CHAPTER 6

Visualizing Random Samples

CONCEPTS

- Statistical Inference, Sample, Population, Infinite Population, Finite Population, Sample Size, Random Sample, Random Variation, Sample with Replacement, Sample without Replacement, Outliers

OBJECTIVES

- Understand the variability of samples in relation to their parent populations and the role sample size plays in that variability

- Recognize how the number of histogram intervals and the labeling of the horizontal axis affect the appearance of a sample

- Learn to infer a population's shape, mean, and standard deviation from a sample

- Understand the effects of sampling with or without replacement

- Recognize how outliers affect histograms and what they represent

Overview of Concepts

A **population** represents all possible outcomes of an experiment or a process. A **finite population** consists of N possible outcomes that can be enumerated or listed (e.g., the number on a roulette wheel). An **infinite population** has an inexhaustible and uncountable number of possible outcomes (e.g., waiting times for pizza deliveries). Some finite populations can be treated as if they are infinite because they are so large (e.g., prices of all new automobiles sold in California last year). Because many populations are very large, statisticians often make a judgment about a population based upon a small portion of the population called a **sample**. This judgment is called a **statistical inference**. For example, you are probably familiar with exit polls that make inferences about election results based upon a sample of voters.

If the sample represents the population, it can provide insight into the population at a fraction of the cost of analyzing the entire population. However, if a sample does not represent the population, it can lead to poor and even embarrassing decisions. Two related cases illustrate this point. The *Literary Digest* predicted that Alfred Landon would easily defeat the incumbent, Franklin D. Roosevelt, in the 1936 Presidential election. One week later, Roosevelt won by a record 11 million votes. What went wrong? The *Literary Digest* drew its sample from phone records, car registrations, and its own subscription list. This sample represented wealthier Americans who were more likely to vote Republican. The *Digest* incorrectly believed that because they had a very large **sample size** their results were very accurate. However, they had actually selected a biased sample, albeit a very large one.

In another case, George Gallup predicted that Thomas E. Dewey would defeat the incumbent Harry S. Truman in the 1948 Presidential election. Ten days later Truman won the election. That forecast was based upon the views of voters who preferred a candidate when the poll was taken — ten days before the election. However, there were a large number of undecided voters at the time of the poll, most of whom voted for Truman. By ignoring the undecided voters Gallup inadvertently biased his sample.

Ironically, in 1936, George Gallup, in his weekly newspaper column, had been critical of the *Literary Digest* poll because it over-represented wealthier Americans and hence was not a **random sample** (every outcome in the population has an equal chance of being selected). Yet 12 years later he inadvertently committed a similar error.

Even if you draw a random sample there will be **random variation**. This is the normal variation expected in a sample. However, if you find an **outlier** (a sample point that is more than three standard deviations from the mean) it is difficult to know if you are seeing random variation or if you have an observation from a different population. For example, you select 10 high school students at a football game and ask them the approximate size of their homes (in square feet). You collect the following data: 1500, 2000, 2200, 1800, 1600, 2300, 5000, 2800, 1900, and 2700. The 5000 is an outlier. Does it represent normal variation, or is it an observation from another population? You would not know unless you asked the subject where she lived. Only then might you find out that she was a student from another community.

In sampling finite populations you can sample with or without replacement. If you **sample with replacement**, after selecting an observation it is returned to the population, possibly to be drawn again. In contrast, if you **sample without replacement** the observation drawn is removed from the population, changing the probability of selecting the next observation. If the sample size is very small relative to the population size the methods are equivalent.

Illustration of Concepts

Consider the distribution of household size in the United States. Every three months the U.S. Bureau of Labor Statistics surveys consumers using the Consumer Expenditure Survey. This study used 5,153 households from the fourth quarter 1989 and the first quarter 1990. This is the **finite population** that will be sampled. A **random sample** of 25 consumers is drawn from this **population**. Since the population is large relative to our **sample size,** we can treat it as an **infinite population.** Therefore, it doesn't matter if we **sample with** or **without replacement**. A histogram of the **sample** is shown in Figure 1. What **statistical inferences** can be made about the population? From the histogram we would infer that the population is right skewed and is triangular in shape with a mode at a one-person household. For this sample size, we have more confidence in our inference that the population is right skewed than we do in its specific shape or its mode. Its sample mean is 2.2 and standard deviation is 1.27.

We draw a second and third sample of 25 from the same population (Figures 2 and 3). The second sample suggests that the population is not as skewed and is bimodal with modes at 2 and 4 persons per household. Its sample mean is 2.6 and standard deviation is 1.20. The third sample suggests that the population is uniform in the one- to four-person range with a slight tail to the right. Its sample mean is 2.60 and standard deviation is 1.26. The only similarity in the three histograms is that all suggest some positive skewness. None of the three samples has an **outlier** (an observation more than 3 standard deviations from the mean).

