Department of Business Administration
University of Global Village

Business Statistics

Course Code	**:**	0542-625	**Total Credit**	**:**	3 **(**Three**)**
			CIE Marks	**:**	90
Semester End Exam (SEE) Hours	**:**	03	SEE Marks	**:**	60

Course Learning Outcomes: At the end of the Course, the Student will be able to-

CLO 1	➢ Understand the meaning, definition, nature, importance and limitations of statistics. ➢ Acquire the knowledge about "An Introduction to Business Statistics".
CLO 2	➢ Understanding of the calculations and main properties of measures of central tendency, including mean, mode, median, quartiles, percentiles, etc. ➢ Impart the knowledge of measures of dispersion and skewness and to enable the students to distinguish between average, dispersion, skewness, moments and kurtosis.
CLO 3	➢ Understanding of bivariate linear correlation, thereby enabling you to understand the importance as well as the limitations of correlation analysis.
CLO 4	➢ Enabling the students to appreciate the relevance of probability theory in decision-making under conditions of uncertainty. After successful completion of the lesson the students will be able to understand and use the different approaches to probability as well as different probability rules for calculating probabilities in different situations.
CLO 5	➢ Able to understand: various terms associated with sampling; various methods of probability and non-probability sampling and how to determine sample size. ➢ Specify the most appropriate test of hypothesis in a given situation, apply the procedure and make inferences from the results.

Department of Business Administration
University of Global Village

Course plan specifying Topics, Teaching time and CLOs

Chapter No.	Content	Hours	CLOs
01	Introduction to Statistics	**8**	CLO 1
02	Measure of Desperation	**10**	CLO 2
03	Correlation	**8**	CLO 3
04	Probability	**8**	CLO 4
05	Hypothesis	**8**	CLO 5

Course plan specifying content, CLOs, Teaching Learning, and Assessment Strategy Mapped with CLOs:

WeeK	Content	Teaching Learning Strategy	Assessment Strategy	
1,2 &3	**Introduction:** Meaning and Definitions of Statistics, Types of Data and Data Sources, Types of Statistics, Scope of Statistics, Importance of Statistics in Business, Limitations of statistics, Summary, Self-Test Questions, Suggested Readings	Class Lecture, Open discussion	Question & Answer (Oral) Class Test Written Test	CLO 1
4,5,6 &7	**Measure of Desperation:** Arithmetic Mean, Median, Mode, Relationships of the Mean, Median and Mode, The Best Measure of Central Tendency, Geometric Mean, Harmonic Mean, Quadratic Mean, Meaning and Definition of Dispersion, Significance and Properties of Measuring Variation, Measures of Dispersion, Range, Interquartile Range or Quartile Deviation, Mean Deviation, Standard Deviation, Lorenz Curve, Skewness:	Class Lecture, Open discussion	Question & Answer (Oral) Class Test Written Test	CLO 2

	Meaning and Definitions, Tests of Skewness, Measures of Skewness, Moments, Kurtosis			
8,9 &10	**Correlation:** What is Correlation, Correlation Analysis, Scatter Diagram, Correlation Graph, Pearson's Coefficient of Correlation, Spearman's Rank Correlation, Concurrent Deviation Method, Limitations of Correlation Analysis	Class Lecture, Open discussion	Question & Answer (Oral) Class Test Written Test	CLO 3
11, 12 & 13	**Probability:** Some Basic Concepts, Approaches to Probability Theory, Probability Rules, Bayes' Theorem, Some Counting Concepts, Discrete Probability Distribution, Bernoulli Random Variable, The Binomial Distribution, The Poisson Distribution, Continuous Probability Distribution, The Normal Distribution, The Standard Normal Distribution, The Transformation of Normal Random Variables	Class Lecture, Open discussion	Question & Answer (Oral) Class Test Written Test	CLO 4
14,15, 16 & 17	**Hypothesis:** Introduction, The Null and the Alternative Hypothesis, Some Basic Concepts, Critical Region in Terms of Test Statistic, General Testing Procedure, Tests of Hypotheses about Population Means, Tests of Hypotheses about Population Proportions, Tests of Hypotheses about Population Variances, The Comparison of Two Populations, Null and Alternative Hypotheses, Outcomes and the Type I and Type II Errors, Distribution Needed for Hypothesis Testing, Full Hypothesis Test Examples, Comparing Two Independent Population Means, Cohen's Standards for Small, Medium, and Large Effect Sizes, Test for Differences in Means: Assuming Equal Population Variances, Comparing Two Independent Population Proportions, Two Population Means with Known Standard Deviations, Matched or Paired Samples	Class Lecture, Open discussion	Question & Answer (Oral) Class Test Written Test Presentation	CLO 5

Department of Business Administration
University of Global Village

Text Books:

[1] Business Statistics by S. P. Gupta and M. P. Gupta, (ii) = Business Statistics by Surinder Kundu and DR. B. S. Bodla,

[2] Introductory Business Statistics by Alexander Holmes, Barbara Illowsky, and Susan Dean

1) Assessment Strategy: Group Discussion, Class tests, Case Study, Term Paper, Presentation.

