

# Research Methods Workshops

## Data Analysis: Statistical Analysis Principles and Probability

### Readings

Essential Readings:

- Berenson et al. Chapter 1, 3 and 4
- Groebner et al. Chapter 4
- *Research Methods for Business Students*, Mark Saunders, Philip Lewis, Adrian Thornhill, Pearson Education, 2009

# Learning Outcomes

By the end of this session you should be able to:

- Understand what statistics is.
- Explain what is meant by descriptive statistics and inferential statistics.
- Calculate the arithmetic mean, the weighted mean (not all data points contribute equally to the mean), the geometric mean (measure of central tendency), the median, and the mode
- Compute and interpret the range, the variance, the standard deviation, covariance and correlation coefficient
- Understand basic probability concepts, rules and population / sample distribution measures.

## What is statistical analysis?

- **Statistics** is the science of collecting, organizing, presenting, analyzing, and interpreting numerical/non-numerical data for the purpose of assisting in making a more effective decision or forming a judgment as to the likelihood of a hypothesis being true.
- It is also a branch of mathematics which transform numbers (especially from very large data sets) into useable information for researchers, decision makers and interpreters of those data sets.
- It also provides several means of validating whether our analysis of data justifies the inferences we make, based on that data and the confidence we can place in those inferences.

# Types of Statistical Analysis

- The branch of mathematics that transforms data into useful information to be presented as evidence.



## Descriptive Statistics

Collecting, summarizing, and describing data - in order to present evidence for findings.

Converting numerical data sets into graphical representations.

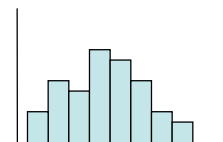
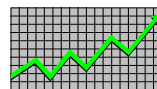


## Inferential Statistics

Drawing conclusions and/or establishing relationships amongst two or more variables, concerning a population based only on sample data.

# Descriptive Statistics

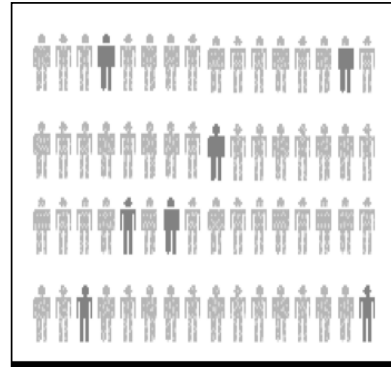
- Collect data
  - e.g., Survey
- Present data
  - e.g., Tables and graphs



- Characterize data
  - e.g., Sample mean =  $\frac{\sum X_i}{n}$  (arithmetic mean)

# Inferential Statistics

- Estimation
  - e.g., Estimate the population mean weight using the sample mean weight
- Hypothesis testing
  - e.g., Test the claim that the population mean weight is 120 pounds
- Testing a possible relationship between two or more variables



Drawing a conclusion about a population based on data sampled from a subset of the population

Basic Business Statistics, 11e © Prentice-Hall, Inc. 2009

## Measure of Central Tendency

**Measure of Central Tendency** (“the mean”, “average”):

A single value that summarises a set of data. It locates the centre of the values -- and indicates (in a normally distributed population) the most likely value – one that is most typical (or representative).

The **central tendency** is the extent to which all the data values group around a typical or central value.

### **Note:**

There is not just one measure of central tendency; in fact, there are many. i.e. the arithmetic mean, the weighted mean, the median, the mode, the geometric mean and the harmonic mean.  
[See Exercise 2].

# Arithmetic Mean

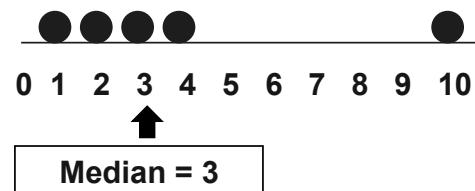
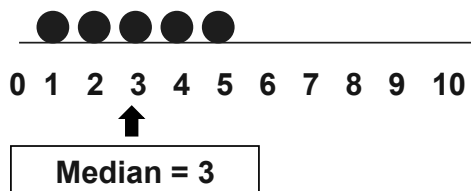
- The arithmetic mean (often (lazily) called the “average”) is the most common measure of central tendency
  - For a sample of size  $n$ :

The diagram shows the formula for the arithmetic mean,  $\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{X_1 + X_2 + \dots + X_n}{n}$ . Four labels with arrows point to parts of the formula: 'Pronounced x-bar' points to  $\bar{X}$ ; 'The i<sup>th</sup> value' points to  $X_i$ ; 'Sample size' points to  $n$  in the denominator; and 'Observed values' points to the sum of  $X_1$  through  $X_n$ .

