomly selected cars at this interchange?

c. What is the probability that 700 cars will carry more than 1075 people? (Hint: Restate the problem in terms of the average number of people per car.)

3. Recall that the distribution of the lengths of calls coming into a Boston, Massachusetts, call center each month is strongly skewed to the right. The mean call length is μ = 90 seconds and the standard deviation is σ = 120 seconds.

a. Let $\bar{x}$ be the sample mean from 10 randomly selected calls. What is the mean and standard deviation of $\bar{x}$? What, if anything, can you say about the shape of the distribution of $\bar{x}$? Explain.

b. Let $\bar{x}$ be the sample mean from 100 randomly selected calls. What is the mean and standard deviation of $\bar{x}$? What, if anything, can you say about the shape of the distribution of $\bar{x}$? Explain.

c. In a random sample of 100 calls from the call center, what is the probability that the average length of these calls will be over 2 minutes?