2) Marks distribution:

a) Continuous Assessment:

- Class attendance is mandatory. Absent of 70% of classes; disqualify the student for final examination only authority recommendations will be accepted with highly reasonable causes.
- Late submission of assignments is not allowed. Late submission of assignments
- will be only taken with highly reasonable causes and a 20% mark will be deducted.
- To pass this course students will have to appear in mid-term and final examinations.

b) Summative:

CIE- Continuous Internal Evaluation (45 Marks)

Bloom's Category Marks (out of 45)	Tests (25)	Assignments (15)	Quizzes (05)	External Participation in Curricular/Co-Curricular Activities (5)
Remember			05	
Understand		05		
Apply	08			5
Analyze	09			
Evaluate	08	05		
Create		05		

SEE- Semester End Examination (60 Marks)

Bloom's Category	Tests
Remember	10
Understand	10
Apply	20
Analyze	10

Evaluate	10
Create	10

3) Make-up Procedures: Dates for exams will be strictly followed. No makeup exam (Normal case), for exceptional cases university rules and regulations should be followed.

What is Statistics

Chapter 1

Week 1

GOALS

- Understand why we study statistics.
- Explain what is meant by descriptive statistics and inferential statistics.
- Distinguish between a qualitative variable and a quantitative variable.
- Describe how a discrete variable is different from a continuous variable.
- Distinguish among the nominal, ordinal, interval, and ratio levels of measurement.

What is Meant by Statistics?

Statistics is the science of collecting, organizing, presenting, analyzing, and interpreting numerical data to assist in making more effective decisions.

Who Uses Statistics?

Statistical techniques are used extensively by marketing, accounting, quality control, consumers, professional sports people, hospital administrators, educators, politicians, physicians, etc...

Types of Statistics

1. Descriptive Statistics - methods of organizing, summarizing, and presenting data in an informative way.

   EXAMPLE 1: A Gallup poll found that 49% of the people in a survey knew the name of the first book of the Bible. The statistic 49 describes the number out of every 100 persons who knew the answer.

Types of Statistics – Descriptive Statistics

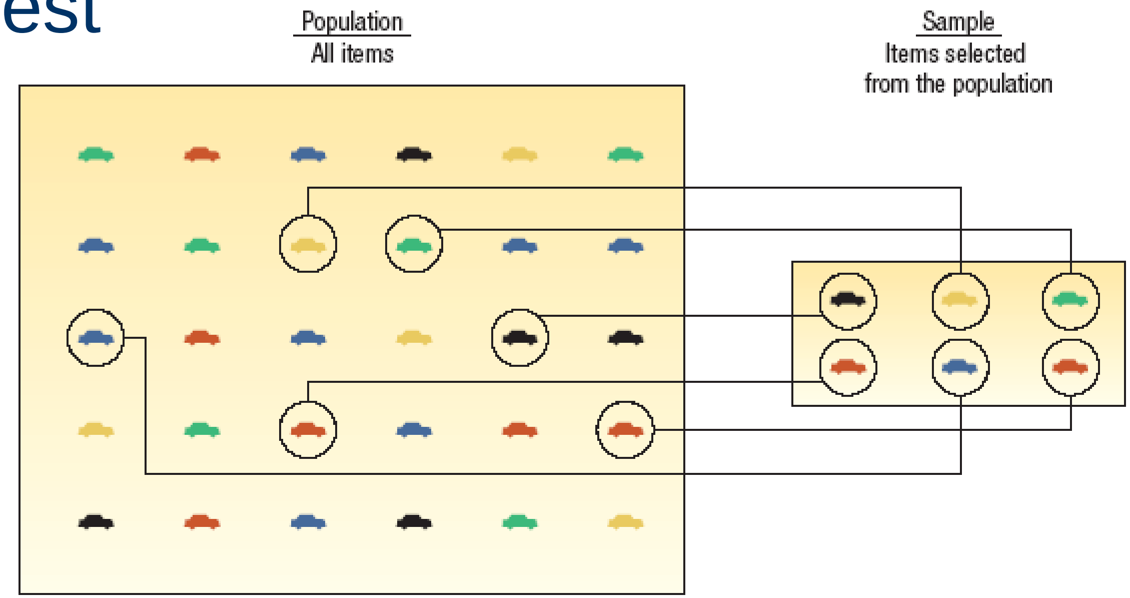
EXAMPLE 2: According to Consumer Reports, General Electric washing machine owners reported 9 problems per 100 machines during 2001. The statistic 9 describes the number of problems out of every 100 machines.

2. Inferential Statistics: A decision, estimate, prediction, or generalization about a population, based on a sample.

Population versus Sample

A population is a collection of all possible individuals, objects, or measurements of interest.

A sample is a portion, or part, of the population of interest



Types of Variables

A. Qualitative or Attribute variable - the characteristic being studied is nonnumeric.

EXAMPLES: Gender, religious affiliation, type of automobile owned, state of birth, eye color are examples.

B. Quantitative variable - information is reported numerically.

EXAMPLES: balance in your checking account, minutes remaining in class, or number of children in a family.

Quantitative Variables - Classifications

Quantitative variables can be classified as either discrete or continuous.

A. Discrete variables: can only assume certain values and there are usually "gaps" between values.

EXAMPLE: the number of bedrooms in a house, or the number of hammers sold at the local Home Depot (1,2,3,…,etc).

B. Continuous variable can assume any value within a specified range.

EXAMPLE: The pressure in a tire, the weight of a pork chop, or the height of students in a class.