These histograms illustrate **random variation**. This differing picture of the population is not unusual, especially if the sample size is small. In contrast, note the similarity in the sample means and standard deviations. These statistics vary much less than the histograms.

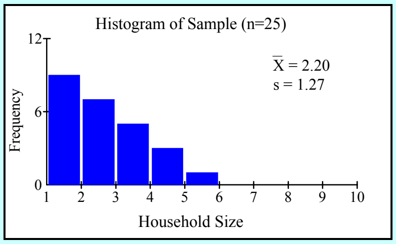


Figure 1: Histogram of First Sample

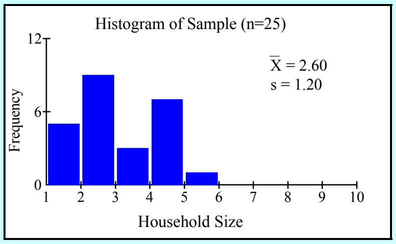


Figure 2: Histogram of Second Sample

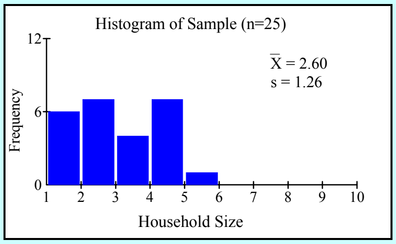


Figure 3: Histogram of Third Sample

For comparison, a histogram of the entire finite population is displayed in Figure 4. The second sample is most like the population, even though it incorrectly led us to suspect a bimodal distribution. Although it was not possible to infer the exact shape of the population from our samples of n = 25, our belief that the population was positively skewed was correct, and our estimates of the population mean and standard deviation were good. Figure 5 shows a histogram of a larger sample (n = 100). Compare its shape with the population histogram, and its sample statistics (sample mean is 2.66 and standard deviation is 1.50) to the population parameters ($\mu = 2.74$, $\sigma = 1.53$). Inferences from larger samples generally are more accurate.

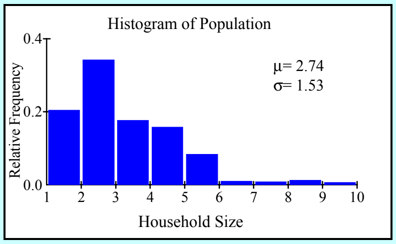


Figure 4: Histogram of Population

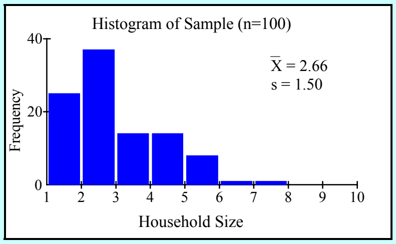


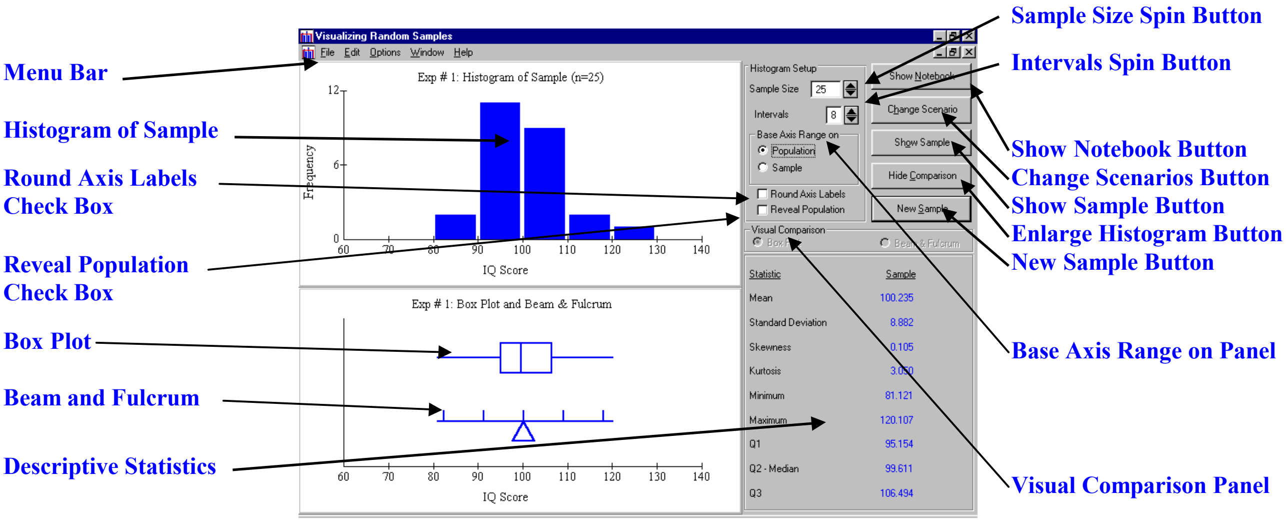
Figure 5: Histogram of Large Sample

Orientation to Basic Features

This module allows you to sample from a population distribution or a finite population that you select and create a histogram to display the sample. You can change the number of intervals in the histogram and superimpose the population for comparison.

