Unit 22: Sampling Distributions

PREREQUISITES

This unit serves as the transition from descriptive statistics to inferential statistics. The concepts in this unit are more sophisticated than in earlier units. Students should have considerable exposure to data before viewing this unit. They need familiarity with the mean and standard deviation (Unit 4, Measures of Center, and Unit 6, Standard Deviation). Most importantly, students need background on the normal distribution (Units 7 – 9).

ADDITIONAL TOPIC COVERAGE

Additional coverage of sampling distributions can be found in *The Basic Practice of Statistics*, Chapter 11, Sampling Distributions. More in-depth information on $\bar{x}$ charts can be found in the Companion Chapters (CD insert), Chapter 17, Statistics for Quality: Control and Capability. Unit 23, Control Charts, continues the discussion of control charts. There are a number of applets that allow students to conduct simulations that give them insight into the Central Limit Theorem. For example, take a look at the Central Limit Theorem applet at www.causeweb.org/repository/statjava/

ACTIVITY DESCRIPTION

This activity supports Learning Objectives for this unit A, B, and C. In this activity, students draw random samples by hand from an approximate normal distribution. After calculating the sample means, students compare a histogram of individual observations to a histogram of the sample means.

Drawing samples by hand makes the process of random sampling more concrete to students. Since drawing 100 samples by hand is time-consuming, after several samples have been selected, students can be working on other material until it is their turn to draw one or more samples. If students have access to statistical software, samples can be generated using the

software's normal random number generator (see Extension for Minitab instructions). If you decide to use technology to create the samples, it is still a good idea to begin the activity by having students draw several samples by hand.

MATERIALS

Prepare 100 identical slips of stiff poster board.
Write the numbers as described in Table T22.1 on the slips.

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 T22.1. Distribution table for approximate normal data.

These 100 numbered slips form a population. The distribution of the numbers in this population is roughly normal with mean μ = 50 and standard deviation σ = 4. Before students begin sampling, they should make a graphic display of the population distribution (question 1).

Before students can answer question 2, they need to collect the data and the instructor needs to provide the instructions. The numbered slips should be put into a container and mixed. A student should be asked to draw a sample of size nine for Sample 1 by drawing a slip of paper, recording its number, and returning it to the container for mixing before drawing the second number, and so on until nine numbers have been drawn. The values should be recorded either in a copy of Table T22.2 or in an Excel spreadsheet. This process should be repeated as many times as convenient – around 100 times if possible. The data should be distributed to students so that they can complete the activity. **Students should save the data in Table T22.2 for use in the activity for Unit 24, Confidence Intervals.**

Extension

Students use technology (such as Excel or Minitab) to generate 100 (or more) samples of size 9 from a uniform distribution on the interval from 0 to 1. A graphic display for this population's distribution is box shaped and centered at 0.5. However, the sampling distribution of $\bar{x}$ appears roughly normal in shape but remains centered at 0.5.

Using Excel to Generate the Samples

- Label the first row of columns A – K with the following headings: Sample, X1, X2, X3, X4, X5, X6, X7, X8, X9, and Mean.
- In the column Sample, generate the numbers 1 – 100 as follows: enter 1 in cell A2. In cell A3, enter the formula = a2+1 and then press Enter. Click cell A2. Then click the bottom right corner of the cell and drag down to row 101.
- In cell B2, enter the formula =rand() and press Enter. Click cell B2 and then click the bottom right corner of the cell and drag across the row to cell J2 to form the first sample.
- With the first sample still highlighted, click the bottom right corner and drag down to form the other 99 samples.
- To find the mean of the first sample, click cell K2 (the first entry in the column labeled Mean). Enter the formula =AVERAGE(B2:J2) and press Enter.
- To calculate the means for the remaining 99 samples, click cell K2. Then click the bottom right corner and drag down to cell K101.

Using Minitab to Generate the Samples

- Label columns C1 – C11 as Sample, X1, X2, X3, X4, X5, X6, X7, X8, X9, and Mean.
- Select Calc > Make Patterned Data > Simple Set of Numbers. Then complete the dialog box as follows:

  Store patterned data in C1
  From first value 1
  To last value 100
  In steps of 1
  Click OK

You should see consecutive integers 1 – 100 under Sample.

