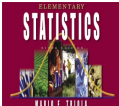


Statistics

- ❖ **Data (sing., datum)**
observations (such as measurements, counts, survey responses) that have been collected.
- ❖ **Statistics**
a collection of methods for planning experiments, obtaining data, and then then organizing, summarizing, presenting, analyzing, interpreting, and drawing conclusions based on the data.

<http://www.youtube.com/watch?v=fsRYkRqQagp&feature=related>



Data Sampling

- ❖ **Population**
the complete collection of all elements (scores, people, measurements, and so on) to be studied. The collection is complete in the sense that it includes all subjects to be studied.

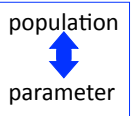
Data collecting tools sample data from a smaller part of a larger group so we can learn something about the larger group.
- ❖ **Sample**
a sub-collection of elements drawn from a population.


Key Concepts

- ❖ Sample data must be collected in an appropriate way, such as through a process of **random** selection.
- If sample data are not collected in an appropriate way, the data may be so completely useless that no amount of statistical torturing can salvage them.

Definitions

- ❖ **Parameter**
a numerical measurement describing some characteristic of a **population**
- ❖ **Statistic**
a numerical measurement describing some characteristic of a **sample**.





Definitions

- ❖ **Quantitative data**
Numbers representing counts or measurements.
Example: weights, number of individuals, etc.
Should have a scale with a "true" zero
0°C is *not* an absence of heat!

Quantitative data can further be distinguished between **discrete** and **continuous** types.

Definitions

- ❖ **Quantitative data**
 - ❖ **Discrete**
Data result when the number of possible values is either a finite number or a 'countable' number of possible values.

0, 1, 2, 3, . . .

Example: The number of eggs that hens lay.

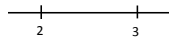
Definitions

❖ Quantitative data

❖ Continuous

(Numerical) data result from infinitely many possible values that correspond to some continuous scale that covers a range of values without gaps, interruptions, or jumps.

Example: The amount of milk that a cow produces;
e.g. 2.343115 gallons per day.



Definitions

❖ Qualitative (or categorical or attribute) data

Can be separated into different categories that are distinguished by some nonnumeric characteristics.

Example: genders (male/female) or color

Overview

❖ Descriptive Statistics

summarize or describe the important characteristics of a known set of population data

❖ Inferential Statistics

use sample data to make inferences (predictions or generalizations) about a population

Descriptive Statistics

Important Characteristics of Data

1. **Size:** The number of values in a data set
2. **Center:** A representative or average value that indicates where the middle of the data set is located
3. **Variation:** A measure of the amount that the values vary among themselves
4. **Distribution:** The nature or shape of the distribution of data (such as bell-shaped, uniform, or skewed)

Notation

- n represents the number of values in a sample
- N represents the number of values in a population
- Σ (*sigma* = sum) denotes the addition of a set of values
- x is the variable, usually used to represent the individual data values

Definition

❖ Measure of Center

The value at the center or middle of a data set

Definition

Arithmetic Mean
(Mean)

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

Notation

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

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

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

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

Definition

❖ **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

❖ 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

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**

Definition

❖ **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

- 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

Definition

❖ Midrange

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

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

Round-off Rule for Measures of Center

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

Best Measure of Center

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	$\bar{x} = \frac{\sum x}{n}$	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).

http://www.youtube.com/watch?v=sydzT_WIRz4&feature=related

Definitions

❖ Symmetric

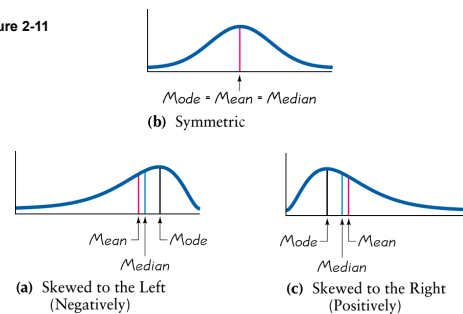
Data is symmetric if the left half of its histogram is roughly a mirror image of its right half.

❖ Skewed

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

Skewness

Figure 2-11



Measures of Variation

- Range
- Standard Deviation
- Variance

Definition

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

$$\text{highest value} - \text{lowest value}$$

Definition

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

Population Standard Deviation

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

Calculated using the population mean and population size

Sample Standard Deviation Formula

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

Note: divide by $n-1$ rather than n .

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

Definition

- ❖ 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^2

- ❖ **Population variance**: Square of the population standard deviation

σ^2

Round-off Rule for Measures 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.