1. Opening Screen

Start the module by clicking on the module's icon, title, or chapter number in the *Visual Statistics* menu and pressing the Run Module button. When the module is loaded, you will be on the introduction page of the Notebook. Read the questions this module covers and then click the Concepts tab to see the concepts that you will learn. Click on the Scenarios tab. Select Normal Populations from the list of choices. Select the IQ Scores scenario, read it and press OK. The upper left of the screen depicts a histogram with eight intervals for a sample of size 25. The Control Panel appears on the right. The bottom left of the screen shows a box plot and a beam and fulcrum diagram of the sample. Descriptive statistics from the sample are shown to the right. Other features are controlled from the menu bar at the top of the screen.



Histogram Setup

Sample Size 25

Intervals 8

Base Axis Range on

● Population

○ Sample

□ Round Axis Labels

□ Reveal Population

Statistic	Sample
Mean	100.235
Standard Deviation	8.882
Skewness	0.105
Kurtosis	3.050
Minimum	81.121
Maximum	120.107
Q1	95.154
Q2 - Median	99.611
Q3	106.494

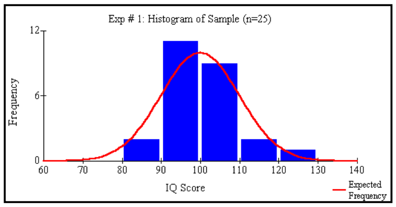
2. Control Panel

a. Press the Intervals spin button. The histogram is automatically updated to reflect the number of intervals desired (2 to 20). Double clicking on the number in the spin box enables you to change it from your keyboard.

b. Press the Sample Size spin button. A new sample size (2 to 1000) can be selected. The sample is drawn when the flashing New Sample button is pushed. Double-click the number in the spin box to change it from your keyboard.

c. The Base Axis Range On panel contains two option buttons. Click Sample to base the histogram labels on the sample drawn. Click Population to base the axis scale on the population sampled. The Population option is useful if you are comparing the sample with the population from which it is drawn because it will not change with every sample.

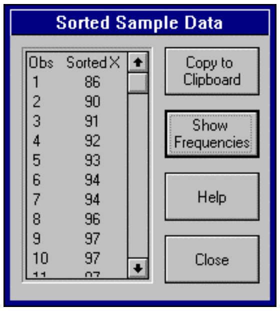
d. Click on the Round Axis Labels check box to improve horizontal axis labeling, making it more pleasing to the eye. When the histogram is based on the sample, this option usually improves the roundness of the numbers on the axis.



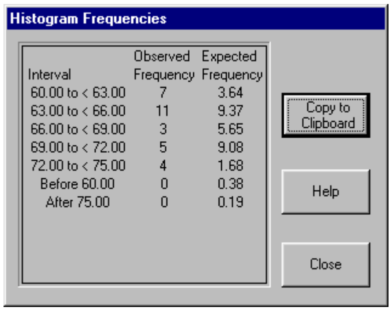
e. Click on the Reveal Population check box to see the population superimposed upon the sample histogram, a visual comparison of the sample and population, and the population statistics listed next to the sample statistics. The Visual Comparison panel is now active, allowing you to select a Box Plot or Beam & Fulcrum display.

f. Press the Hide Comparison button to enlarge the histogram vertically. Press the Show Comparison button to return the histogram to normal size and show both displays.

g. Press the Show Sample button to see the sample in sorted order. Press the Show Frequencies button to see the observed frequency in each interval. If the population is being revealed, the expected frequency is also shown. Press the Copy to Clipboard button to copy a table to other applications.


Sorted Sample Data

Obs	Sorted X
1	86
2	90
3	91
4	92
5	93
6	94
7	94
8	96
9	97
10	97


Histogram Frequencies

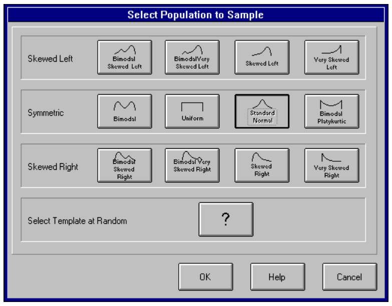
Interval	Observed Frequency	Expected Frequency
60.00 to < 63.00	7	3.64
63.00 to < 66.00	11	9.37
66.00 to < 69.00	3	5.65
69.00 to < 72.00	5	9.08
72.00 to < 75.00	4	1.68
Before 60.00	0	0.38
After 75.00	0	0.19

3. Changing the Population

There are two easy ways to select the population being sampled.

a. You can use a scenario. The purpose of a scenario is to provide a context showing how a statistician would use a sample. A different scenario can be selected by pressing the Change Scenario button and turning the pages of the Notebook.

b. You can sample from 12 predefined populations. Press the Show Notebook button, select the Pop. Templates tab and click OK. A template like the one to the right appears. Click any population button or the random population button (?) and click OK. The Change Scenario button becomes the Change Templates button, providing a shortcut to the templates in the Notebook.



