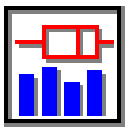
Solutions to Worktext Exercises



Chapter 1

Visualizing Univariate Data Analysis

Basic Learning Exercises

1. a) The variable is the average price of homes in smaller cities. b) The units are dollars. c) Data was probably collected by going to a national real estate company. d) It is discrete since there are a countable number of different prices. However, the range of possible values is so large that it behaves like continuous data.

2. Tulsa, OK –1.191 Santa Barbara, CA 4.125

3. An observation that is more than 3 standard deviations from the mean is an outlier. Yes, Santa Barbara is an outlier because its standardized value is greater than 3.

4. Mean 109,630 Median 100,988 Midrange 153,289 Midhinge 105,834
5% Trimmed Mean 107,080 The measures are quite different suggesting skewness. Because the mean and midrange exceed the median and the trimmed mean is smaller than the mean, positive skewness is suggested

5. Standard deviation 29,757 Range 158,199 First quartile 87,956
Third quartile 123,712 Interquartile range 35,756
Standard deviation, range and interquartile range will increase with increased dispersion, while Q1 and Q3 are used to calculate the interquartile range.

6. A dot plot gives a better representation of the data, but if you are creating both displays by hand a stem and leaf is a much easier display to create.

7. The median (centrality) is the line within the box. The distance between the two whiskers (range) and the length of the box (interquartile range) illustrate dispersion.

8. All three displays show a positively skewed distribution with an outlier on the high side. A frequency histogram shows the number of occurrences in each interval. A relative frequency histogram shows the proportion of the sample within each interval. A frequency polygon is generally used if there are a large number of intervals.

9. A cumulative histogram shows the number of occurrences less than the interval's upper limit. A standardized Z value histogram is used to compare two very different data sets or as a visual check for outliers.

10. a) The appearance changes because the number of observations in a class can change a lot with a change in the histogram's range or its class width. b) The cumulative changes less dramatically because it's the summation from left to right. This hides or masks many differences. c) The standardized histogram does not change because its classes are set at 1, 2, 3, and 4 standard deviations from the mean.

Intermediate Learning Exercises

11. The mean (centrality) is the fulcrum. The distance between any two vertical lines (standard deviation) and the length of the beam (range) illustrate dispersion.

12. Skewness 1.77 Kurtosis 7.431
    % within 1 SD 77 % within 2 SD 98 % within 3 SD 98
    The data is positively skewed and relatively peaked compared to a normal distribution. The percents within 1, 2, and 3 SD confirm that the data is not normally distributed.

13. In the box plot skewness is seen by comparing the lengths of the two whiskers (a longer right whisker shows right or positive skewness) and seeing where the median lies within the box (if the median is to the left of the box's center there is right or positive skewness). In the beam and fulcrum skewness is seen by comparing the length of the beam on either side of the fulcrum (a longer beam on the right side shows right or positive skewness).

14. The box plot illustrates peakedness by the ratio of the length of the box (interquartile range) to the length of the whiskers (range). For symmetric mesokurtic data (kurtosis = 3), it is about 0.25, for leptokurtic data (kurtosis > 3) it is usually less than 0.20, and for platykurtic data (kurtosis < 3) it is generally greater than 0.33. The beam and fulcrum illustrates peakedness by the number of standard deviations needed to cover the range. For mesokurtic symmetric data (kurtosis = 3) it is about 5, for leptokurtic data (kurtosis > 3) it is generally over 6, and for platykurtic data (kurtosis < 3) it is usually under 4.5.

15. The quantile plot has the lazy “S” shape that is characteristic of data having a bell-shape distribution. The quantile plot is almost a 45-degree line that is characteristic of data having a rectangular distribution.

16. The runs chart displays randomness when the plotted data shows no particular pattern. In both data sets, there is a pattern. In the first case, early observations have higher ability scores. In the second case, Popes served longer terms in later centuries.

17. When nice limits is used the second interval is much taller than the first interval. When cover range is used, they are almost equal in height. This happens because with nice limits the first class is from 60,000 to 80,000. However, since the lowest house price is 74,189, there are few observations in this first interval.

18. The appearance changes less when cover range is used because with cover range only the interval width is changing whereas with nice limits both the interval width and the starting and ending points for the histogram change.

19. To do EDA, covering the range may be better because fewer factors influence the data’s appearance. However, to display the results to others, nice limits are more pleasing.

Advanced Learning Exercises

20. The answer should discuss the skewness of the histogram, the existence of outliers, its modal classes (why there is only one or why there are multiple), and perhaps its peakedness.

21. There is no incorrect answer to any serious attempt to improve upon the histogram. The tradeoffs are between the niceness of the labels, fairness in representing the data, and the amount of wasted space in the leading and ending intervals. Often there is a particular number of classes that provides a nice set of labels. The problem is that this number of classes usually is not the number that should be used to display the data.

22. Sturges' Rule says to add a class every time you double the sample size. It is only a beginning. It is easy to create rules for nice classes, but it is hard to reconcile problems when combining them with the need to adequately represent the data. For example, the "right" number of classes may make it difficult to choose nice limits, which do not waste too much space in the two end classes. Rules might be devised to set limits on such waste (for example, don't waste more than half an interval on either end). In devising a rule, the student may refer to "round" intervals (divisible by a multiple of 2, 5, or 10 multiplied by a power of 10) without defining "round." "Divisible by 5 or 10" is also unclear if data values are measured in small units (e.g., interest rates) or large units (e.g., annual income).

23. The student should discuss the symmetry and peakedness of the data. Skewness is a measure of relative symmetry and is based on the distribution's third sample moment. Kurtosis is a measure of relative peakedness and is based on the fourth sample moment. The skewness p-value is the probability that the analyzed data is from a symmetric population. The kurtosis p-value is the probability that the data is from a population with the peakedness of a normal or bell-shaped distribution. If the data is not from a symmetric population (small skewness p-value), the kurtosis p-value is invalid.

24. The coefficient of variation is the ratio of the sample standard deviation to the sample mean. It scales the standard deviation. Statisticians use it when comparing the relative variation of different data sets. The 5% trimmed mean is the sample mean after the 5% highest and lowest observations are removed. It is robust against large or small data values. Since the sample mean is not robust against such values, if the two statistics are similar then a statistician knows that the sample mean is unaffected by these values.

25. Data measured over time and data measuring different groups are likely to suffer from nonrandomness. This occurs in data measured over time because over time things change; especially over a long time span. For example, the Years Served by Popes covers almost 500 years. During this time, medical knowledge and the general health of society changed dramatically. Nonrandomness occurs in data for different groups since there is a tendency to order the data by these groups. For example, in the Importance of Job Abilities, the researcher had inadvertently arranged the important job abilities first.

25. The beam and fulcrum and the box plot give a picture of the shape of the distribution based on descriptive statistics. These displays do not show the actual data; this must be visualized by the user. In contrast, the stem and leaf and the dot plot give a picture of the actual data. These displays do not show the statistics; these must be visualized by the user. The two types of displays complement rather than substitute for one another.