- Generate 100 samples of size 9: Calc > Random Data > Uniform. Then complete the dialog box as follows:

  Number of rows of data to generate 100 (or more)
  Store in columns: Select C2 – C10 for X1 – X9
  Click OK

- To calculate the sample means for each sample: Calc > Row Statistics and then complete as follows:

  Select Mean for the statistic
  Select C2 – C10 for the input variables
  Store results in: C11
  Click OK.

Note: The Minitab instructions above can be adapted to replace drawing by hand from a normal population. To do so, adapt the second bullet above as follows:

- Generate 100 samples of size 9: Calc > Random Data > Normal and then complete the dialog box as follows:

  Number of rows of data to generate 100 (or more)
  Store in columns: Select C2 – C10 for X1 – X9
  Mean: 50
  Standard deviation: 4
  Click OK

Samples of Size 9, page 1

Sample	X1	X2	X3	X4	X5	X6	X7	X8	X9	Mean
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										
11										
12										
13										
14										
15										
16										
17										
18										
19										
20										
21										
22										
23										
24										
25										
26										
27										
28										
29										
30										
31										
32										
33										
34										
35										
36										
37										
38										
39										

Samples of Size 9, page 2

Sample	X1	X2	X3	X4	X5	X6	X7	X8	X9	Mean
40										
41										
42										
43										
44										
45										
46										
47										
48										
49										
50										
51										
52										
53										
54										
55										
56										
57										
58										
59										
60										
61										
62										
63										
64										
65										
66										
67										
68										
69										
70										
71										
72										
73										
74										
75										
76										
77										
78										

Samples of Size 9, page 3

Sample	X1	X2	X3	X4	X5	X6	X7	X8	X9	Mean
79										
80										
81										
82										
83										
84										
85										
86										
87										
88										
89										
90										
91										
92										
93										
94										
95										
96										
97										
98										
99										
100										

Table T22.2. Data table for samples of size 9.

Table T22.3 contains sample data that you might use if you don't want to have students draw all 100 samples. The sample solution will be based on these data.

X1	X2	X3	X4	X5	X6	X7	X8	X9	Sample Mean
40	55	45	53	41	50	54	60	56	
53	55	50	53	51	43	54	53	50	
57	48	46	50	46	48	53	51	47	
44	50	55	47	44	44	47	59	53	
48	59	51	49	51	47	47	45	55	
58	41	46	56	46	55	49	50	51	
51	46	51	40	49	47	48	60	46	
50	47	51	52	51	51	60	53	48	
59	54	49	49	49	41	56	46	56	
52	44	46	57	43	46	50	50	47	
51	45	53	54	46	48	49	46	41	
52	48	58	57	56	44	52	50	49	
47	50	53	58	47	44	48	47	48	
52	54	50	49	53	43	53	59	49	
54	44	51	43	46	52	53	58	47	
51	43	49	51	46	49	48	46	48	
51	58	48	45	51	50	59	53	55	
45	51	45	46	52	48	52	48	54	
52	56	51	52	53	53	43	47	48	
53	47	52	50	55	54	46	49	55	
50	40	57	51	52	52	48	46	49	
51	46	47	47	52	55	46	51	60	
56	41	51	52	48	44	51	48	47	
50	50	49	56	60	42	52	45	57	
48	48	54	49	53	49	51	49	47	
48	50	51	52	51	50	47	50	50	
54	53	55	46	48	48	52	52	47	
45	51	53	52	46	40	49	57	43	
49	56	53	48	50	54	44	52	50	
45	52	44	57	48	41	49	51	46	
50	53	52	49	50	53	53	52	55	
49	52	44	40	43	50	51	50	55	
47	51	50	56	47	54	55	50	46	
52	51	51	53	42	47	47	54	45	
53	49	40	49	49	49	48	52	48	
54	48	57	54	48	55	46	52	55	
50	47	55	46	49	52	45	53	46	

