Objectives

- Calculate proportions
- To calculate the coefficient of variation and compare variations
- To calculate the mean, variance, and standard deviation of grouped data
- To use the empirical rule and Chebyshev's Theorem to describe data variability

Oct 9-3:52 PM

Applying the Standard Deviation

Although the standard deviation is not an intuitive concept, knowing the mean and standard deviation of a data set provides a great deal of information about the data. If the histogram of the measurements is bell-shaped, the empirical rule describes the variability of a set of measurements. Chebyshev's Theorem is a more general rule describing the variability of any set of data regardless of the shape of its distribution.

Oct 9-3:52 PM

Empirical Rule

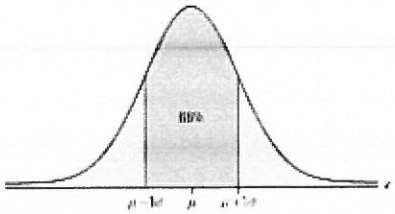


Figure 4.14 – One Sigma

One sigma rule: If the distribution of the data is bell-shaped, about 68 % of the data should lie within one standard deviation of the mean.

A deviation of more than one sigma from the mean is to be expected about once in every three observations.

Oct 9-3:53 PM

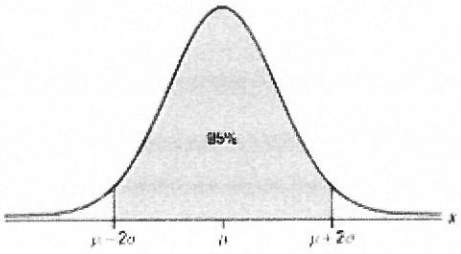


Figure 4.15 – Two Sigma

Two sigma rule: If the distribution of data is bell-shaped, about 95 % of the data should lie within two standard deviations of the mean.

A deviation of more than two sigma from the mean is to be expected about once in every twenty observations.

Oct 9-6:13 PM

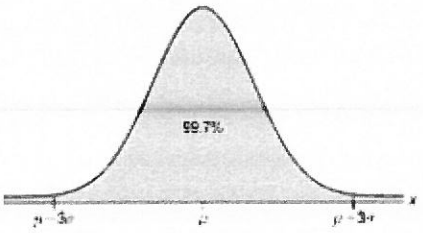


Figure 4.16 – Three Sigma

Three sigma rule: If the distribution of the data is bell-shaped, about 99.7 % of the observations should be within three standard deviations of the mean.

A deviation of more than three sigma from the mean is to be expected about once in every 333 observations, slightly less than 0.3 % of the time.

Oct 9-6:14 PM

Who is King of the Hill?

In 1961, Wilt Chamberlain was the National Basketball Association (NBA) rebounding leader with 27 rebounds per game. In 1992, the colorful Dennis Rodman won the same honor with 18.7 rebounds per game. Common sense suggests that professional basketball in the 1990s was played at a much higher level than in the 1960s. So why have the rebounding leader's number fallen? Is it another case of "less is more"?

Researchers investigating this interesting puzzle considered two other variables: the number of rebounding opportunities (this had gone down since the field goal percentage has increased historically) and the average number of minutes played per game, which has also fallen.

Thus, when we adjust the actual rebounds obtained by the rebounding leaders to the number of minutes played and the total number of rebounding opportunities, we see a completely different picture. The adjusted rebound numbers for Chamberlain and Rodman are 35.42and 51.06 respectively.

Oct 9-6:14 PM

Example 4.23

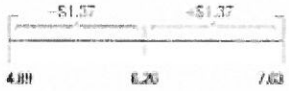
Suppose a group of high technology stocks has an average earnings per share of $6.26, with a standard deviation of $1.37. If the data possess a bell-shaped distribution, which interval contains 68% of the earnings? Which interval contains 95% of the earnings?

average = 6.26
std dev = 1.37

Solution

Using the one sigma rule, we will capture 68% of the observations.

$\mu \pm 1\sigma = \$6.26 \pm \$1.37$



Using the one sigma rule results in an interval from $6.26−$1.37 to $6.26+$1.37. Doing the arithmetic produces an interval from $4.89 to $7.63.

To capture 95% of the earnings, use the two sigma rule, $6.26 ± 2·$1.37. Doing the arithmetic results in an interval from $3.52 to $9.00.

Note that to increase the percentage of data captured from 68% to 95% requires an interval that is twice as large.

