Chapter 3
Describing, Exploring, and Comparing Data

Slide 1

A) Measures of Center

B) Measures of Variation

C) Measures of Relative Standing

D) Exploratory Data Analysis

Definition

Slide 2

❖ **Measure of Center**

The value at the center or middle of a data set

Definition

Slide 3

Arithmetic Mean

(Mean)

the measure of center obtained by adding the values and dividing the total by the number of values

Notation

Slide 4

- Σ denotes the addition of a set of values
- x is the variable usually used to represent the individual data values
- n represents the number of values in a sample
- N represents the number of values in a population

Notation

Slide 5

$\overline{x}$ is pronounced 'x-bar' and denotes the mean of a set of sample values

$$\overline{x} = \frac{\Sigma x}{n}$$

μ is pronounced 'mu' and denotes the mean of all values in a population

$$\mu = \frac{\Sigma x}{N}$$

Definitions

Slide 6

❖ **Median**

the middle value when the original data values are arranged in order of increasing (or decreasing) magnitude

❖ **often denoted by $\tilde{x}$ (pronounced 'x-tilde')**

❖ **is not affected by an extreme value**

Finding the Median

Slide 7

- If the number of values is odd, the median is the number located in the exact middle of the list
- If the number of values is even, the median is found by computing the mean of the two middle numbers

Slide 8

5.40 1.10 0.42 0.73 0.48 1.10

0.42 0.48 0.73 1.10 1.10 5.40

(even number of values – no exact middle shared by two numbers)

$\frac{0.73 + 1.10}{2}$ MEDIAN is 0.915

5.40 1.10 0.42 0.73 0.48 1.10 0.66

0.42 0.48 0.66 0.73 1.10 1.10 5.40

(in order - odd number of values)

exact middle MEDIAN is 0.73

Definitions

Slide 9

- Mode

the value that occurs most frequently

The mode is not always unique. A data set may be:

Bimodal

Multimodal

No Mode

- denoted by M

the only measure of central tendency that can be used with nominal data

Examples

Slide 10

a. 5.40 1.10 0.42 0.73 0.48 1.10 ⇦ Mode is 1.10

b. 27 27 27 55 55 55 88 88 99 ⇦ Bimodal - 27 & 55

c. 1 2 3 6 7 8 9 10 ⇦ No Mode

Definitions

Slide 11

- Midrange

the value midway between the highest and lowest values in the original data set

$$\text{Midrange} = \frac{\text{highest score + lowest score}}{2}$$

Round-off Rule for Measures of Center

Slide 12

Carry one more decimal place than is present in the original set of values

Mean from a Frequency Distribution

Slide 13

Assume that in each class, all sample values are equal to the class midpoint

Mean from a Frequency Distribution

Slide 14

use class midpoint of classes for variable *x*

$$\bar{x} = \frac{\Sigma (f \cdot x)}{\Sigma f}$$ Formula 2-2

x = class midpoint

f = frequency

$\Sigma f = n$

Weighted Mean

Slide 15

In some cases, values vary in their degree of importance, so they are weighted accordingly

$$\bar{x} = \frac{\Sigma (w \cdot x)}{\Sigma w}$$

Best Measure of Center

Slide 16

Table 2-10 Comparison of Mean, Median, Mode, and Midrange

Measure of Center	Definition	How Common?	Existence	Takes Every Value into Account?	Affected by Extreme Values?	Advantages and Disadvantages
Mean	$\bar{x} = \frac{\Sigma x}{n}$	most familiar "average"	always exists	yes	yes	used throughout this book; works well with many statistical methods
Median	middle value	commonly used	always exists	no	no	often a good choice if there are some extreme values
Mode	most frequent data value	sometimes used	might not exist; may be more than one mode	no	no	appropriate for data at the nominal level
Midrange	$\frac{\text{high} + \text{low}}{2}$	rarely used	always exists	no	yes	very sensitive to extreme values

General comments:
- For a data collection that is approximately symmetric with one mode, the mean, median, mode, and midrange tend to be about the same.
- For a data collection that is obviously asymmetric, it would be good to report both the mean and median.
- The mean is relatively *reliable*. That is, when samples are drawn from the same population, the sample means tend to be more consistent than the other measures of center (consistent in the sense that the means of samples drawn from the same population don't vary as much as the other measures of center).

Definitions

Slide 17

- **Symmetric**

  Data is symmetric if the left half of its histogram is roughly a mirror image of its right half.

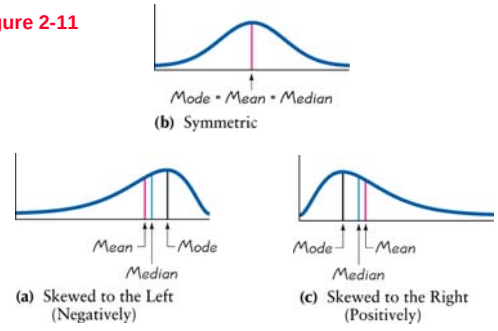
- **Skewed**

  Data is skewed if it is not symmetric and if it extends more to one side than the other.