48	48	50	54	54	45	49	51	55	
52	49	49	46	53	48	43	55	54	
46	51	51	53	56	53	60	58	48	
56	56	52	45	46	53	51	59	54	
46	55	50	53	53	54	43	47	43	
54	50	53	48	54	46	47	41	41	
50	46	52	45	50	57	50	51	50	
55	50	58	45	49	51	47	57	46	
47	52	52	54	49	48	51	47	53	
46	49	53	54	53	47	52	55	44	
56	43	47	54	50	50	59	54	52	
41	55	51	51	52	53	48	49	55	
51	54	47	46	50	52	52	50	52	
59	47	42	44	50	44	41	45	53	
55	46	48	44	48	56	48	47	46	
52	54	43	54	43	54	51	42	50	
46	52	46	50	40	51	49	52	53	
53	50	46	48	53	47	52	52	54	
49	56	50	46	58	47	58	55	57	
46	60	45	48	49	48	51	53	54	
46	51	46	54	48	53	49	51	47	
56	47	48	52	48	49	52	55	50	
50	46	57	45	46	46	55	52	45	
52	41	52	46	51	51	48	49	40	
49	46	44	51	58	49	41	48	49	
47	56	52	43	47	50	52	50	48	
45	47	52	49	49	45	54	56	46	
53	50	45	49	51	52	49	51	51	
49	46	44	45	48	45	53	44	49	
51	54	48	47	53	49	59	46	47	
48	51	48	53	54	48	51	59	43	
44	44	46	50	58	52	57	53	56	
52	51	53	51	48	58	51	51	47	
60	40	52	43	48	52	55	43	60	
50	51	51	52	55	50	53	55	48	
52	52	54	49	46	57	48	46	55	
51	49	45	53	51	56	53	52	54	
50	46	45	47	53	50	48	49	47	
51	52	44	46	53	44	47	46	42	

49	50	49	54	52	48	53	51	50	
46	47	48	52	52	48	49	48	55	
52	46	51	50	60	50	47	56	52	
46	52	54	45	45	60	56	50	50	
51	50	54	47	47	45	56	51	54	
52	48	50	49	50	51	44	49	52	
40	48	44	48	47	49	60	47	47	
48	53	55	51	48	52	51	51	41	
49	58	53	47	58	50	53	47	52	
52	46	49	47	51	48	49	44	47	
51	50	53	50	52	52	49	54	43	
49	48	52	48	60	54	49	45	50	
59	53	49	45	46	45	50	42	51	
46	53	51	54	60	54	50	57	54	
49	44	47	48	48	54	51	50	58	
44	51	56	53	52	47	45	48	51	
48	47	44	48	48	51	52	52	53	
49	56	45	51	51	54	53	46	48	
53	52	48	47	45	53	51	48	50	
51	51	52	50	47	44	48	50	49	
56	51	55	53	52	53	49	49	48	
53	48	51	53	49	44	51	55	48	
54	53	54	54	44	48	49	49	51	
42	57	44	50	43	59	51	45	55	

Table T22.3. Sample Data

THE VIDEO SOLUTIONS

1. Parameters describe an entire population and are generally unknown. Statistics are computed from samples.

2. No, it might be too expensive to inspect them all. In addition, it is too costly to wait until the end to examine the finished product. If somewhere in the process things are out of control, it doesn't make sense to put in additional money to finish a defective product.