Basic Business Statistics, 11e © Prentice-Hall, Inc. 2009

# The Median

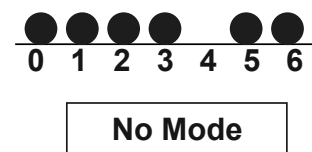
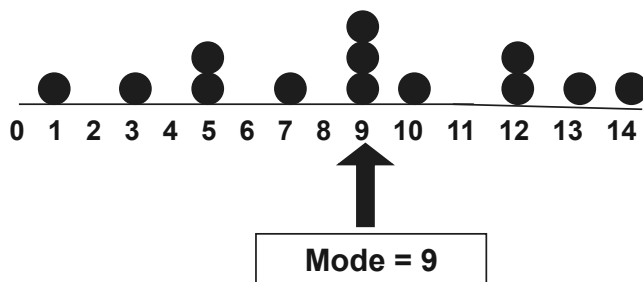
- In an ordered array, the median is the “middle” number (50% above, 50% below).



- It is not affected by extremely large or small values and is therefore a valuable *measure of central tendency* when such values occur in a data sample.
- It can be misleading in non-normal (multi-modal) distributions.

# The Mode

- Value that occurs most often
- Not affected by extreme values
- Used for either numerical or categorical (nominal) data
- There may may be no mode if data is evenly distributed.
- There may be several modes, if there are peaks of distribution



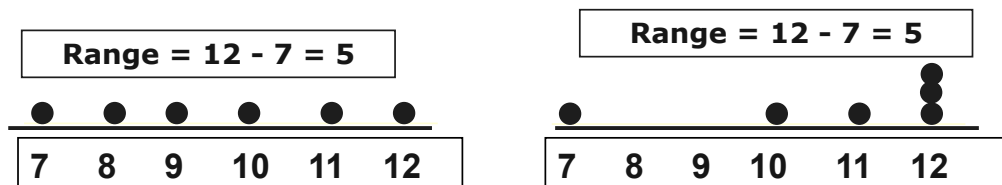
# The Range

Range is the simplest measure of variation

$$\text{Range} = X_{\text{largest}} - X_{\text{smallest}}$$

The Range can be misleading

- Ignores the way in which data are distributed



- Sensitive to outliers

**1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,3,3,3,3,4,5**

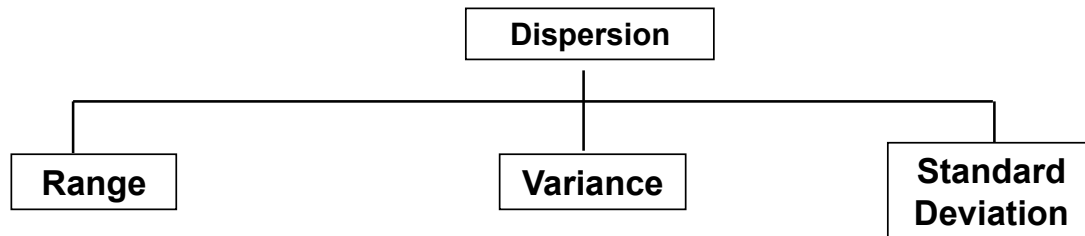
$$\text{Range} = 5 - 1 = 4$$

**1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,3,3,3,3,4,120**

$$\text{Range} = 120 - 1 = 119$$

# Measures of Dispersion

- We will consider several measures of dispersion.



- Measures of variation give information on the spread or variability or dispersion of the data values.

## The Variance

- Average of squared deviations of values from the mean
  - Sample variance:

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

Where  $\bar{X}$  = arithmetic mean  
 $n$  = sample size  
 $X_i$  =  $i^{\text{th}}$  value of the variable  $X$

- Population variance ( $p$ ):

$$p^2 = \frac{\sum_{i=1}^N (X_i - m)^2}{N}$$

Where  $m$  = population mean  
 $N$  = population size  
 $X_i$  =  $i^{\text{th}}$  value of the variable  $X$

# The Standard Deviation (STDEV)

- Most commonly used measure of variation
- Shows variation about the mean
- Is the square root of the variance
- Has the same units as the original data

– Sample standard deviation:

$$S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}}$$

– Population standard deviation:  
where  $p$  = STDEV of population and  
 $m$  = population mean (arithmetic),  
 $N$  = population size.

$$p = \sqrt{\frac{\sum_{i=1}^N (X_i - m)^2}{N}}$$

## The Covariance

- The covariance measures the strength of the linear relationship between **two numerical variables** (X & Y).
- The sample covariance:

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

- Only concerned with the strength of the relationship.
- No causal effect is implied.