Summary of Types of Variables

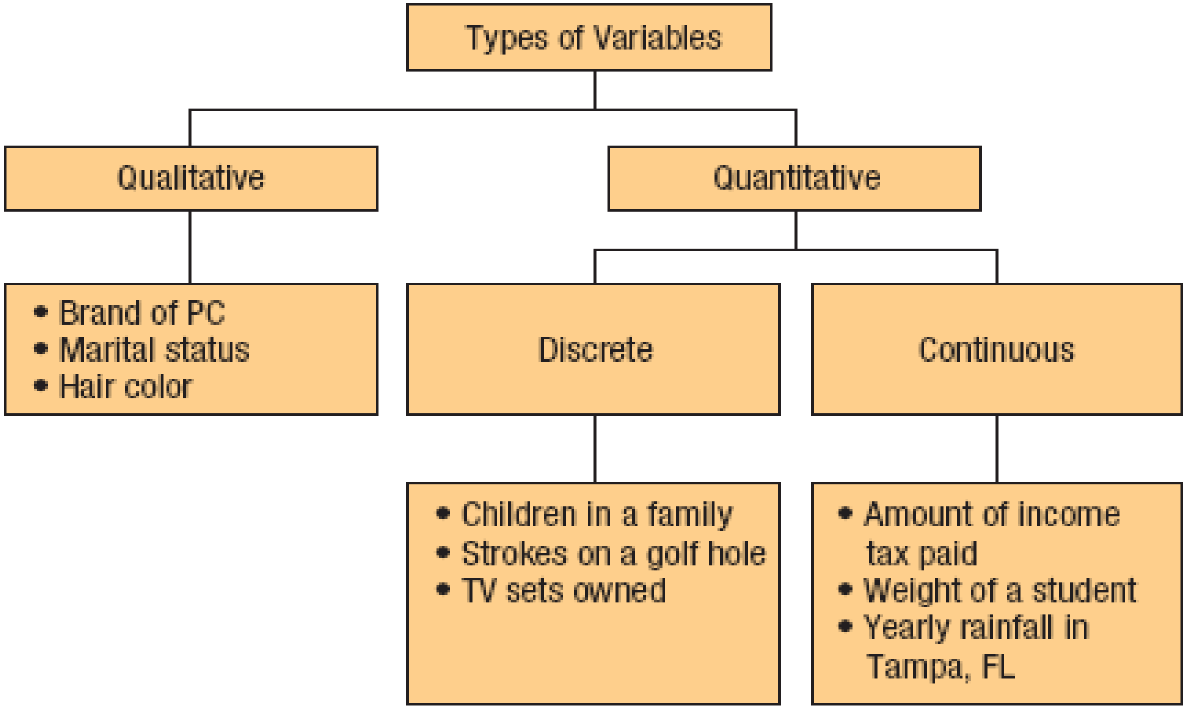


CHART 1–2 Summary of the Types of Variables

Four Levels of Measurement

Nominal level - data that is classified into categories and cannot be arranged in any particular order.

EXAMPLES: eye color, gender, religious affiliation.

Ordinal level – involves data arranged in some order, but the differences between data values cannot be determined or are meaningless.

EXAMPLE: During a taste test of 4 soft drinks, Mellow Yellow was ranked number 1, Sprite number 2, Seven-up number 3, and Orange Crush number 4.

Interval level - similar to the ordinal level, with the additional property that meaningful amounts of differences between data values can be determined. There is no natural zero point.

EXAMPLE: Temperature on the Fahrenheit scale.

Ratio level - the interval level with an inherent zero starting point. Differences and ratios are meaningful for this level of measurement.

EXAMPLES: Monthly income of surgeons, or distance traveled by manufacturer's representatives per month.

Summary of the Characteristics for Levels of Measurement

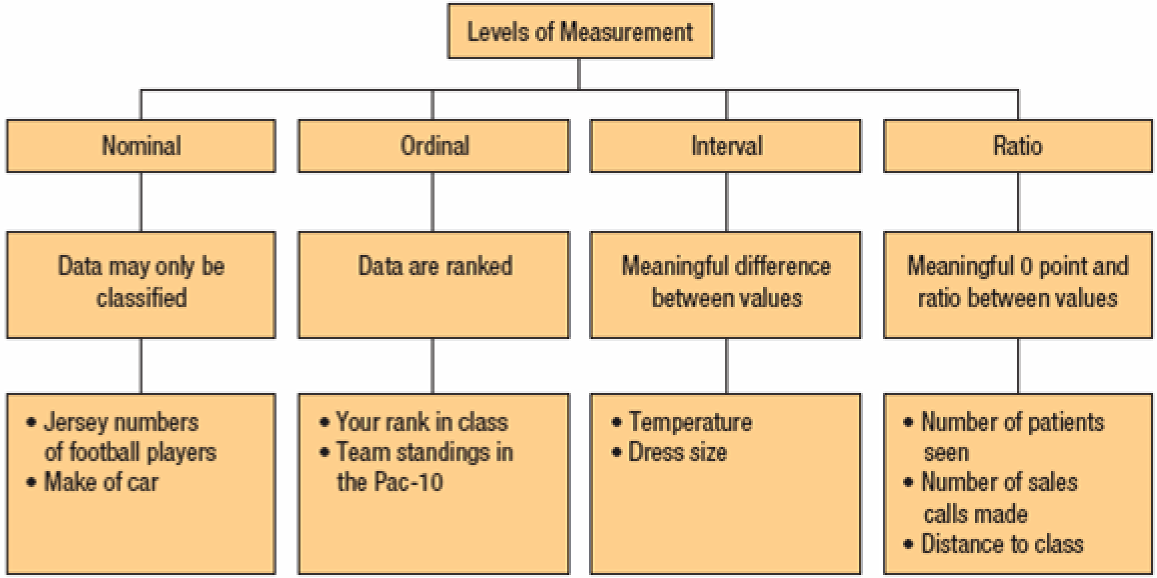


CHART 1–3 Summary of the Characteristics for Levels of Measurement

Describing Data:
Frequency Tables, Frequency Distributions, and Graphic Presentation