Oct 9-6:15 PM

Chebyshev's Theorem

It is important to remember that the empirical rule applies only to bell-shaped distributions. For any distribution, regardless of shape, Chebyshev's Theorem may be used, although its results are much more approximate.

Theorem: Chebyshev's Theorem

The proportion of any data set lying within k standard deviations of the mean is at least

$$1-\frac{1}{k^2}, \text{ for } k>1.$$

Oct 9-6:17 PM

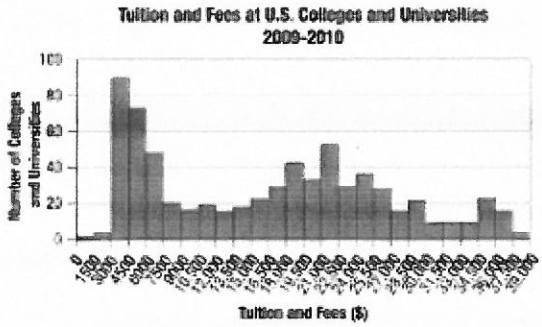
For example, if $k=2$ at least $1 - \frac{1}{2^2} = \frac{3}{4}$ (or 75%) of the data values lie within 2 standard deviations of the mean, for any data set. Similarly, if $k=3$ at least $1 - \frac{1}{3^2} = \frac{8}{9}$ (or approximately 88.9%) of the data values lie within 3 standard deviations of the mean, for any data set.

Also note that k does not have to be an integer value. If $k=1.5$, at least $1 - \frac{1}{1.5^2} = \frac{5}{9}$ (or approximately 55.6%) of the data values will lie within 1.5 standard deviations of the mean, for any data set.

Oct 9-6:19 PM

Example 4.24

The tuition and fees of colleges and universities distribution histogram for the United States in 2009–2010 is shown in the figure below. The mean of the data is \$16,442, while the standard deviation is \$10,014. What can we conclude from Chebyshev's Theorem using $k=2$?



mean = \$16,442 std dev = \$10,014 k=2

Oct 9-6:20 PM

mean = \$16,442 std dev = \$10,014 k=2

$$\mu+2\sigma=16422+2(10014)=\$36,450$$

$$\mu-2\sigma=16422-2(10014)=-\$3606.$$

Therefore, by Chebyshev's Theorem, we can say that at least 75% of the tuition and fees of colleges and universities in the United States are between \$0 and \$36,450.

Oct 9-6:22 PM

Data Subsetting

Looking at the tuition data presented in Example 4.24 with a histogram using fewer intervals gives a slightly different picture. Figure 4.17 provides some idea about location and dispersion of the data, but nothing specific. Histograms are outstanding at defining the shape of the data.

Oct 9-6:25 PM

Where is the Central Value of the Tuition Data?

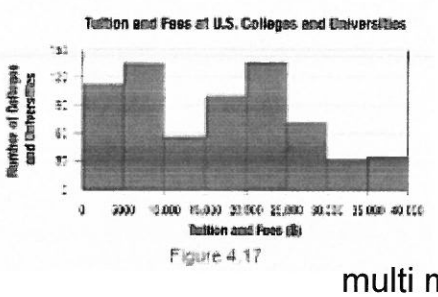


Figure 4.17

- no mode
- median is $17,468
- mean is $16,442
- trimmed mean $16,087

multi modal - not symmetric?

mean & trimmed mean close - no outliers?

Can justify either - just be specific and provide reasons

Oct 9-6:26 PM

Table 4.16 - Tuition Data

Measure of Location	Value
Mean	$16,442
Median	$17,468
10% Trimmed Mean	$16,087

Table 4.17 - Tuition Data Percentiles

Percentile	Value
100^{th} (Max)	$38,140
95^{th}	$35,015
90^{th}	$29,982
75^{th} (Q_3)	$23,800
50^{th} (Q_2)	$17,468
25^{th} (Q_1)	$6049
10^{th}	$4166
5^{th}	$3558
0 (Min)	$790