4. Copying a Display

Click on the display you wish to copy. Its window title will be highlighted. Select Copy from the Edit menu (on the menu bar at the top of the screen) or Ctrl-C to copy the display. It can then be pasted into other applications, such as Word or WordPerfect, so it can be printed.

5. Help

Click on Help on the menu bar at the top of the screen. Search for Help lets you search a topic index, Contents shows a table of contents, Using Help gives instructions on how to use Help, and About gives licensing and copyright information about this *Visual Statistics* module.

6. Exit

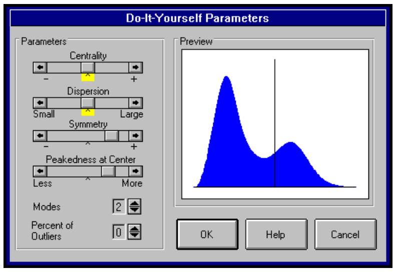
Close the module by selecting Exit in the File menu (or click ☒ in the upper right-hand corner of the window). You will be returned to the *Visual Statistics* main menu.

Orientation to Additional Features

1. Changing the Population

There are three advanced ways to change the population.

a. You can sample from a continuous population of your own creation. Press the Show Notebook button, select the Do-It-Yourself tab, click on User Created Distributions and click OK. Four scroll bars control the population's parameters. The Modes spin button sets the number of modes (0, 1, or 2) in the distribution being sampled. The Percentage of Outliers spin button sets the percent of outliers (1, 2, 3, 4, or 5). These are sampled from a second population with a mean 4 standard deviations away from the original mean. Click OK to sample from the population created. The Change Templates button becomes the Change Parameters button, providing a shortcut to this control in the Notebook.

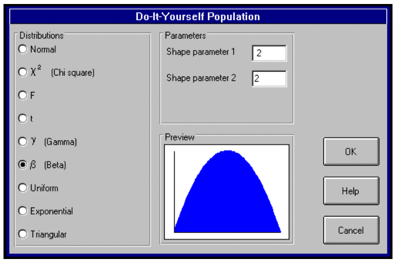


Do-It-Yourself Parameters

Modes 2

Percent of Outliers 0

b. You can sample from 9 known distributions. Press the Show Notebook button and click on Next page in the lower right corner. Click OK. Select a population from those listed. Change the parameter values using the text boxes. The range of values appears if the cursor is placed on a text box. The distribution is shown on the accompanying graph. Click on OK to sample from the population selected. The Change Parameters button becomes the Change Populations button, providing a shortcut to this control in the Notebook.



Do-It-Yourself Population

Distributions

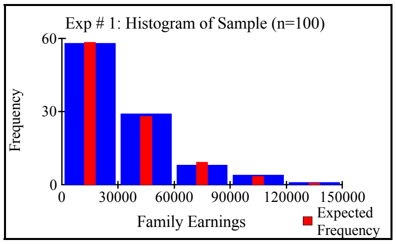
○ Normal
○ χ^2 (Chi square)
○ F
○ t
○ γ (Gamma)
● β (Beta)
○ Uniform
○ Exponential
○ Triangular

Parameters

Shape parameter 1 2

Shape parameter 2 2

c. You can sample from either of two databases. Press the Show Notebook button and select the Databases tab. Read the descriptions and select a database. Each database is organized by categories. Click on the + symbol of any category that sounds interesting to expand the category and list its variables (the + symbol will become a – symbol). Click on the – symbol to shrink a category and hide its variables. Click on any variable and read its description in the text window at the right. Click on OK to sample from this finite population. You can sample with or without replacement. The Change Parameters button becomes Change Populations. Population frequencies can be superimposed on the sample histogram.



2. Options

Two options are available from the Options menu:

a. Select Change Title from the Options menu to retitle the current display.

b. Select Full Window Graph from the Options menu to extend both graphs to the full screen width. Deselect Full Window Graph to bring back the Control Panel.

3. Second Display

Under Window, click Copy Default Window or Copy Current Window from the Windows menu on the menu bar. This creates a second graph that can be tiled or cascaded. Using the Change Title option to retitle the displays, so you can keep track of them.

Basic Learning Exercises

Name ________________

Sampling Variation

Select Sample in the Base Axis Range On panel. Press the Show Notebook button, select the Scenarios tab and click on Non-normal populations with one mode. Select The Sum of Two Random Numbers scenario. Read the scenario. Click OK. *Don't* select Population or Reveal Population yet.

1. Set Sample Size to 25. Use the Intervals spin box to create a histogram that best illustrates the sample (you may select Round Axis Labels if you wish). a) How many intervals did you select? Using only the sample histogram, sketch your impression of the shape of the population's distribution. Repeat this for three more samples. b) How do your sketches differ from one another? How are they similar? c) Why don't the sketches portray a consistent picture of the population's distribution?