# The Coefficient of Correlation

- Measures the relative strength of the linear relationship between two numerical variables
- Sample coefficient of correlation:

where 
$$r = \frac{\text{cov}(X, Y)}{S_X S_Y}$$

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1} \quad S_X = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}} \quad S_Y = \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}}$$

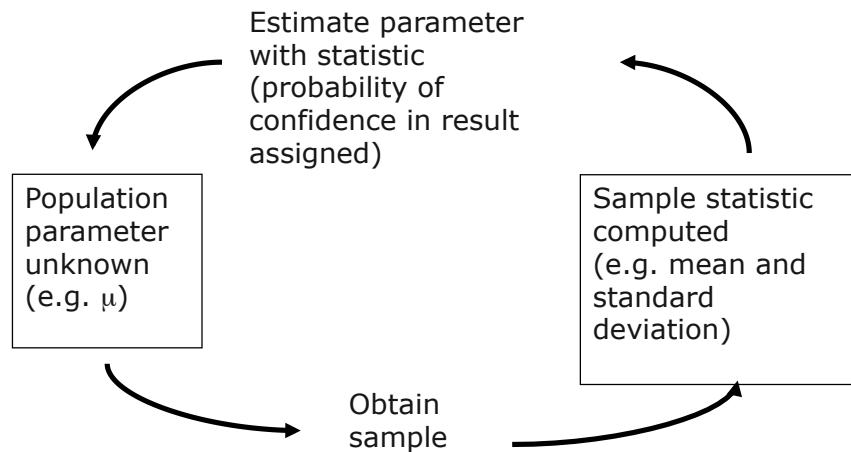
Basic Business Statistics, 11e © Prentice-Hall, Inc. 2009

## Introduction to Probability

- Much statistical analysis is inferential, and probability is the basis for inferential statistics. Remember that inferential statistics involves taking a sample from a population, calculating a statistical analysis on the sample, and inferring from the statistic the value of the corresponding parameter of the population.
- Because the value of the parameter is unknown, we have to conduct the inferential process under uncertainty. However, by applying rules and laws, we can often assign a probability of obtaining accurate results (within a confidence interval).

# Introduction to Probability, Cont.

## Probability in the Process of Inferential Statistics



Statistical Techniques in Business & Economics, ©The McGraw-Hill Companies, Inc. 2008

## Determining Probability

- Classical Probability Assessment

→ All outcomes equally likely to occur

$$P(E_i) = \frac{\text{Number of ways } E_i \text{ can occur}}{\text{Total number of experimental outcomes}}$$

- Relative Frequency of Occurrence

→ Based on actual observations

$$\text{Relative Freq. of } E_i = \frac{\text{Number of times } E_i \text{ occurs}}{N}$$

- Subjective Probability Assessment

An opinion or judgment by a decision maker about the likelihood of an event based on their expertise

# Rules of Probability

Individual Values

$$0 \leq P(E_i) \leq 1$$

For any event  $E_i$

Rule 1

The value of a probability is between 0 and 1  
0 = no chance of occurring  
1 = 100% change of occurring

Sum of All Values

$$\sum_{i=1}^k P(e_i) = 1$$

where:  
 $k$  = Number of individual outcomes in the sample space

$e_i$  =  $i^{\text{th}}$  individual outcome

Rule 2

The sum of the probabilities of all the outcomes in a sample space must = 1 or 100%

Business Statistics: A Decision-Making Approach, 8e©Pearson Education, Inc. 2011

## Rules of Probability, Cont.

### *Addition Rule*

- The probability of an event  $E_i$  is equal to the sum of the probabilities of the individual outcomes forming  $E_i$ .
- That is, if:  $E_i = \{e_1, e_2, e_3\}$

then:  $P(E_i) = P(e_1) + P(e_2) + P(e_3)$

Rule 3

Business Statistics: A Decision-Making Approach, 8e©Pearson Education, Inc. 2011

# Rules of Probability, Cont.

## *Multiplication Rule*

- The special rule of multiplication requires that two events  $A$  and  $B$  are independent.
- Two events  $A$  and  $B$  are independent if the occurrence of one has no effect on the probability of the occurrence of the other.
- This rule is written:  $P(A \text{ and } B) = P(A)P(B)$