3. No, there will be variability in the mean scores. Sometimes the mean score will be above 100 and sometimes below 100.

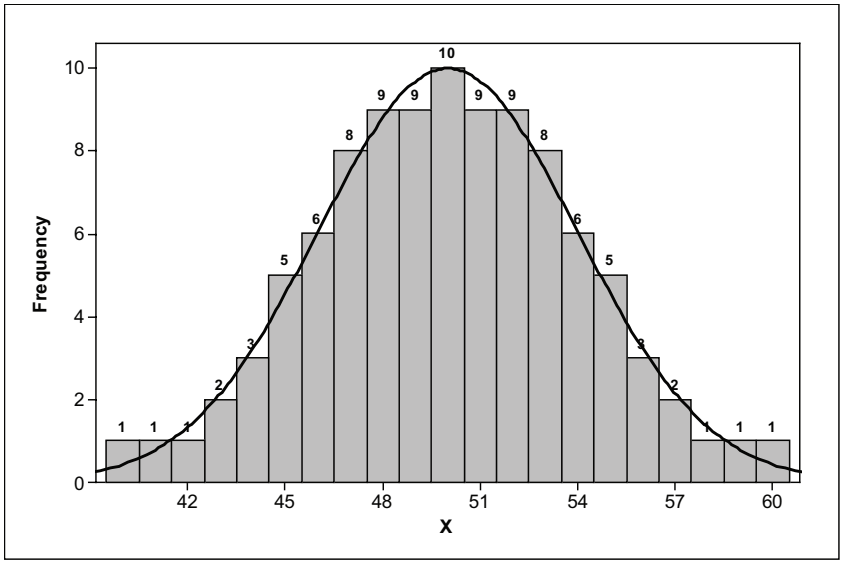
4. The distribution of $\bar{x}$ will be normal with mean 100 and standard deviation $4/\sqrt{5}$.

5. The sample mean is less variable than individual observations. That's because when averaging, high observations will balance out low observations, which makes the mean less variable.

6. The sampling distribution of the sample mean is approximately normally distributed if the sample size is sufficiently large.

ACTIVITY SOLUTIONS

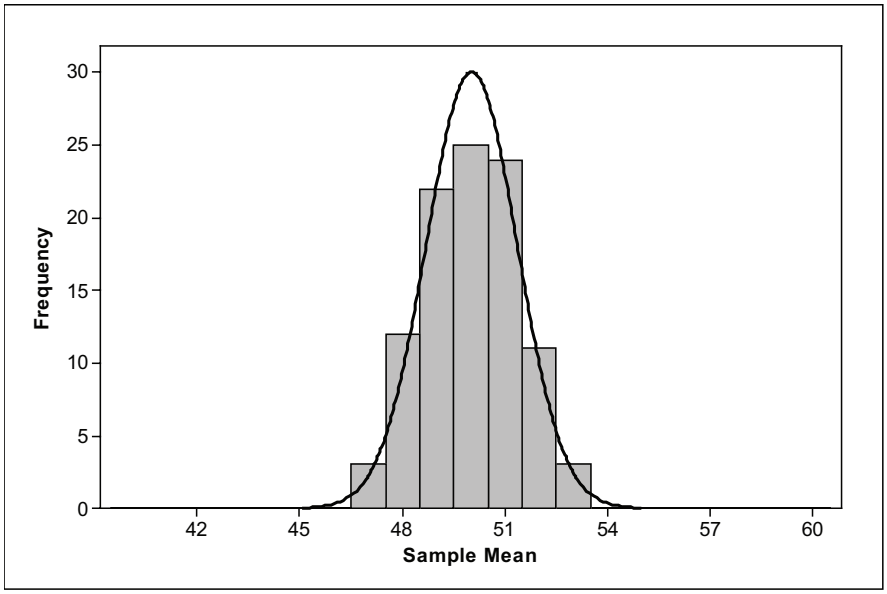
1. The shape is roughly symmetric and mound-shaped. Except for the fact that the population consists of whole numbers so that the distribution can't be normal, it is approximately normally distributed.



2. a. Sample answer (based on sample data):

50.44	51.33	49.56	49.22	50.22	50.22	48.67	51.44	51.00	48.33
48.11	51.78	49.11	51.33	49.78	47.89	52.22	49.00	50.56	51.22
49.44	50.56	48.67	51.22	49.78	49.89	50.56	48.44	50.67	48.11
51.89	48.22	50.67	49.11	48.56	52.11	49.22	50.44	49.89	52.89
52.44	49.33	48.22	50.11	50.89	50.33	50.33	51.67	50.56	50.44
47.22	48.67	49.22	48.78	50.56	52.89	50.44	49.44	50.78	49.11
47.78	48.33	49.44	49.22	50.11	47.00	50.44	50.56	51.11	51.33
50.33	51.67	51.00	51.56	48.33	47.22	50.67	49.44	51.56	50.89
50.56	49.44	47.78	50.00	51.89	48.11	50.44	50.56	48.89	53.22
49.89	49.67	49.22	50.33	49.67	49.11	51.78	50.22	50.67	49.56