2. Take a new sample. Record the sample mean, standard deviation, skewness, kurtosis, and median. Repeat the process for three more samples. Subtract the smallest mean from the largest (this is the range of the sample means in this experiment). Calculate the ranges of the other sample statistics.

Sample	Mean	Std. Dev.	Skewness	Kurtosis	Median
1					
2					
3					
4					
Range					

3. Why is sample variation more evident in the histograms (exercise 1) than in the sample mean or sample median (exercise 2)?

Sample Size and Sample Variation

4. Increase the sample size to 500 (use the spin button or enter 500 into the Sample Size box) and press the New Sample button. Using only the sample histogram, sketch your impression of the shape of the population's distribution. Repeat this for three more samples. Why are the sample histograms more consistent than in exercise 1?

5. Repeat exercise 2 using a sample size of 500.

Sample	Mean	Std. Dev.	Skewness	Kurtosis	Median
1					
2					
3					
4					
Range					

6. a) Compare the range of each statistic in exercises 2 and 5. What do you observe? b) Compare your sketches in exercises 1 and 4. What do you observe? c) Give a general rule regarding the relationship between sample size, sample variation, variation in histograms, and variation in sample statistics. d) Why did increasing the sample size increase your confidence in your impressions about the population?

7. Click the Reveal Population check box. Select Population in the Base Axis Range On panel. Were your impressions about the shape of the distribution and its statistics correct?

Intermediate Learning Exercises Name ________________

Making Inferences

One key to making inferences about unknown populations is seeing how the shape of the population's distribution affects the appearance of a sample histogram, while keeping the sample size and number of histogram intervals the same.

8. Push the Show Notebook button and select the IQ Test Scores scenario. Read the scenario and click OK. Change the sample size to 20, set the number of histogram intervals to 5, and take a sample. Sketch your impression of the shape of the population's distribution. Repeat this for four more samples. Then select Reveal Population. Were your sketches correct?

   IQ Test Scores: Population Distribution __________

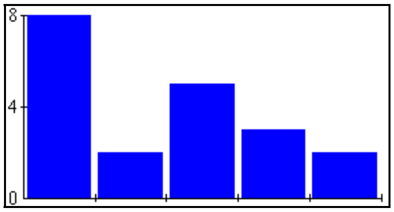
9. Repeat exercise 8 using The Accuracy of an Archer scenario from the Notebook.

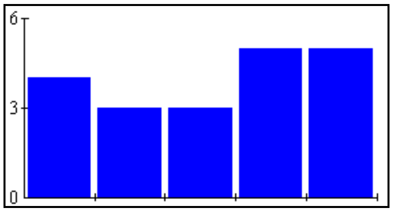
   Accuracy of an Archer: Population Distribution ____________

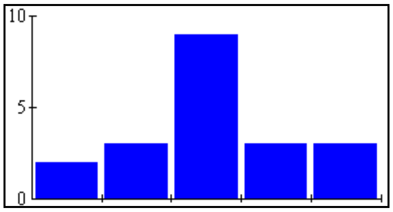
10. Repeat exercise 8 using the Choosing a Lottery Number scenario from the Notebook.

    Choosing a Lottery Number: Population Distribution ___________

11. All of the following three histograms were generated by one of the three populations you analyzed in exercises 8–10. Which scenario is most likely to have generated these histograms? The least likely? Explain your reasoning.







Sampling from a Finite Population

12. Press the Show Notebook button and select the Databases tab. Read about each database and select one that interests you. Read the descriptions of several variables and select one that interests you. Which variable did you select? Give its definition and population size (N).

13. Make sure that Sample with Replacement is checked. Click OK. Reveal Population should *not* be selected. Use a sample size of n = 40 for the grocery data or n = 100 for the expenditures data. Record the sample size and draw five samples, recording each mean. Calculate the range of the 5 sample means. Triple the sample size and repeat the process.

n	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Range

14. Return the sample size to its original size (either 40 or 100). Press the Change Database button. Click on Sample with Replacement to deselect it. Press OK. Repeat exercise 13.

n	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Range

15. a) For each sample size, calculate the ratio of the range of sample means when you sampled with and without replacement. b) What do you conclude from these comparisons? c) Why would the range of sample means be zero if you sampled without replacement and n = N?

Advanced Learning Exercises Name ________________

Sampling with Outliers

16. Click Reveal the Population to deselect this option. Press the Show Notebook button, select the Do-It-Yourself tab, click on User Created Distributions and click OK. A display will let you create your own distribution by changing the scroll bars and the Number of Modes spin button. Create any distribution you desire. Sketch and describe the shape of the distribution you created. What data might have this type of distribution? Defend your selection.

17. Use the Outliers spin button to set the percentage of outliers to 3%. Click OK to sample from the distribution selected. What is an outlier? (**Hint:** Use Help.) What could cause an outlier in a real sampling experiment? What warning does it give the statistician?