Chapter 2

Week 2

LEARNING OBJECTIVES

LO1 Make a frequency table for a set of data.
LO2 Organize data into a bar chart.
LO3 Present a set of data in a pie chart.
LO4 Create a frequency distribution for a data set.
LO5 Understand a relative frequency distribution.
LO6 Present data from a frequency distribution in a histogram or frequency polygon.

LO1 Make a frequency table for a set of data

Frequency Table

FREQUENCY TABLE A grouping of qualitative data into mutually exclusive classes showing the number of observations in each class.

Location	Number of Cars
Kane	52
Olean	40
Sheffield	45
Tionesta	43
Total	180

Bar Charts

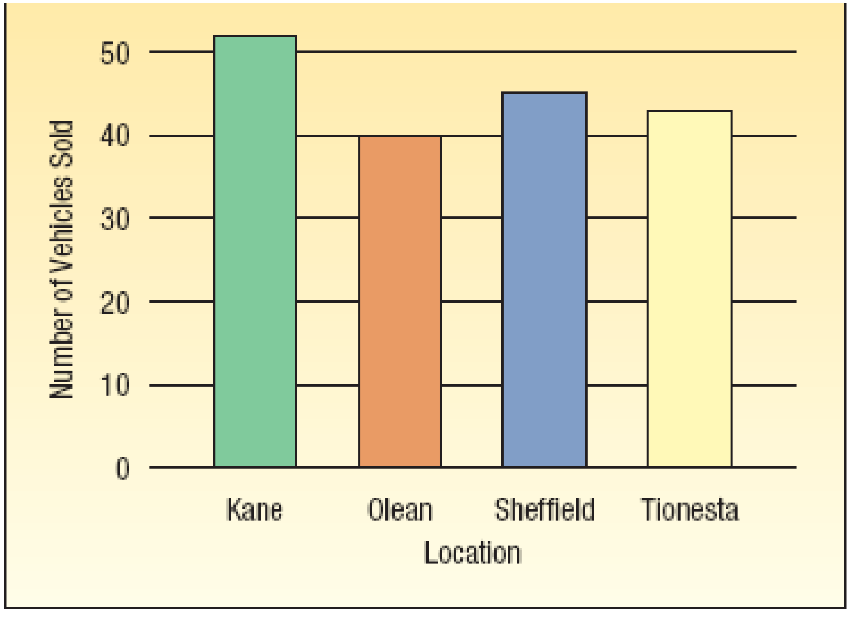
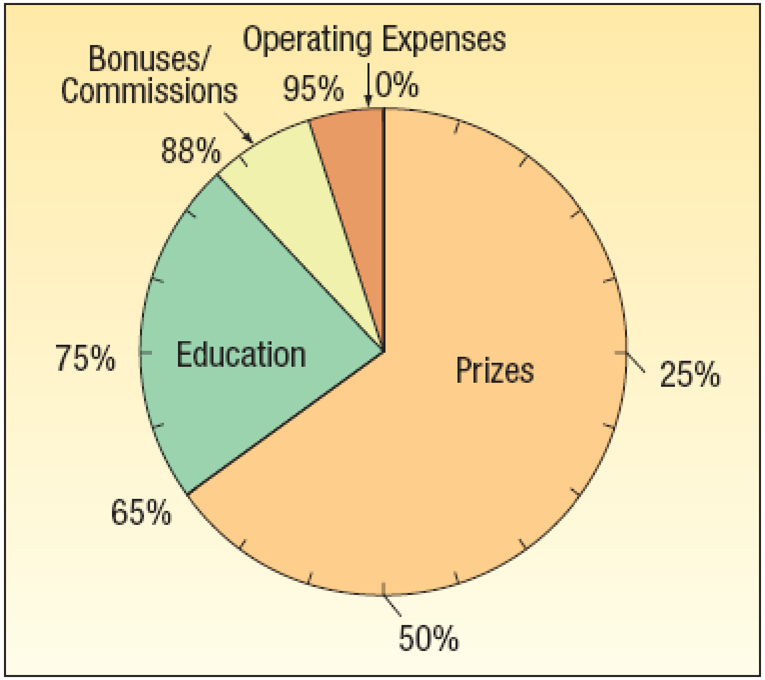


CHART 2–1 Number of Vehicles Sold by Location

BAR CHART A graph in which the classes are reported on the horizontal axis and the class frequencies on the vertical axis. The class frequencies are proportional to the heights of the bars.

Pie Charts

PIE CHART A chart that shows the proportion or percent that each class represents of the total number of frequencies.