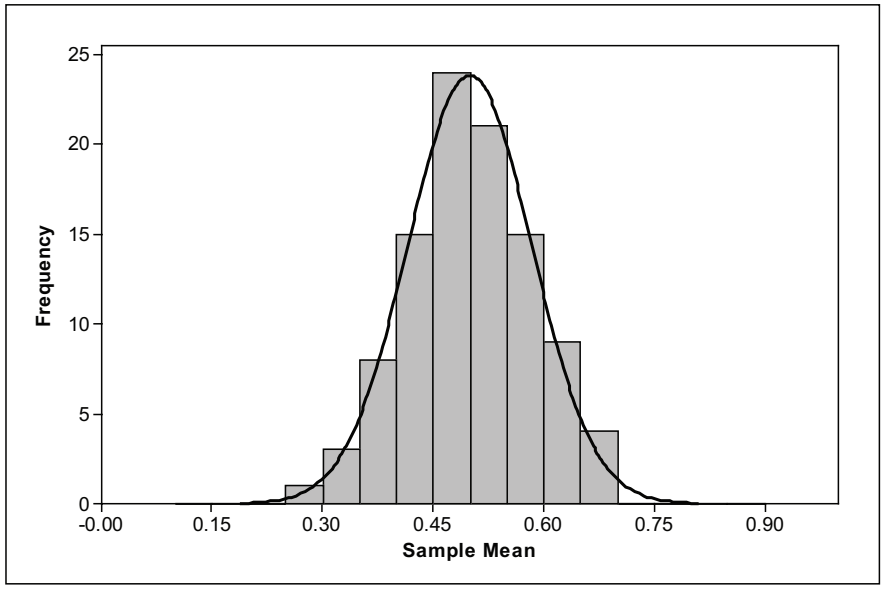
b.



Both the population distribution and sampling distribution of $\bar{x}$ are fairly symmetric and mound-shaped, and appear approximately normal in shape. Both are centered at around 50. However, the sampling distribution is more concentrated about its center, and not as spread out as the population distribution.

Extension

3. c. The distribution will be roughly symmetric and mound-shaped – or approximately normally distributed – similar to the one that follows.



The histogram should be centered close to 0.5, which is the center of the population distribution curve shown in Figure 22.10. However, the sampling distribution will be less spread out than the population distribution, which is evenly spread between 0 and 1. Instead it will be concentrated around 0.5 and trail off on either side of 0.5.

EXERCISE SOLUTIONS

1. a. The mean of the three measurements has standard deviation

$$\frac{\sigma}{\sqrt{n}} = \frac{0.08}{\sqrt{3}} \approx 0.046$$

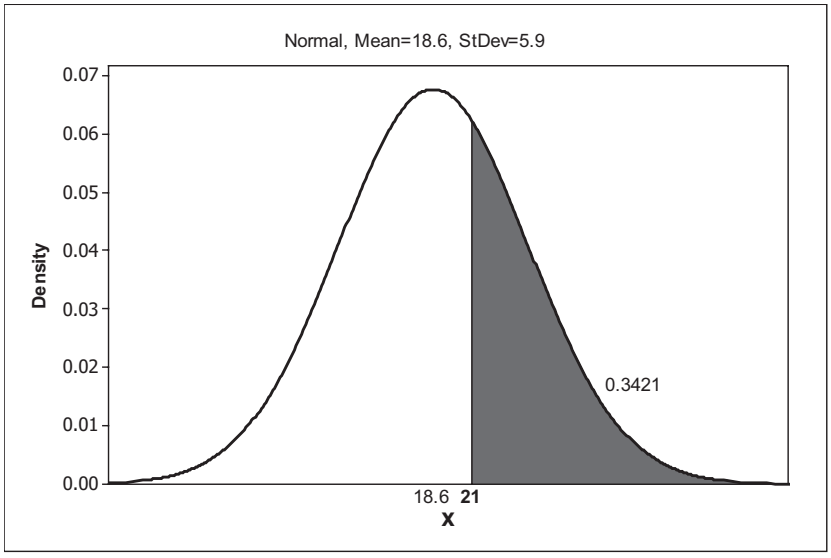
b. The mean of three measurements is less variable than a single measurement. That is why it is good practice to repeat a laboratory measurement several times independently and report the average.

