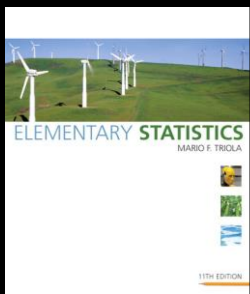
Lecture Slides



Elementary Statistics
Eleventh Edition

and the Triola Statistics Series

by Mario F. Triola

PEARSON

Preview

- **Descriptive Statistics**

  In this chapter we'll learn to summarize or describe the important characteristics of a known set of data

- **Inferential Statistics**

  In later chapters we'll learn to use sample data to make inferences or generalizations about a population

---

Measure of Center Or Measures of Central Tendency

- **Measure of Center**

  the value at the center or middle of a data set

Arithmetic Mean

- **Arithmetic Mean (Mean)**

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

What most people call an *average*.

Notation

Σ denotes the sum of a set of values.

x is the variable usually used to represent the individual data values.

n represents the number of data values in a sample.

N represents the number of data values in a population.

Notation

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

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

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

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

Mean

- **Advantages**
  Is relatively reliable, means of samples drawn from the same population don't vary as much as other measures of center
  Takes every data value into account
- **Disadvantage**
  Is sensitive to every data value, one extreme value can affect it dramatically; is not a *resistant* measure of center

Median

- **Median**
  the **middle value** when the original data values are arranged in order of increasing (or decreasing) magnitude
- is not affected by an extreme value -- is a resistant measure of the center

5.40	1.10	0.42	0.73	0.48	1.10
0.42	0.48	0.73	1.10	1.10	5.40

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

$$\frac{0.73 + 1.10}{2}$$

MEDIAN is ______

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 _________

Mode

- **Mode**
  the value that occurs with the **greatest frequency**
- Data set can have one, more than one, or no mode

Bimodal two data values occur with the same greatest frequency

Multimodal more than two data values occur with the same greatest frequency

No Mode no data value is repeated

Mode - Examples

a. 5.40 1.10 0.42 0.73 0.48 1.10

Mode is _____

b. 27 27 27 55 55 55 88 88 99 _________

c. 1 2 3 6 7 8 9 10 _________

Round-off Rule for Measures of Center

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

Skewed and Symmetric

- **Symmetric**

  distribution of data is symmetric if the left half of its histogram is roughly a mirror image of its right half

- **Skewed**

  distribution of data is skewed if it is not symmetric and extends more to one side than the other

Skewed Left or Right

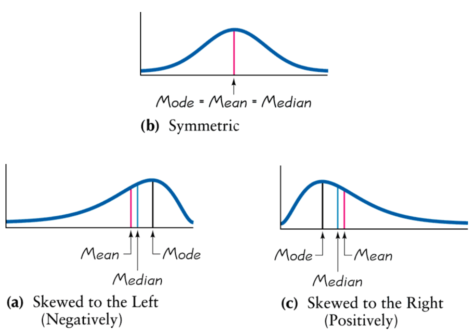
- **Skewed to the left**

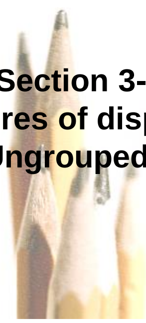
  (also called negatively skewed) have a longer left tail, mean and median are to the left of the mode

- **Skewed to the right**

  (also called positively skewed) have a longer right tail, mean and median are to the right of the mode

Skewness





Section 3-2 Measures of dispersion for Ungrouped Data

Definition

The range of a set of data values is the difference between the maximum data value and the minimum data value.

Range = (maximum value) – (minimum value)

It is very sensitive to extreme values; therefore not as useful as other measures of variation.

Round-Off Rule for Measures of Variation

When rounding the value of a measure of variation, 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

The standard deviation of a set of sample values, denoted by s, is a measure of variation of values about the mean.

Population Standard Deviation (Shortcut Formula)

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

Sample Standard Deviation (Shortcut Formula)

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

Standard Deviation - Important Properties

- 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.

Properties of the Standard Deviation

- Measures the variation among data values
- Values close together have a small standard deviation, but values with much more variation have a larger standard deviation
- Has the same units of measurement as the original data

Properties of the Standard Deviation

- For many data sets, a value is *unusual* if it differs from the mean by more than two standard deviations
- Compare standard deviations of two different data sets only if the they use the same scale and units, and they have means that are approximately the same

Variance

- The variance of a set of values is a measure of variation equal to the square of the standard deviation.
- Sample variance: s^2 - Square of the sample standard deviation s
- Population variance: σ^2 - Square of the population standard deviation σ

Variance - Notation

s = *sample* standard deviation

s^2 = *sample* variance

σ = *population* standard deviation

σ^2 = *population* variance

Unbiased Estimator

The sample variance s^2 is an unbiased estimator of the population variance σ^2, which means values of s^2 tend to target the value of σ^2 instead of systematically tending to overestimate or underestimate σ^2.

Section 3-4
Use of Standard Deviation

By using the mean and standard deviation, we can find the proportion or percentage of total observations that fall within a given interval about the mean.