LO4 Make a frequency table for a set of data

Frequency Distribution

FREQUENCY DISTRIBUTION A grouping of data into mutually exclusive classes showing the number of observations in each class.

Class interval: The class interval is obtained by subtracting the lower limit of a class from the lower limit of the next class.

Class frequency: The number of observations in each class.

Class midpoint: A point that divides a class into two equal parts. This is the average of the upper and lower class limits.

LO4 Create a frequency distribution for a data set.

EXAMPLE – Creating a Frequency Distribution Table

Applewood Auto Group wants to develop tables, charts, and graphs to show the typical selling price on various dealer lots. The table on the right reports only the price of the 180 vehicles sold last month.

TABLE 2–4 Profit on Vehicles Sold Last Month by the Applewood Auto Group

$1,387	$2,148	$2,201	$ 963	$ 820	$2,230	$3,043	$2,584	$2,370
1,754	2,207	996	1,298	1,266	2,341	1,059	2,666	2,637
1,817	2,252	2,813	1,410	1,741	3,292 (Highest)	1,674	2,991	1,426
1,040	1,428	323	1,553	1,772	1,108	1,807	934	2,944
1,273	1,889	352	1,648	1,932	1,295	2,056	2,063	2,147
1,529	1,166	482	2,071	2,350	1,344	2,236	2,083	1,973
3,082	1,320	1,144	2,116	2,422	1,906	2,928	2,856	2,502
1,951	2,265	1,485	1,500	2,446	1,952	1,269	2,989	783
2,692	1,323	1,509	1,549	369	2,070	1,717	910	1,538
1,206	1,761	1,638	2,348	978	2,454	1,797	1,536	2,339
1,342	1,919	1,961	2,498	1,238	1,606	1,955	1,957	2,700
443	2,357	2,127	294 (Lowest)	1,818	1,680	2,199	2,240	2,222
754	2,866	2,430	1,115	1,824	1,827	2,482	2,695	2,597
1,621	732	1,704	1,124	1,907	1,915	2,701	1,325	2,742
870	1,464	1,876	1,532	1,938	2,084	3,210	2,250	1,837
1,174	1,626	2,010	1,688	1,940	2,639	377	2,279	2,842
1,412	1,761	2,165	1,822	2,197	842	1,220	2,626	2,434
1,809	1,915	2,231	1,897	2,646	1,963	1,401	1,501	1,640
2,415	2,119	2,389	2,445	1,461	2,059	2,175	1,752	1,821
1,546	1,766	335	2,886	1,731	2,338	1,118	2,058	2,487

LO4

Constructing a Frequency Table - Example

- **Step 1: Decide on the number of classes.**
  A useful recipe to determine the number of classes (*k*) is the "2 to the *k* rule." such that $2^k > n$.

  There were 180 vehicles sold, so $n = 180$. If we try $k = 7$, then $2^7 = 128$, somewhat less than 180. Hence, 7 is not enough classes. If we let $k = 8$, then $2^8 = 256$, which is greater than 180. So the recommended number of classes is 8.

- **Step 2: Determine the class interval or width.**
  The formula is: $i \geq (H-L)/k$ where *i* is the class interval, *H* is the highest observed value, *L* is the lowest observed value, and *k* is the number of classes.

$$i \geq \frac{H - L}{k} = \frac{\$3{,}292 - \$294}{8} = \$374.75$$

Round up to some convenient number, such as a multiple of 10 or 100. Use a class width of $400

Constructing a Frequency Table - Example

- Step 3: Set the individual class limits

Classes
$ 200 up to $ 600
600 up to 1,000
1,000 up to 1,400
1,400 up to 1,800
1,800 up to 2,200
2,200 up to 2,600
2,600 up to 3,000
3,000 up to 3,400

- Step 4: Tally the vehicle selling prices into the classes.

TABLE 2–7 Frequency Distribution of Profit for Vehicles Sold Last Month at Applewood Auto Group

Profit	Frequency
$ 200 up to $ 600	𝍸 III
600 up to 1,000	𝍸 𝍸 I
1,000 up to 1,400	𝍸 𝍸 𝍸 𝍸 III
1,400 up to 1,800	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 III
1,800 up to 2,200	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸
2,200 up to 2,600	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 II
2,600 up to 3,000	𝍸 𝍸 𝍸 IIII
3,000 up to 3,400	IIII
Total	

- Step 5: Count the number of items in each class.

Profit	Frequency
$ 200 up to $ 600	8
600 up to 1,000	11
1,000 up to 1,400	23
1,400 up to 1,800	38
1,800 up to 2,200	45
2,200 up to 2,600	32
2,600 up to 3,000	19
3,000 up to 3,400	4
Total	180

LO5 Understand a relative frequency distribution.

Relative Class Frequencies

- Class frequencies can be converted to **relative class frequencies** to show the fraction of the total number of observations in each class.
- A relative frequency captures the relationship between a class total and the total number of observations.

TABLE 2–2 Relative Frequency Table of Vehicles Sold By Type At Whitner Autoplex Last Month

Vehicle Type	Number Sold	Relative Frequency
Domestic	50	0.625
Foreign	30	0.375
Total	80	1.000

Week 3

Relative Frequency Distribution

To convert a frequency distribution to a *relative* frequency distribution, each of the class frequencies is divided by the total number of observations.