2. This exercise contrasts the distribution of one score with that of the average of 55 scores.

a. We need to calculate the probability that a normal random variable *x* has value 21 or greater. We first convert to *z*-scores so that we can use standard normal tables:

$$P(x \geq 21) = P\left(\frac{x - 18.6}{5.9} \geq \frac{21 - 18.6}{5.9} \right) = P(z \geq 0.41) = 1 - 0.6591 = 0.3409$$

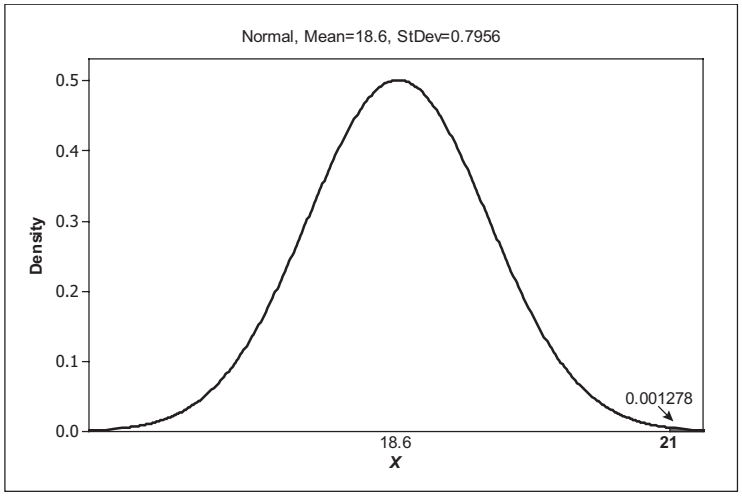
Note: We could also use software such as Minitab or a graphing calculator to compute the probability directly, as shown below. This gives a more accurate answer.



b. The sampling distribution of $\bar{x}$ for samples of size 25 is normally distributed with mean 18.6 and standard deviation $5.9/\sqrt{55} \approx 0.7956$. Below are the calculations for computing the probability using a standard normal table:

$$P(\bar{x} \geq 21) = P\left(\frac{\bar{x} - 18.6}{0.7956} \geq \frac{21 - 18.6}{0.7956} \right) = P(z \geq 3.02) = 1 - 0.9987 = 0.0013$$

We can calculate this probability directly using software:



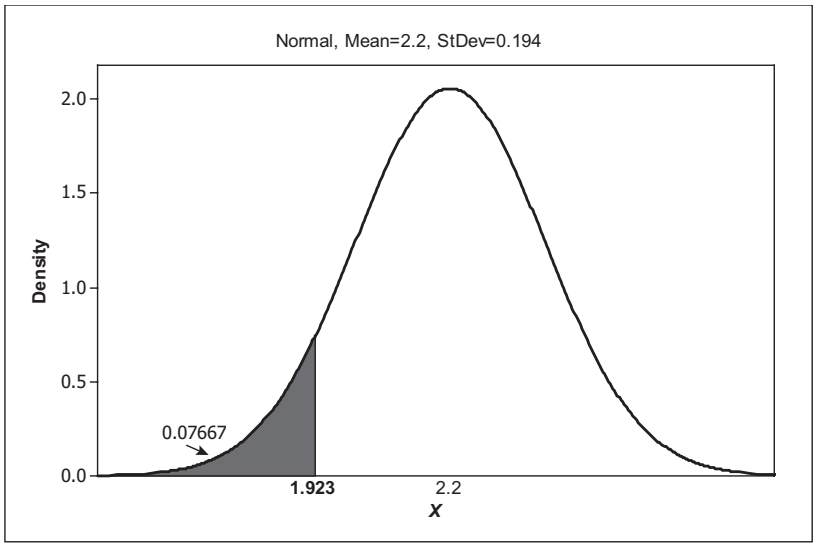
The mean of many scores is less likely to be far away from the population mean than is a single score.