Skewness

Slide 18

Figure 2-11



Measures of Variation

Because this section introduces the concept of variation, this is one of the most important sections in the entire book

Definition

The range of a set of data is the difference between the highest value and the lowest value

highest value – lowest value

Definition

The standard deviation of a set of sample values is a measure of variation of values about the mean

Sample Standard Deviation Formula

$$s = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n - 1}}$$

Formula 2-4

Sample Standard Deviation (Shortcut Formula)

$$s = \sqrt{\frac{n(\Sigma x^2) - (\Sigma x)^2}{n(n - 1)}}$$

Formula 2-5

Standard Deviation - Key Points

- The standard deviation is a measure of variation of all values from the mean
- The value of the standard deviation s is usually positive
- The value of the standard deviation s can increase dramatically with the inclusion of one or more outliers (data values far away from all others)
- The units of the standard deviation s are the same as the units of the original data values

Population Standard Deviation

Slide 25

$$\sigma = \sqrt{\frac{\Sigma (x - \mu)^2}{N}}$$

This formula is similar to Formula 2-4, but instead the population mean and population size are used

Definition

Slide 26

- The variance of a set of values is a measure of variation equal to the square of the standard deviation.
- Sample variance: Square of the sample standard deviation s
- Population variance: Square of the population standard deviation σ

Variance - Notation

Slide 27

standard deviation squared

Notation
- s^2 Sample variance
- σ^2 Population variance

Round-off Rule for Measures of Variation

Slide 28

Carry one more decimal place than is present in the original set of data.

Round only the final answer, not values in the middle of a calculation.

Definition

Slide 29

The coefficient of variation (or CV) for a set of sample or population data, expressed as a percent, describes the standard deviation relative to the mean

Sample

$$CV = \frac{s}{\bar{x}} \cdot 100\%$$

Population

$$CV = \frac{\sigma}{\mu} \cdot 100\%$$

Standard Deviation from a Frequency Distribution

Slide 30

Formula 2-6

$$s = \sqrt{\frac{n[\Sigma(f \cdot x^2)] - [\Sigma(f \cdot x)]^2}{n(n-1)}}$$

Use the class midpoints as the x values

Estimation of Standard Deviation Range Rule of Thumb

Slide 31

For estimating a value of the standard deviation s,

Use

$$s \approx \frac{\text{Range}}{4}$$

Where range = (highest value) – (lowest value)

Estimation of Standard Deviation Range Rule of Thumb

Slide 32

For interpreting a known value of the standard deviation s, find rough estimates of the minimum and maximum "usual" values by using:

Minimum "usual" value ≈ (mean) – 2 X (standard deviation)

Maximum "usual" value ≈ (mean) + 2 X (standard deviation)

Definition

Slide 33

Empirical (68-95-99.7) Rule

For data sets having a distribution that is approximately bell shaped, the following properties apply:

- **About 68% of all values fall within 1 standard deviation of the mean**
- **About 95% of all values fall within 2 standard deviations of the mean**
- **About 99.7% of all values fall within 3 standard deviations of the mean**

The Empirical Rule

Slide 34

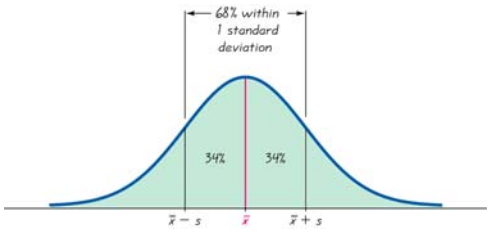


FIGURE 2-13

The Empirical Rule

Slide 35

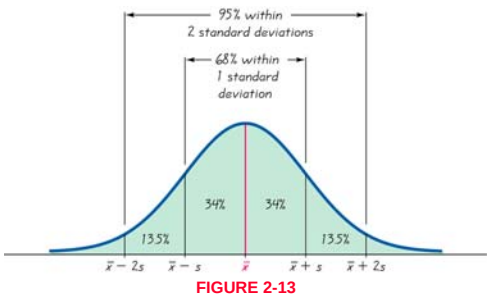


FIGURE 2-13

The Empirical Rule

Slide 36

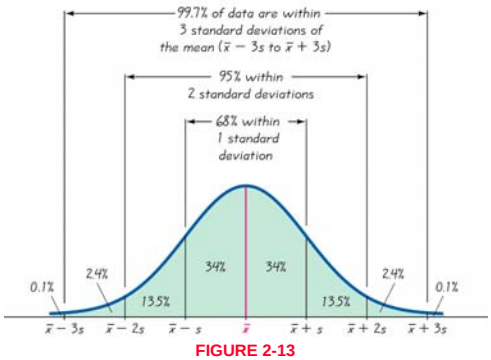


FIGURE 2-13

Definition

Slide 37

Chebyshev's Theorem

The proportion (or fraction) of any set of data lying within K standard deviations of the mean is always at least $1-1/K^2$, where K is any positive number greater than 1.