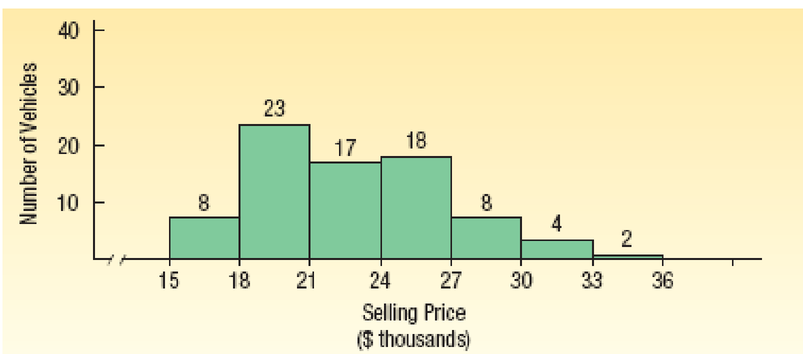
TABLE 2–8 Relative Frequency Distribution of Profit for Vehicles Sold Last Month at Applewood Auto Group

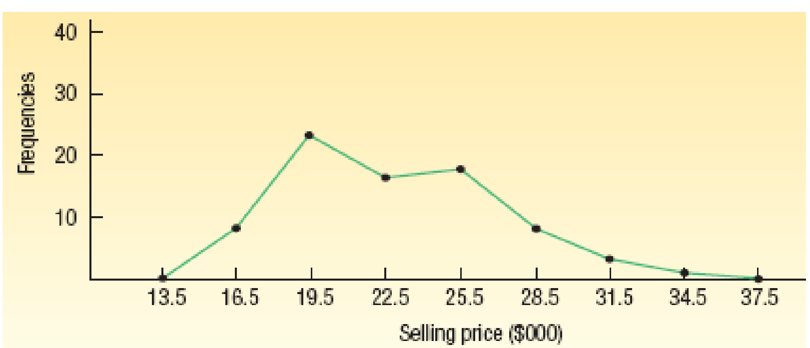
Profit	Frequency	Relative Frequency	Found by
$ 200 up to $ 600	8	.044	8/180
600 up to 1,000	11	.061	11/180
1,000 up to 1,400	23	.128	23/180
1,400 up to 1,800	38	.211	38/180
1,800 up to 2,200	45	.250	45/180
2,200 up to 2,600	32	.178	32/180
2,600 up to 3,000	19	.106	19/180
3,000 up to 3,400	4	.022	4/180
Total	180	1.000	

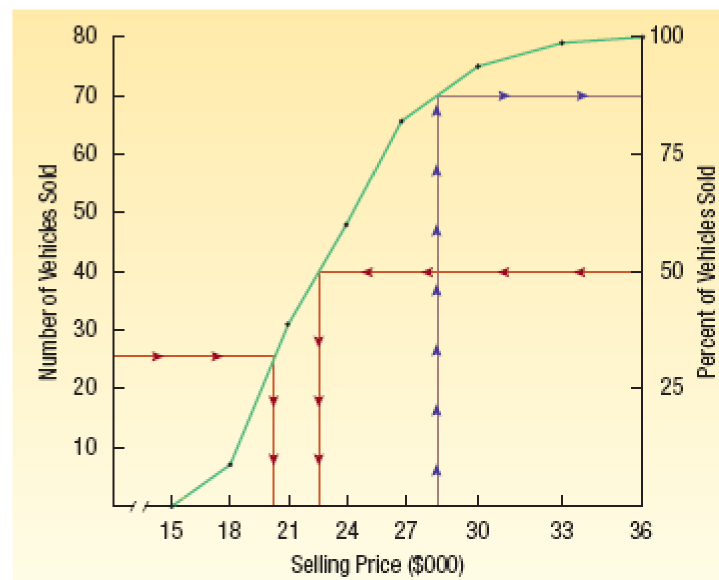
Graphic Presentation of a Frequency Distribution

The three commonly used graphic forms are:

- Histograms
- Frequency polygons
- Cumulative frequency distributions







Histogram

HISTOGRAM A graph in which the classes are marked on the horizontal axis and the class frequencies on the vertical axis. The class frequencies are represented by the heights of the bars and the bars are drawn adjacent to each other.