3. a. The distribution is approximately normal with mean 2.2 and standard deviation $1.4/\sqrt{52} \approx 0.194$.

b. Again, transforming in order to use the standard normal table gives:

$$P(\bar{x} < 2) = P\left(\frac{\bar{x} - 2.2}{0.194} < \frac{2 - 2.2}{0.194} \right) = P(z < -1.03) = 0.1515$$

c. To say that the total is less than 100 is exactly the same as saying that the mean per week is less than 100/52 or 1.923. Now, we can proceed to find $P(\bar{x} < 1.923)$, which we find directly using software:



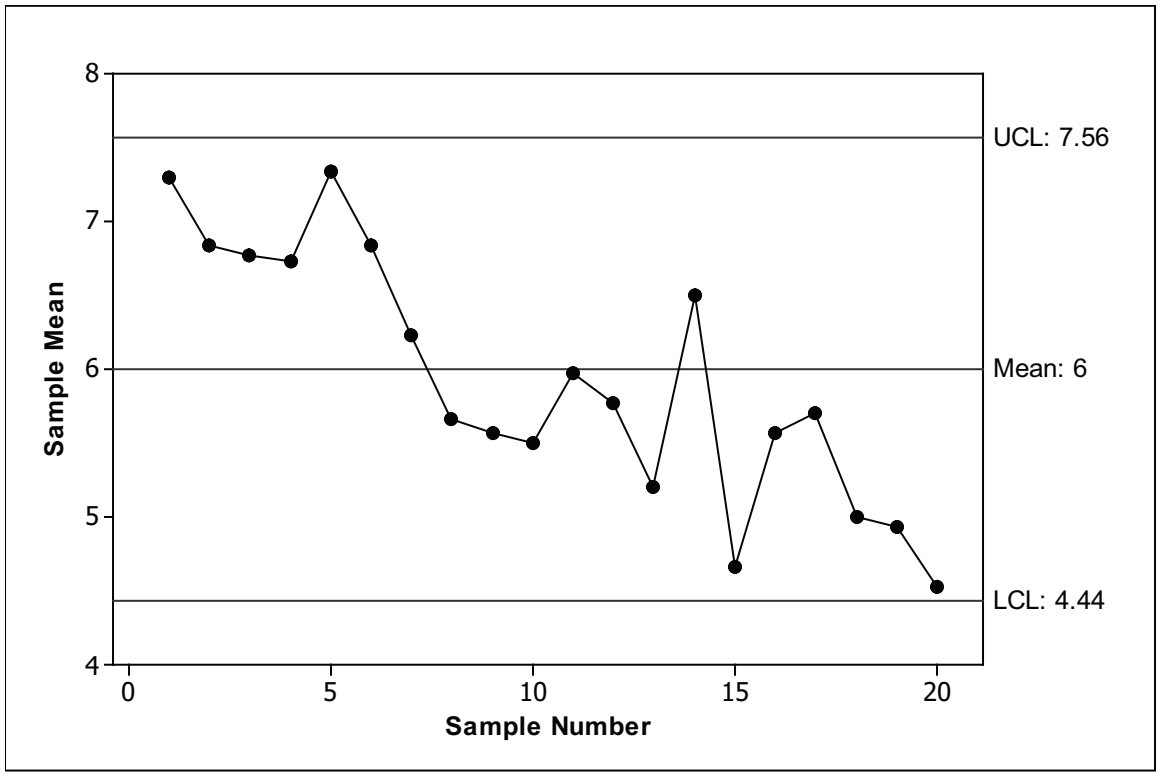
4. a. The lower and upper control limits are $\mu \pm 3\sigma/\sqrt{n}$. In this case, the limits are $6 \pm (3)(0.9/\sqrt{3})$ or LCL ≈ 4.44 and UCL ≈ 7.56 (rounded to two decimals).

b. Samples collected over a 24-hour time period appear in Table 22.3. The sample means appear below.