18. Not all samples will contain an outlier. Set your sample size to 100. Select the number of intervals that you believe is appropriate. Press the New Sample button. How can you tell if the data displayed contains an outlier? If it doesn't, draw new samples until it does (it would be very unusual to take more than two samples).

19. Click on Reveal the Population. From this display how can you tell that the outlier did not come from the population displayed?

20. Can a statistician assume that if a sample contains an outlier that the sample has an observation from another population?

Sampling from Known Distributions

21. Press the Show Notebook button. Select the Do-It-Yourself tab and click on Known Distributions. Press OK to bring up its Control Panel. Nine different continuous distributions can be sampled. The parameters of the selected distribution are listed in the Parameters panel. A sketch of the distribution is shown for reference. Select each and change its parameter values. The range of valid values is displayed in the tool tip (rest the cursor on a parameter value box). For each distribution, describe its shape and how its parameter(s) affect its shape.

    Normal

    Chi Square

    F

    Student's t

    Gamma

    Beta

    Uniform

    Exponential

    Triangular

22. An exponential distribution is often used to model processes that have constant failure rates. Its single parameter is the failure rate, whose reciprocal is the mean of the distribution. For example, many 60W light bulbs have an advertised life of 1000 hours, which implies a failure rate of 0.001 or that 0.1% of the light bulbs fail every hour. Select an exponential distribution and enter 0.001 for its parameter value. Click OK. Take a sample of 100 light bulbs. Based on this sample, how many would fail in the first 250 hours? The first 500 hours? The first 1000 hours? How many hours before the last light bulb burns out?

Bimodal Populations

23. Bimodal distributions usually represent situations where two different populations are being sampled. Press the Show Notebook button, select the Scenarios tab and click on Populations with 0 or 2 modes. Read Heights of Students. Why is this distribution bimodal? Give your own example of a bimodal population. Press OK to study the sample. Can the two modes be identified in the sample? If a sample is bimodal does this mean that the population is also?

Individual Learning Projects

Write a report on one of the three topics listed below. Use the cut-and-paste facilities of the module to place the appropriate graphs and tables in your report.

1. Investigate the importance of sample size in reducing sample variability. Select a data set from one of the two databases to investigate. Describe the variable you selected. What do you think its upper and lower bound would be and why? Start your investigation with a sample size of 12. Draw three samples. For each sample provide a histogram (based upon the sample). Do the three samples provide a consistent impression of the shape of the distribution being sampled? If not, increase the sample size to 25 and reevaluate the histograms. Continue doubling the sample size until all three samples give a consistent impression of the distribution's shape. Include the histograms of these three samples (the histograms should all use the same sample size) in your report along with a histogram of the population. Repeat this process for a different data set that you believe will have a different distribution shape. Discuss the importance of sample size in overcoming random variation and how random variation was affected by the distribution's shape.
2. Investigate the effect sampling with or without replacement has on the variability of the sample mean. Select a data set to investigate. Conduct six experiments, sampling with and without replacement using three different sample sizes: less than 5% of the finite population size, 15 to 35% of the finite population size, and 60 to 85% of the finite population size. In each case take 10 samples and record the sample mean. For each experiment calculate the average and the variance of the 10 sample means. Your project should show that the variability of the sample means decreases as sample size increases and is further reduced if you sample without replacement, especially if the sample size is a large portion of the population size. Your report should include the histogram of the sample with the most variability and the histogram of the sample with the least variability from each of the six experiments.
3. Illustrate statistical inference. Select three different data sets (you must use both databases). Describe the variables you selected. Tell what you believe their upper and lower bounds would be and why. Use a sample size of 40 for the Grocery Expenditure Survey and 100 for the Consumer Expenditure Survey. Decide if you are going to sample with or without replacement and if you are going to use or not use Round Axis Labels. For each data set take one sample. Create a histogram (based upon the sample data) with the appropriate number of intervals. Analyze the box plot and beam and fulcrum diagrams. Evaluate the descriptive statistics. Make an inference about the population's shape and population statistics (mean, standard deviation, etc.) and your degree of confidence in the inferences. Reveal the population and evaluate your answer. Your answer will be judged on your explanation, not on whether you correctly identified the shape of the distribution. Your report should include evaluations of the usefulness of the box plot, the beam and fulcrum, histograms, and descriptive statistics in making inferences. Discuss the feasibility of creating a histogram based the population's upper and lower bound.

Team Learning Projects

Select one of the three projects listed below. In each case produce a team project that is suitable for an oral presentation. Use presentation software or large poster boards to display your results. Graphs should be large enough for your audience to see. Each team member should be responsible for producing some of the graphs. Ask your instructor if a written report is also expected.