- **For K = 2, at least 3/4 (or 75%) of all values lie within 2 standard deviations of the mean**
- **For K = 3, at least 8/9 (or 89%) of all values lie within 3 standard deviations of the mean**

Rationale for Formula 2-4

Slide 38

The end of Section 2- 5 has a detailed explanation of why Formula 2- 4 is employed instead of other possibilities and, specifically, why n – 1 rather than n is used. The student should study it carefully

Definition

Slide 39

- ***z* Score (or standard score)**

  the number of standard deviations that a given value *x* is above or below the mean.

Measures of Position z score

Slide 40

Sample **Population**

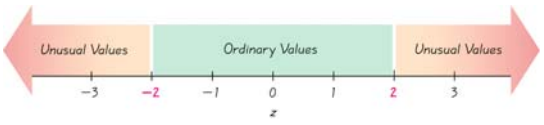
$$z = \frac{x - \bar{x}}{s} \qquad z = \frac{x - \mu}{\sigma}$$

Round to 2 decimal places

Interpreting Z Scores

Slide 41

FIGURE 2-14



Whenever a value is less than the mean, its corresponding z score is negative

Ordinary values: ***z* score between –2 and 2 sd**

Unusual Values: ***z* score < -2 or *z* score > 2 sd**

Definition

Slide 42

- **Q_1 (First Quartile) separates the bottom 25% of sorted values from the top 75%.**
- **Q_2 (Second Quartile) same as the median; separates the bottom 50% of sorted values from the top 50%.**
- **Q_1 (Third Quartile) separates the bottom 75% of sorted values from the top 25%.**

Quartiles

Q_1, Q_2, Q_3

divides ranked scores into four equal parts

25% 25% 25% 25%

(minimum) Q_1 Q_2 Q_3 (maximum)

(median)

Percentiles

Just as there are quartiles separating data into four parts, there are 99 percentiles denoted P_1, P_2, . . . P_{99}, which partition the data into 100 groups.

Finding the Percentile of a Given Score

$$\text{Percentile of value } x = \frac{\text{number of values less than } x}{\text{total number of values}} \cdot 100$$

Converting from the *k*th Percentile to the Corresponding Data Value

Notation

$$L = \frac{k}{100} \cdot n$$

- n total number of values in the data set
- k percentile being used
- L locator that gives the *position* of a value
- P_k *k*th percentile

Converting from the *k*th Percentile to the Corresponding Data Value

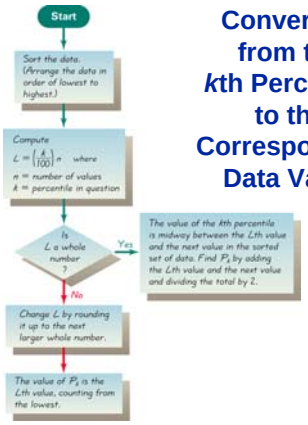


Figure 2-15

Some Other Statistics

- **Interquartile Range (or IQR):** $Q_3 - Q_1$
- **Semi-interquartile Range:** $\frac{Q_3 - Q_1}{2}$
- **Midquartile:** $\frac{Q_3 + Q_1}{2}$
- **10 - 90 Percentile Range:** $P_{90} - P_{10}$

Definition

Slide 49

- **Exploratory Data Analysis** is the process of using statistical tools (such as graphs, measures of center, and measures of variation) to investigate data sets in order to understand their important characteristics

Definition

Slide 50

- An **outlier** is a value that is located very far away from almost all the other values

Important Principles

Slide 51

- An outlier can have a dramatic effect on the mean
- An outlier have a dramatic effect on the standard deviation
- An outlier can have a dramatic effect on the scale of the histogram so that the true nature of the distribution is totally obscured

Definitions

Slide 52

- For a set of data, the **5-number summary** consists of the minimum value; the first quartile Q_1; the median (or second quartile Q_2); the third quartile, Q_3; and the maximum value
- A **boxplot** (or **box-and-whisker-diagram**) is a graph of a data set that consists of a line extending from the minimum value to the maximum value, and a box with lines drawn at the first quartile, Q_1; the median; and the third quartile, Q_3

Boxplots

Slide 53

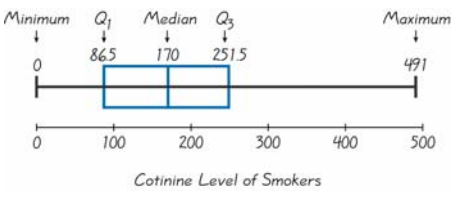


Figure 2-16

Boxplots

Slide 54

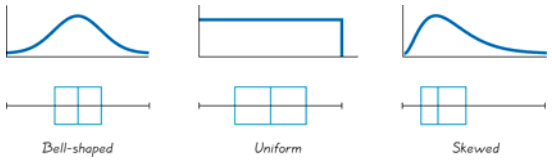


Figure 2-17