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

21	6.6	5.2	6.8	6.2
22	6.4	6.0	5.9	6.1
23	6.4	4.6	6.7	5.9
24	7.0	6.3	7.4	6.9

c.



d. No, none of the sample means is above the upper control limit or below the lower control limit.

e. Although the sample means lie within the control limits, the process appears to be changing. Initially, the sample means were fairly close to the upper control limit. Over time, the sample means tended to decrease. If this pattern continues, the sample means will begin to fall below the lower control limit. So, this process does not appear to be stable.

REVIEW QUESTIONS SOLUTIONS

1. a. Using the 68-95-99.7 rule, 95% of the caps will have diameters within two standard deviations of the mean – hence, between 0.497 and 0.503 inches. Thus, around 5% of the bottles will have diameters outside of the chemical manufacturer's specification limits.

b. $\bar{x}$ will have a normal distribution with mean 0.500 inch and standard deviation $0.0015/\sqrt{9}$ = 0.0005 inch.

c. The endpoints of the acceptance interval can be written as 0.500 ± 0.001, which is equivalent to 0.005 ± 2(0.0005). Hence, 95% of the samples will have means within this interval. The production process will be stopped 5% of the time (or a proportion of 0.05).

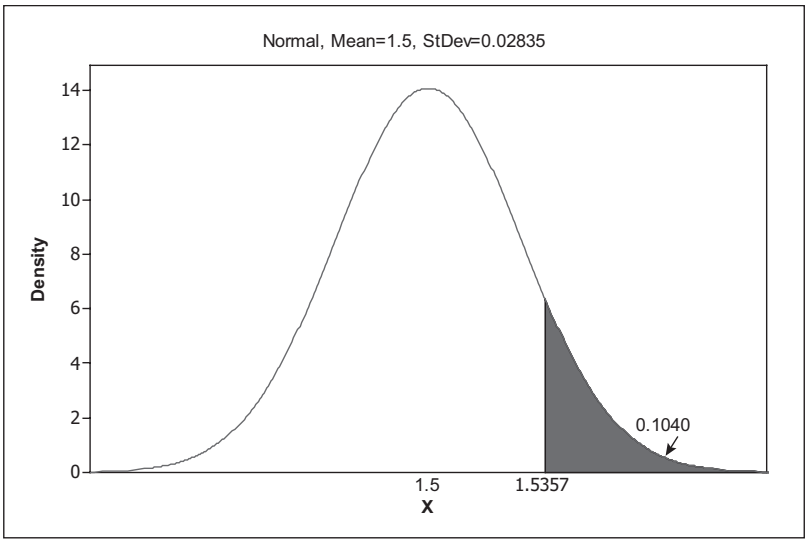
2. a. No. The number of people in a car must be a whole number. A normal random variable can take any value, not just whole number values. (Normal distributions are *continuous* distributions.)

b. The sample mean $\bar{x}$ has an approximately normal distribution with mean 1.5 and standard deviation $0.75/\sqrt{700} \approx 0.02835$

c. The total in 700 cars exceeds 1075 people exactly when the mean $\bar{x}$ exceeds 1075/700 ≈1.5357 persons per car. So, the probability we want is:

$$P(\bar{x} > 1.5357) = P\left(\frac{\bar{x} - 1.5}{0.02835} > \frac{1.5357 - 1.5}{0.02835}\right) = P(z > 1.26) = 1 - 0.8962 = 0.1038$$

We can compute this directly using software (see chart next page):



3. a. $\mu_{\bar{x}} = 90$ seconds; $\sigma_{\bar{x}} = 120/\sqrt{10} \approx 37.9$ seconds. The shape will be less skewed to the right than the original distribution. Given that the sample size is relatively small, you can't say much more than that.

b. $\mu_{\bar{x}} = 90$ seconds; $\sigma_{\bar{x}} = 120/\sqrt{100} = 12$ seconds. Since the sample size is large, by the Central Limit Theorem we can say that the shape of the distribution will be approximately normal.

c. First, we convert 2 minutes into 120 seconds. We need to calculate $P(\bar{x} > 120)$, which we determine using software, which gives $P(\bar{x} > 120) \approx 0.0062$ as shown below.