1. A team of three to five individuals should investigate the importance of the number of intervals in producing a histogram. Each team member should choose a different shape distribution by using the Pop. Templates tab or the Do-It-Yourself tab in the Notebook (select distributions as diverse as possible). Give an example of data that would have such a distribution. Set Sample Size to 25, set Base Axis Range on to Sample, and, if you wish, click on Round Axis Labels. Draw four samples for each distribution. For each sample create two histograms, each with a different number of intervals. The project should illustrate how the number of intervals affects the impression of a distribution's shape and how variability is affected by the shape of the distribution that generates the sample.
2. A team of three to five should investigate the usefulness of the box plot and the beam and fulcrum diagrams in making inferences about a distribution's shape. Each team member should choose a different shape distribution by using the Pop. Templates tab or the Do-It-Yourself tab in the Notebook (select distributions as diverse as possible in skewness and peakedness). Give an example of data that would have such a distribution. Set Sample Size to 15, set Base Axis Range on to Sample, and, if you wish, click on Round Axis Labels. Using only the box plot, make an inference about the shape of the distribution that generated the data. Using only the beam and fulcrum, make an inference. Draw two more samples and repeat the process. Increase your sample size to 60 and repeat the process with 3 more samples. Which diagram is more useful in making an inference when n = 15? When n = 60? Consider the variability as well as the reliability of each diagram at each sample size. As a team, discuss how the shape of the distribution that generated the sample affected these results. Each team member should have a display illustrating six pairs of diagrams as well as a representation of the population.
3. A team of three to five should investigate how sample size, as well as the shape of the distribution that generates a sample, affects random variability. Each team member should choose a different shape distribution by using the Pop. Templates tab or the Do-It-Yourself tab in the Notebook (select distributions as diverse as possible in skewness and peakedness). Give an example of data that would have such a distribution. Set Sample Size to 10, set Base Axis Range on to Sample, and, if you wish, click on Round Axis Labels. Create a histogram that accurately displays the sample. Take two more samples using the same number of intervals. Increase your sample size to 30. Create three more histograms (adjust the number of intervals appropriately). Increase your sample size to 90 and create three more histograms. The project should illustrate how sample size reduces random variability and how variability is affected by the shape of the distribution that generates the sample.

Self-Evaluation Quiz

1. A sample has a histogram shaped like its population
   a. regardless of sample size.
   b. if the sample size is large enough.
   c. if the population is normal.
   d. if the sample is drawn randomly.
   e. More than one of the above are correct.

2. Sample means generally
   a. have the same variability as individual items in the population.
   b. have more variability than individual items in the population (due to sampling error).
   c. have less variability than individual items in the population.
   d. may be more or less variable than population items, depending on the population shape.
   e. may most easily be seen in the sample box plot.

3. Which is true of random variation?
   a. Variation is decreased as sample size is decreased.
   b. Variation is larger if a peaked distribution is sampled than a uniform distribution.
   c. As random variation is reduced, confidence in inferences is increased.
   d. Variation is smaller if a distribution has two modes rather than one.
   e. None of the above is correct.

4. If a normal population is sampled,
   a. the sample box plot should be roughly symmetric.
   b. the sample beam and fulcrum should be roughly symmetric.
   c. the sample median should roughly equal the sample mean.
   d. All of the above are correct.
   e. Only a and c are correct.

5. If we sample a given population, increasing the sample size will cause the
   a. sample means to vary more from sample to sample.
   b. sample standard deviations to vary more from sample to sample.
   c. histogram range for each sample to grow narrower.
   d. histogram of each sample to approach normality.
   e. histogram of each sample to approach the population shape.

6. As we increase the number of classes in a sample histogram, which would be expected?
   a. The population's mode (if any) becomes more apparent.
   b. The population's range becomes more apparent.
   c. The frequencies in each class interval become larger.
   d. Some empty intervals are likely to appear.
   e. All of the above would be expected.

7. A right-skewed sample histogram is *least* likely to be from a
   a. uniform population.
   b. right-skewed population.
   c. normal population.
   d. left-skewed population.
   e. bimodal population.

8. When sampling a uniform population, a bimodal sample histogram
   a. could be due to sampling variation.
   b. could reflect the number of histogram intervals that were used.
   c. could be due to the interval limits that were chosen.
   d. could arise from using a small sample size.
   e. could be due to any of the above.

9. Which of the following statements is true?
   a. Sampling with replacement decreases sample variability.
   b. Sampling with or without replacement is not an issue in infinite populations.
   c. Sampling without replacement decreases the accuracy of the sample mean.
   d. Sampling without replacement is like sampling from an infinite population.
   e. None of the above statements is true.

10. If a sample of 25 items is drawn without replacement from a uniform population of 50 items,
    a. the sample mean will equal the sample median.
    b. the sample beam and fulcrum will contain at least 6 standard deviations.
    c. the sample box plot will have very long whiskers.
    d. the sample histogram should be bell-shaped.
    e. the sample mean is more accurate than if the sample was drawn with replacement.