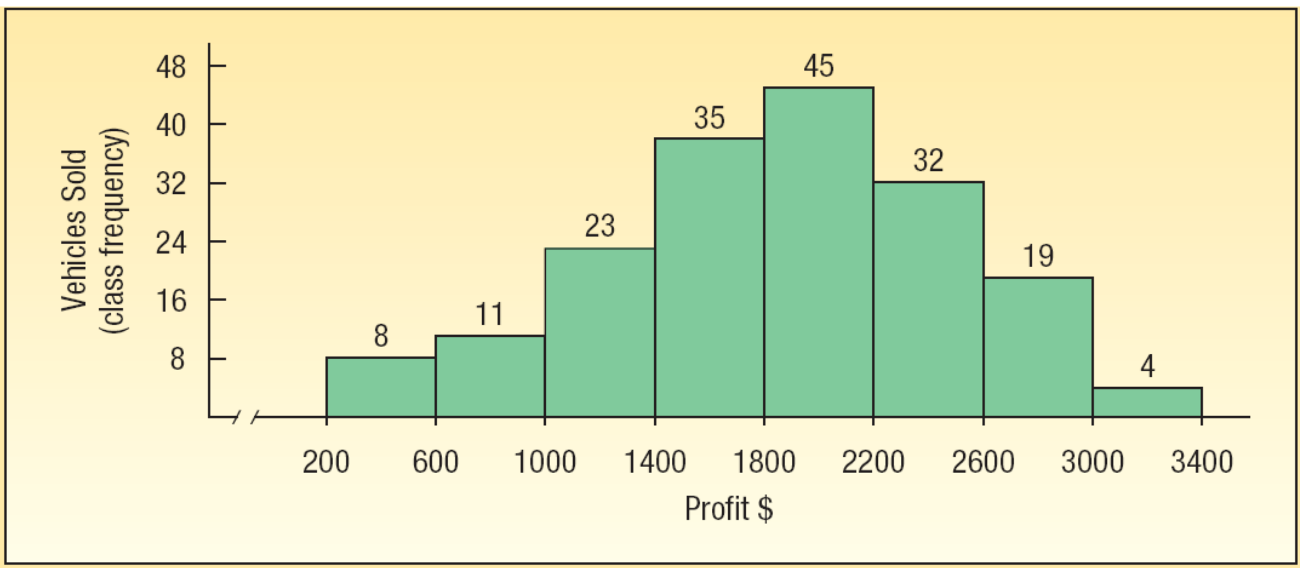


CHART 2–4 Histogram of the Profit on 180 Vehicles Sold at the Applewood Auto Group

LO6

Frequency Polygon

- A **frequency polygon** also shows the shape of a distribution and is similar to a histogram.
- It consists of line segments connecting the points formed by the intersections of the class midpoints and the class frequencies.

Profit	Midpoint	Frequency
\$ 200 up to \$ 600	\$ 400	8
600 up to 1,000	800	11
1,000 up to 1,400	1,200	23
1,400 up to 1,800	1,600	38
1,800 up to 2,200	2,000	45
2,200 up to 2,600	2,400	32
2,600 up to 3,000	2,800	19
3,000 up to 3,400	3,200	4
Total		180

LO6

Histogram Versus Frequency Polygon

- Both provide a quick picture of the main characteristics of the data (highs, lows, points of concentration, etc.)
- The histogram has the advantage of depicting each class as a rectangle, with the height of the rectangular bar representing the number in each class.
- The frequency polygon has an advantage over the histogram. It allows us to compare directly two or more frequency distributions.

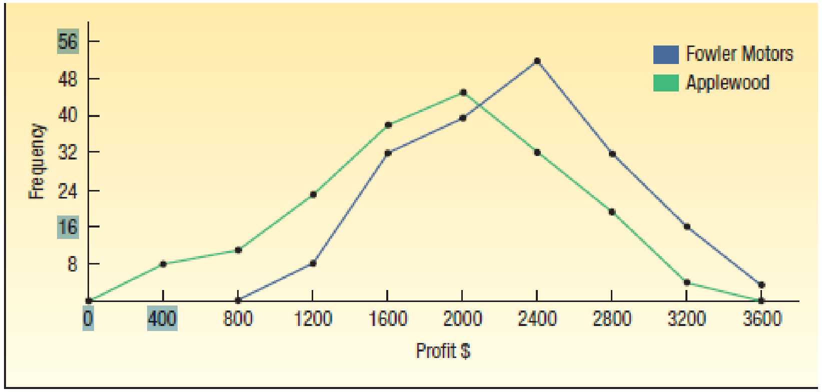
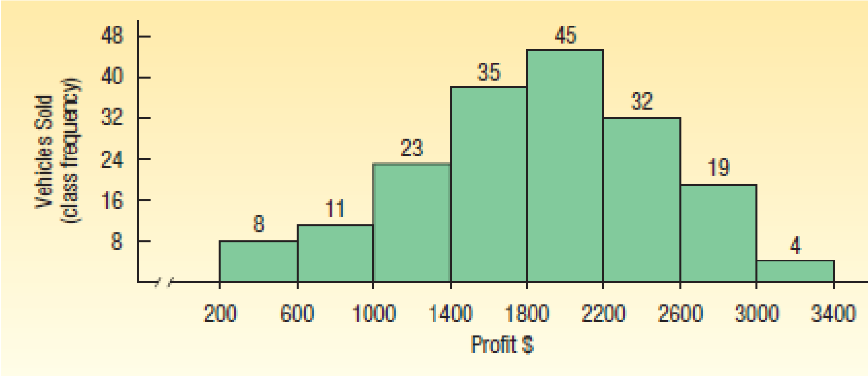


CHART 2–6 Distribution of Profit at Applewood Auto Group and Fowler Motors



LO6

Cumulative Frequency Distribution

TABLE 2–9 Cumulative Frequency Distribution for Profit on Vehicles Sold Last Month at Applewood Auto Group

Profit	Frequency	Cumulative Frequency	Found by
$ 200 up to $ 600	8	8	8
600 up to 1,000	11	19	8 + 11
1,000 up to 1,400	23	42	8 + 11 + 23
1,400 up to 1,800	38	80	8 + 11 + 23 + 30
1,800 up to 2,200	45	125	8 + 11 + 23 + 30 + 45
2,200 up to 2,600	32	157	8 + 11 + 23 + 30 + 45 + 32
2,600 up to 3,000	19	176	8 + 11 + 23 + 30 + 45 + 32 + 19
3,000 up to 3,400	4	180	8 + 11 + 23 + 30 + 45 + 32 + 19 + 4
Total	180		