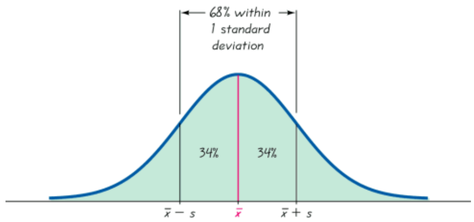
- Empirical Rule
- Chebyshev's Theorem

Empirical (or 68-95-99.7) Rule

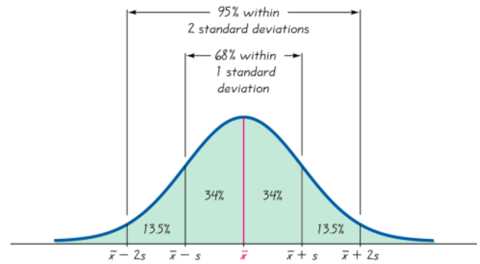
For data sets having a distribution that is approximately bell shaped, the following properties apply:

- 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.

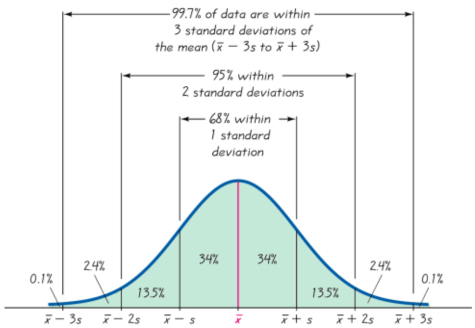
The Empirical Rule



The Empirical Rule



The Empirical Rule



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.**

A sample of 1000 observations has a mean of 64 and a standard deviation of 8.

a) Using Chebyshen's thm., find at least what % of the observations fall in the intervals $\bar{x} \pm 2s$ *and* $\bar{x} \pm 1.5s$

b) Using the empirical rule, find what percentage of the observations fall in the intervals $\mu \pm 1\sigma$ and $\mu \pm 2\sigma$

The average systolic blood pressure for 4000 women who were screened for high blood pressure was found to be 187 with a standard deviation of 22. Using Chebyshev's thm., find at least what percentage of women in this group have a systolic blood pressure between 143 and 231

The age distribution of a sample of 5000 persons is bell-shaped with a mean of 40 years and a standard deviation of 12 years. Determine the approximate percentage of people who are 16 to 64 years old.

Section 3-5 and 3-6 Measures of Position and Box and Whisker Plots

Quartiles

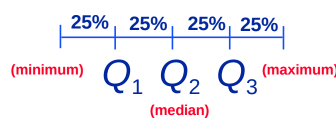
Are measures of location, denoted Q_1, Q_2, and Q_3, which divide a set of RANKED data into four groups with about 25% of the values in each group.

- Q_1 (First Quartile) seperates approx. 25% of sorted values from the top (apprx.) 75%.
- Q_2 (Second Quartile) same as the median; separates the data into two equal parts.
- Q_3 (Third Quartile) separates (apprx.) the bottom 75% of sorted values from the (apprx) top 25%.

Quartiles

$$Q_1, \; Q_2, \; Q_3$$

divide ranked scores into four equal parts



A) Find the values of the three quartiles.
B) Find the interquartile range

47 28 39 51 33 37 59 24 33

Calculating Percentiles

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 kth percentile

35 29 44 72 34 64 41 50 54 104 39 58

Find the value of the 42nd percentile and interpret

Find the value of the 72nd percentile and interpret

Finding the Percentile of a Data Value

Percentile of value x =

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

35 29 44 72 34 64 41 50 54 104 39 58

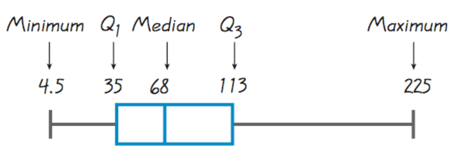
Find the percentile rank of 64 and interpret

Find the percentile rank of 50 and interpret.

Boxplot

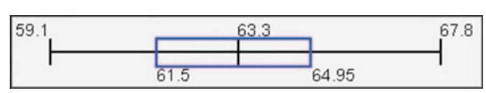
- A **boxplot** (or **box-and-whisker-diagram**) is a graph of a data set that consists of a line extending from the minimum value to the maximum value, and a box with lines drawn at the first quartile, Q_1; the median; and the third quartile, Q_3.

Boxplots



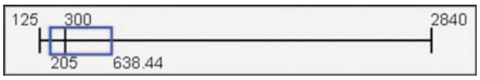
Boxplot of Movie Budget Amounts

Boxplots - Normal Distribution



Normal Distribution:
Heights from a Simple Random Sample of Women

Boxplots - Skewed Distribution



Skewed Distribution:
Salaries (in thousands of dollars) of NCAA Football Coaches

Outliers

- An outlier is a value that lies very far away from the vast majority of the other values in a data set.

Important Principles

- An outlier can have a dramatic effect on the mean.
- An outlier can 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.

Outliers

For purposes of constructing *modified boxplots*, we can consider outliers to be data values meeting specific criteria.

In modified boxplots, a data value is an outlier if it is . . .

above Q_3 by an amount greater than $1.5 \times$ IQR

or

below Q_1 by an amount greater than $1.5 \times$ IQR

Draw a box and whisker plot for the following data

35	29	44	72	34	64	41	50
54	104	39	58				