11. Which of the following is *not* true of an outlier?
    a. An outlier should be discarded since it is not part of the population being sampled.
    b. An outlier is more than 3 standard deviations from the mean.
    c. An outlier can be generated by a different population.
    d. An outlier can be generated by the population being sampled.
    e. All of the above are true of outliers.

12. Which of the following distributions can be either right-skewed or left-skewed?
    a. Normal distribution.
    b. Chi-square distribution.
    c. Student's t distribution.
    d. Triangular distribution.
    e. Exponential distribution.

Glossary of Terms

Beam and fulcrum Display that plots the position of the sample mean (the "fulcrum") and the standard deviation points (Mean ±1 SD, Mean ± 2 SD, Mean ± 3 SD, etc.). This display reveals skewness (the longer tail will indicate the direction of skewness) and kurtosis (the more standard deviations displayed along the beam, the more peaked the data).

Bimodal Population whose probability distribution has two peaks separated by a valley. In reference to a sample histogram, it would refer to two intervals that have higher frequencies than their adjacent intervals. See **Unimodal**.

Box plot Five-number graphical display plotting the positions of the minimum, quartiles (first, second, third), and maximum along a scale representing data values. The box encloses the quartiles and the span of the whiskers indicates the range.

Centrality General reference to measures of the middle of a distribution (for example: mean, median, mode, midrange).

Dispersion General reference to measures of "spread" of data values around the center of a distribution (for example: standard deviation, range).

Finite population Population with N elements.

Infinite population Population whose elements are uncountably large.

Kurtosis Measure of relative peakedness. If a distribution is unimodal and symmetric, then $K = 3$ indicates a normal bell-shaped distribution (mesokurtic); $K < 3$ indicates a platykurtic distribution (flatter than normal with shorter tails); and $K > 3$ indicates a leptokurtic distribution (more peaked than normal with longer tails).

Mean For a population, the mean is the expected value of X, denoted μ. It may be thought of as the probability-weighted average of the X values and may be interpreted as the fulcrum (balancing point) of the distribution along the X-axis. For a sample, the mean (denoted $\overline{X}$) is the sum of the sample items divided by the sample size.

Outlier Any sample observation that differs from the mean by 3 or more standard deviations.

Peakedness See **Kurtosis**.

Population Any collection of data values that are being sampled.

Population distribution The probability density function f(x) defined over the a range of values of a random variable. The ordinate shows the probability associated with each X value. The area under the entire distribution is 1.

Random sample A sample selected from a population by a method that ensures that every population element has the same chance of being chosen. Simple random sampling may be done using tables or computer-generated random numbers.

Random variation Sample items do not consistently lead to a perfect representation of a population. This variation is reduced if the sample size is increased. Such variation is expected, and can be quantified using the rules of statistics.

Replacement Sampling method used with a finite population in which an item is sampled and then returned to the population possibly to be sampled again. Sampling with replacement keeps the probability of drawing each item in the population unchanged. If sampling is done without replacement, the selected item is not returned to the population so the probability of selecting any of the remaining items from the population is changed.

Sample Set of observations taken from a population (usually, but not necessarily, by random sampling). See **Replacement**.

Sample size Number of items in a sample, usually denoted n.

Skewed population A population is skewed right and has a long right tail if its mean exceeds its median (and conversely if the population is skewed left).

Skewness Measure of relative symmetry. Zero indicates symmetry. Positive values show a long right tail. Negative values show a long left tail.

Standard deviation Denoted s (for a sample) or σ (for a population), it is the square root of the variance. It is a measure of dispersion about the mean. The larger the standard deviation, the greater the dispersion. See **Variance**.

Statistical inference Generalization about a population, based on a sample that has been drawn from the population. Often associated with a probability or confidence that the inference is correct.

Symmetry See **Skewness**.

Unimodal Population whose probability density function has one peak. In reference to a sample histogram, it would refer to an interval (bin) with higher frequencies than other intervals. See **Bimodal**.

Variance In a population, the variance is the expected value of $(X - \mu)^2$ and is denoted σ^2. In a sample, the variance is the sum of the squared deviations about the sample mean divided by $n - 1$ and is denoted s^2. It is a measure of dispersion about the mean.

Solutions to Self-Evaluation Quiz

1. b Do Exercises 1–7, and 8–11. Read the Illustration of Concepts.
2. c Do Exercises 2–6. Read the Illustration of Concepts.
3. c Do Exercises 2–5. Read the Illustration of Concepts.
4. d Do Exercise 8. Review Chapter 5.
5. e Do Exercises 1–7. Review Chapter 2.
6. d Do Exercise 1. Do Team Learning Project 1. Review Chapter 1.
7. d Do Exercises 8–11. Review Chapter 2.
8. e Do Exercises 10–11. Review Chapters 2 and 3.
9. b Do Exercises 12–15. Read the Overview of Concepts.
10. e Do Exercises 12–15. Read the Overview of Concepts.
11. a Do Exercises 16–20. Read the Overview of Concepts. Review Chapter 1.
12. d Do Exercise 21.