Oct 9-6:35 PM

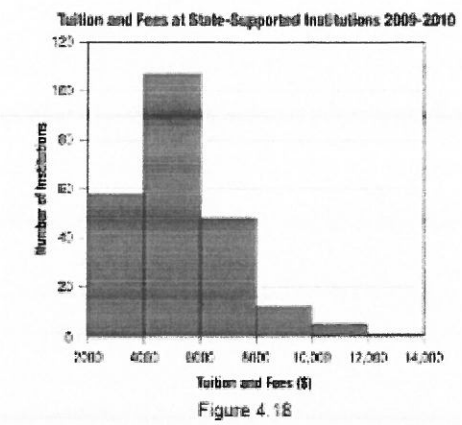


Figure 4.18

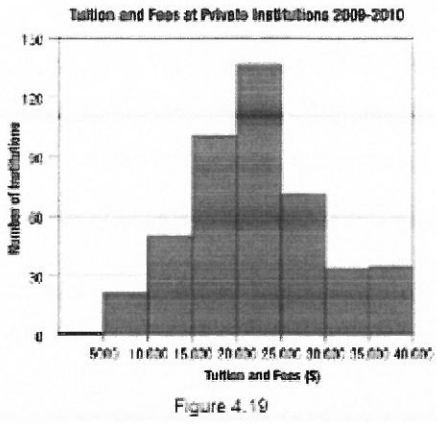


Figure 4.19

The data subsetting suggests that the data clustering revealed in the histogram in Figure 4.17 is the result of merging data from different kinds of educational institutions.

Oct 9-7:07 PM

Formula: Coefficient of Variation

The **coefficient of variation**, another statistical measure, compares the variation in data sets.

For population data, the coefficient of variation is defined as

$$CV = \left(\frac{\sigma}{\mu} \cdot 100\right)\%,$$

and for sample data,

$$CV = \left(\frac{s}{\bar{x}} \cdot 100\right)\%.$$

When comparing the variation of data sets, many times the units of measure will be different. The coefficient of variation standardizes the variation measure by dividing it by the mean. The division has one interesting side effect: the unit of measure is removed from the statistic.

Oct 9-7:09 PM

$$CV_{\text{private}} = \frac{\$7376}{\$22163} \cdot 100\% \approx 33.28\%$$

$$CV_{\text{public}} = \frac{\$1864}{\$5396} \cdot 100\% \approx 34.54\%$$

Oct 9-7:15 PM

Formula: Mean of Grouped Data

The population **mean of grouped data** is given by

$$\mu = \frac{\sum(f_i M_i)}{N}$$

where

f_i = number of observations in the i^{th} class,

N = the total number of observations in all classes, $N = \sum f_i$, and

M_i = midpoint of the i^{th} class.

The sample mean of grouped data is given by

$$\bar{x} = \frac{\sum(f_i M_i)}{n},$$

where n is the number of observations in the sample.

Oct 9-7:16 PM

Proportions

The **proportion** is one of the more common summary measures.

Definition

A **proportion** measures the fraction of a group that possesses some characteristic.

To calculate a proportion, simply count the number in the group that possess the characteristic and divide the count by the total number in the group. Let

X = number of observations that possess the characteristic,

N = number of observations in the population, and

n = number of observations in the sample, then

$p = \frac{X}{N}$ = the population proportion, and

$\hat{p} = \frac{X}{n}$ = the sample proportion.

The symbol $\hat{p}$ is pronounced *p-hat*.

Oct 9-7:16 PM

Example 4.26

Suppose your statistics class is composed of 46 students of which 4 are left-handed. What proportion of the class is left-handed?

Solution

There are 46 pieces of data in the class. Think of data as composed of 0s and 1s. Any left-handed person will be a 1, and any right-handed person will be a 0. In our data set, there will be four 1s and forty-three 0s.

Assigning $x = \begin{cases} 1 & \text{if person is left-handed} \\ 0 & \text{if person is right-handed} \end{cases}$

then $\sum x = 1+1+1+1+0+0+\cdots+0 = 4$.

In the notation we used earlier, X equals the number of observations that possess the characteristic. Therefore,

$X = \sum x = 4$ and $p = \frac{X}{N} = \frac{4}{46} = 0.0869$

Thus, approximately $0.0869(100) = 8.69\%$ of people in the class are left-handed.

Note that we are using p, the population proportion, in this case because we are considering the population of students in your statistics class. If we were using the data from your statistics class to estimate the proportion of all statistics students that are left-handed, we would use the symbol $\hat{p}$, the sample proportion, since we would then be estimating the proportion from sample data.

Oct 9-7:19 PM

This chapter has been devoted to summarizing data. Yet, with the exception of the mode, none of the summary methods discussed should be applied to nominal data. Using proportions is one of the few summary methods available for analyzing qualitative data.

Oct 9-7:20 PM