Rule 4

# Rules of Probability, Cont.

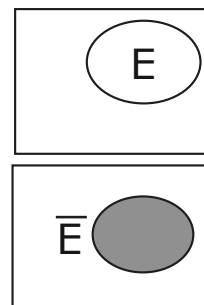
## *Complement Rule*

- The complement of an event  $E$  is the collection of all possible elementary events **not** contained in event  $E$ . The complement of event  $E$  is represented by  $\bar{E}$ .

- Complement Rule:

e.g., If there is a 40% chance it will rain, there is a 60% chance it won't rain!

$$P(\bar{E}) = 1 - P(E)$$



Or,  $P(E) + P(\bar{E}) = 1$

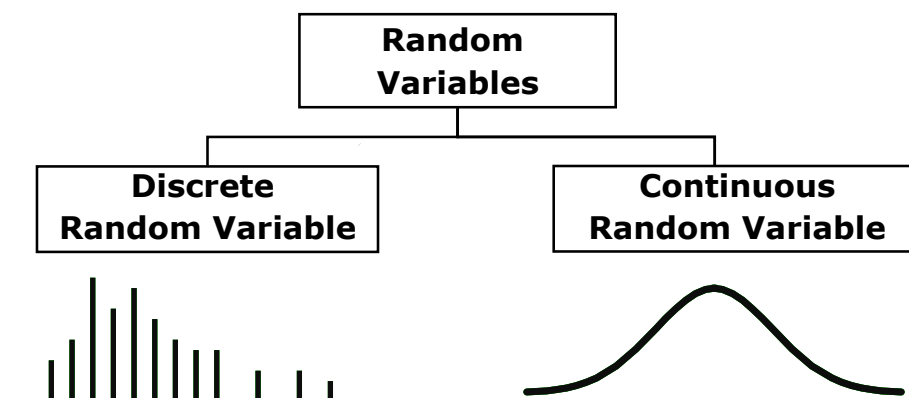
Rule 5

# Definitions: Random Variables

- A **random variable** represents a possible numerical value from an uncertain event.
- **Discrete** random variables produce outcomes that come from a counting process (e.g. number of purchases you are making (in a single shop visit)).
- **Continuous** random variables produce outcomes that come from a measurement (e.g. your annual salary, or your vehicle's fuel consumption).

Basic Business Statistics, 11e © Prentice-Hall, Inc. 2009

## Definitions: Random Variables, Cont.



Basic Business Statistics, 11e © Prentice-Hall, Inc. 2009

# Discrete Probability Distribution

- A probability distribution for a discrete random variable is a mutually exclusive listing of all possible numerical outcomes for that variable and a probability of occurrence associated with each outcome.

| Number of purchases made | Probability |
|--------------------------|-------------|
| 2                        | 0.2         |
| 3                        | 0.4         |
| 4                        | 0.24        |
| 5                        | 0.16        |

Basic Business Statistics, 11e © Prentice-Hall, Inc. 2009

# Continuous Probability Distributions

- A continuous random variable is a variable that can assume any value on a continuum (can assume an uncountable number of values)
  - thickness of an item.
  - time required to complete a task.
  - temperature of a solution.
  - height, in centimeters.
- These can potentially take on any value depending only on the ability to measure precisely and accurately.

Basic Business Statistics, 11e © Prentice-Hall, Inc. 2009

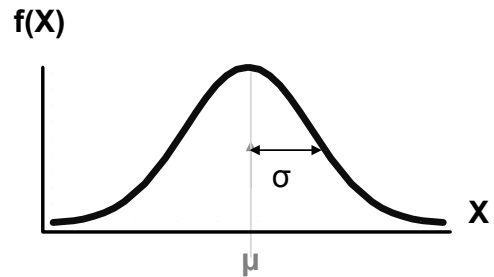
# The Normal Distribution

- 'Bell Shaped'
- Symmetrical
- Mean, Median and Mode are equal
- Location is determined by the mean,  $\mu$

Spread is determined by the standard deviation,  $\sigma$

The random variable has an infinite theoretical range:

$+\infty$  to  $-\infty$



$\mu$  = Mean  
= Median  
= Mode