Estimation of Standard Deviation Range Rule of Thumb

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

Minimum "usual" value \approx (mean) $- 2 \times$ (standard deviation)

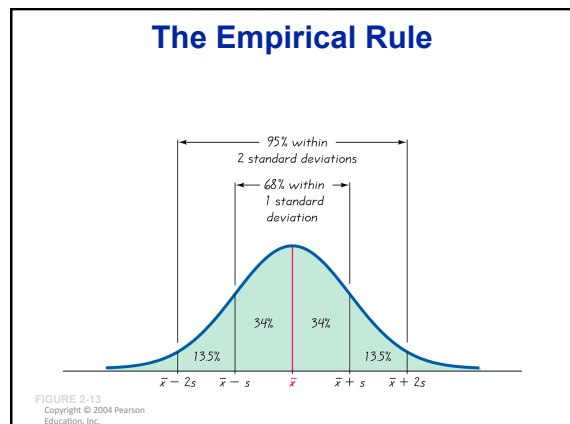
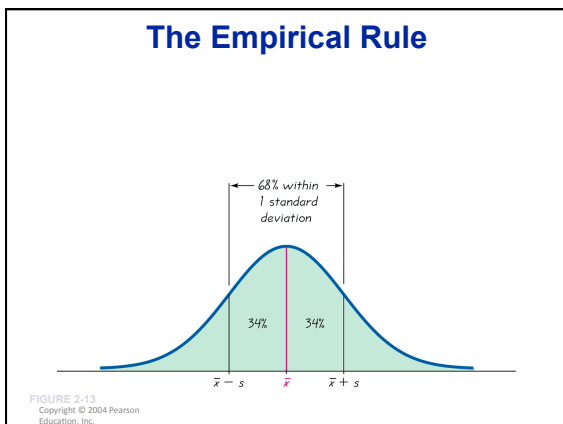
Maximum "usual" value \approx (mean) $+ 2 \times$ (standard deviation)

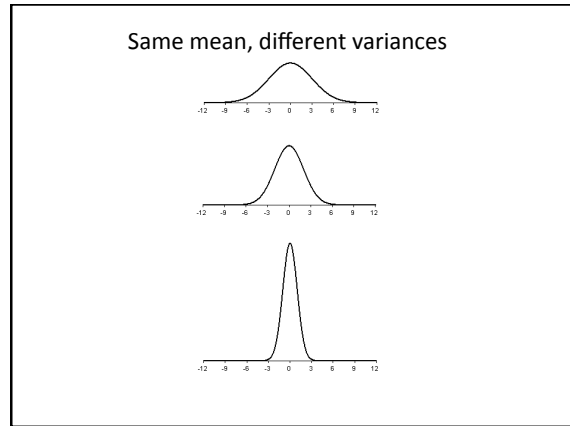
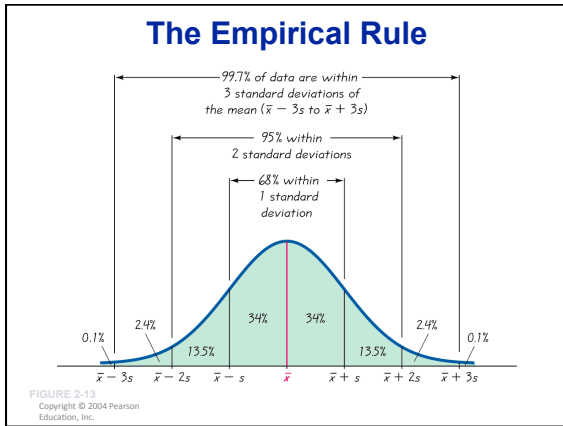
Definition

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





Statistical Confidence

Variation in a **population** exists because of **natural variation**.

Variation in a **sample** results from **natural variation**, compounded by **sampling error** and **sloppy measurements**.

Measurements of confidence:

- **Coefficient of Variation (CV)**: ratio of variation versus the mean
- **Standard error (SE)**: ratio of variation versus the sample size

Definition: Coefficient of Variation

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

Definition: Standard Error

The **standard error** (or **SE**) for a set of sample or population data describes the standard deviation relative to the sample size

$$SE = \frac{s}{\sqrt{n}}$$

Definition: Standard Error

How confident should you be that \bar{x} of your **sample** represents μ of the **population**?

Applying the empirical rule again, we can be **95% confident** that μ lies within $\bar{x} \pm 1.96 SE$. (the "95% confidence interval")