Appendix S

SECOND CANADIAN EDITION

Statistics
A First Course

Donald H. Sanders
Educational Consultant
Fort Worth, Texas

Robert K. Smidt
California Polytechnic State University
San Luis Obispo, California

Aminmohamed Adatia
University of Regina
Regina, Saskatchewan

Glenn A. Larson
University of Regina
Regina, Saskatchewan

McGraw-Hill Ryerson

Toronto Montréal Boston Burr Ridge, IL Dubuque, IA Madison, WI New York San Francisco
St. Louis Bangkok Bogotá Caracas Kuala Lumpur Lisbon London Madrid
Mexico City Milan New Delhi Santiago Seoul Singapore Sydney Taipei

Example 3.16 AIDS in the Americas In Example 3.12 (page 65), we found that the mean rate of incidence of AIDS per 100,000 in 1995 for the Americas was 13.0. Obtain the standard deviation of this data set.

◆ **Solution:**

Two possible ways of answering this question are to use either formula 3.5 or 3.6. But, examining Figures 3.15a and 3.15b, we find that MINITAB and the TI-83 calculator, in addition to calculating the measures of central tendency that we have previously considered, also compute several other items, including the standard deviation, denoted as "StDev" by MINITAB and as "Sx" by the TI-83 calculator. Both agree that for this data, $s = 23.92$. ◆

Interpreting the Standard Deviation

We know that dispersion is the amount of spread or scatter that occurs in a data set. If, for example, the values in the set are clustered tightly about their mean, the measured dispersion—in this case the standard deviation—is small. But if we have other data sets where the values become more and more scattered about their means, the standard deviations for those sets become larger and larger. To summarize, then, if a standard deviation is small, the items in the data set are bunched about their mean, and if the standard deviation is large, the data items are widely dispersed about their mean. To drive home this generalization in a more tangible way, let's first consider *Chebyshev's theorem*.

Russian mathematician P. P. Chebyshev has been dead for a century now, but his theorem lives on.

Chebyshev's Theorem

Chebyshev's theorem states that the proportion of *any* data set that lies within k standard deviations of the mean (where k is any positive number greater than or equal to 1) is *at least* $1 - (1/k^2)$.

Thus, if we substitute 2 for k in the theorem, we get $1 - (1/k^2) = 1 - (1/2^2) = 1 - (1/4) = 3/4$, or, in percentage form, $3/4 \times 100 = 75$ percent. This result means that *at least* 75 percent of the items in *any* data set (no matter how skewed it is) must lie within two standard deviations of the mean. And at least 88.9 percent $[1 - (1/3^2)$ or $8/9]$ of the items in *any* data set must fall within three standard deviations of the mean.

Chebyshev's theorem shows us how the standard deviation is related to the scatter of data items. But it tells us only the minimum percentage of items that must fall within given intervals in any data set. We've seen earlier (and in Figure 3.8 on page 53) though that many data sets have values that are found to be distributed or scattered about their means in reasonably symmetrical ways.

For bell-shaped distributions, known as normal distributions, the *empirical rule* applies and is of greater significance than Chebyshev's theorem.

STATISTICS IN ACTION

Government Statistical Engines

Many government departments collect descriptive statistics that have multiple uses. Every five years, Statistics Canada tries to determine how many people there are in the provinces and in the country. These figures, along with facts about births and deaths that are generated elsewhere to advance public health, are raw materials used by demographers—the scientists who analyze the characteristics of human populations. Similarly, government-generated international trade data and weather facts support the development of new insights in economics and atmospheric sciences.

Empirical Rule

The empirical rule for distributions that are generally bell-shaped or normal is that

- About 68 percent of all data items lie within one standard deviation of the mean ($\mu \pm 1\sigma$ or $\bar{x} \pm 1s$).
- About 95 percent of all data items lie within two standard deviations of the mean ($\mu \pm 2\sigma$ or $\bar{x} \pm 2s$).
- About 99.7 percent of all data items lie within three standard deviations of the mean ($\mu \pm 3\sigma$ or $\bar{x} \pm 3s$).

Let's look at an example of an application of this empirical rule. Suppose that many people are given a new type of IQ test, and the resulting raw scores are organized into a frequency distribution. A frequency polygon is prepared from the distribution and is found to be symmetrical in shape. The arithmetic mean of this mound-shaped distribution is 100 points, and the standard deviation is ten points. In this situation, the mean IQ score is directly under the peak of the curve, and the following relationships exist: (1) about 68 percent of the test scores fall within one standard deviation of the mean—that is, about 68 percent of the people have test scores between 90 and 110 points; (2) about 95 percent of the test scores fall within two standard deviations of the mean—that is, about 95 percent of those taking the test have scores between 80 and 120 points; and (3) virtually all (99.7 percent) of the test scores fall within three standard deviations of the mean (scores between 70 and 130). Figure 3.21 shows these relationships.

The relationships that exist between the mean and the standard deviation in a bell-shaped distribution may also be used for analysis purposes with distributions that are only approximately symmetrical. Let's return to our Slimline Fizzy Cola example and interpret the meaning of the standard deviation of 14.78 litres, since that distribution is approximately normal. We can conclude that about the middle two-thirds of the 30 employees sold syrup quantities between $\bar{x} \pm 1s$, that is, between 115.20 litres ± 14.78

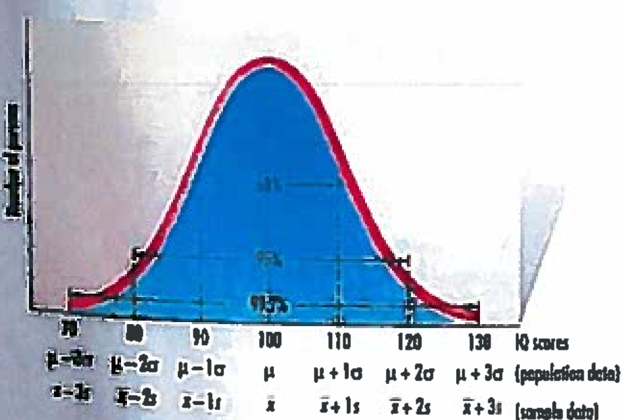

FIGURE 3.21 Illustration of the empirical rule