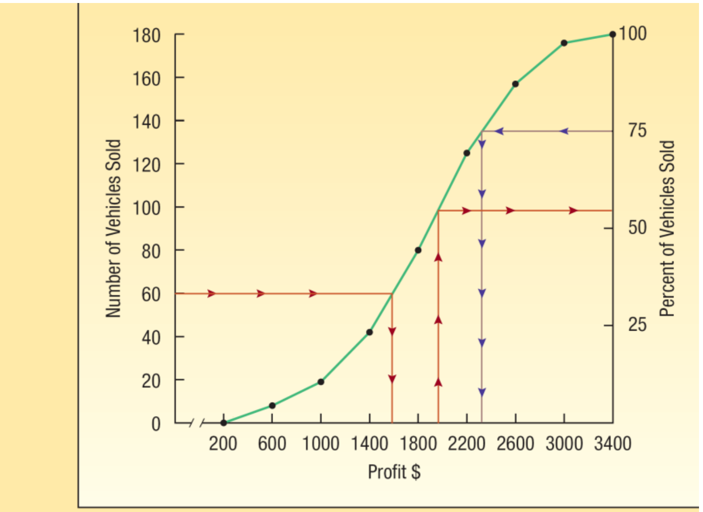


CHART 2–7 Cumulative Frequency Distribution for Vehicle Profit

Describing Data:
Numerical Measures

Chapter 3

McGraw-Hill/Irwin

Week 4

GOALS

- Calculate the arithmetic mean, weighted mean, median, mode, and geometric mean.
- Explain the characteristics, uses, advantages, and disadvantages of each measure of location.
- Identify the position of the mean, median, and mode for both symmetric and skewed distributions.
- Compute and interpret the range, mean deviation, variance, and standard deviation.
- Understand the characteristics, uses, advantages, and disadvantages of each measure of dispersion.
- Understand Chebyshev’s theorem and the Empirical Rule as they relate to a set of observations.

Characteristics of the Mean

The arithmetic mean is the most widely used measure of location. It requires the interval scale. Its major characteristics are:

- All values are used.
- It is unique.
- The sum of the deviations from the mean is 0.
- It is calculated by summing the values and dividing by the number of values.

Population Mean

For ungrouped data, the population mean is the sum of all the population values divided by the total number of population values:

POPULATION MEAN $$\mu = \frac{\Sigma X}{N}$$ **[3–1]**

where:

- μ represents the population mean. It is the Greek lowercase letter "mu."
- N is the number of values in the population.
- X represents any particular value.
- Σ is the Greek capital letter "sigma" and indicates the operation of adding.
- ΣX is the sum of the X values in the population.

EXAMPLE – Population Mean

There are 12 automobile manufacturing companies in the United States. Listed below is the number of patents granted by the United States government to each company in a recent year.

Company	Number of Patents Granted	Company	Number of Patents Granted
General Motors	511	Mazda	210
Nissan	385	Chrysler	97
DaimlerChrysler	275	Porsche	50
Toyota	257	Mitsubishi	36
Honda	249	Volvo	23
Ford	234	BMW	13

Is this information a sample or a population? What is the arithmetic mean number of patents granted?

$$\mu = \frac{\sum X}{N} = \frac{511 + 385 + 275 + \ldots + 36 + 23 + 13}{12} = \frac{2340}{12} = 195$$

Sample Mean

- For ungrouped data, the sample mean is the sum of all the sample values divided by the number of sample values:

SAMPLE MEAN

$$\overline{X} = \frac{\Sigma X}{n} \qquad [3\text{–}2]$$

where:

$\overline{X}$ is the sample mean. It is read "X bar."

n is the number of values in the sample.

EXAMPLE – Sample Mean

SunCom is studying the number of minutes used monthly by clients in a particular cell phone rate plan. A random sample of 12 clients showed the following number of minutes used last month.

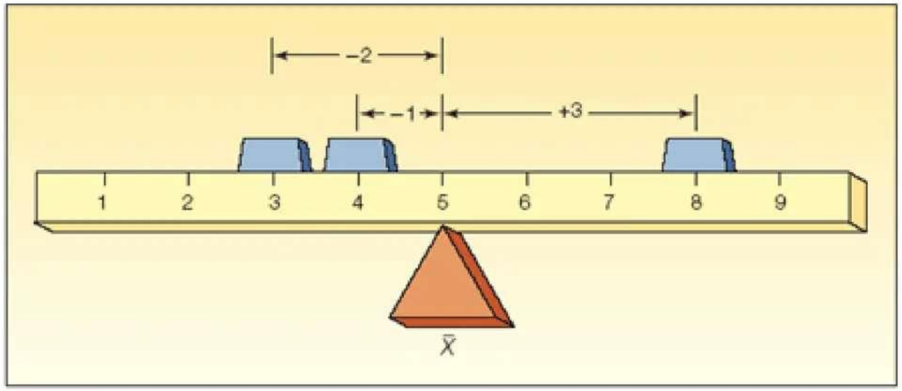
90	77	94	89	119	112
91	110	92	100	113	83

What is the arithmetic mean number of minutes used?

$$\bar{X} = \frac{\Sigma X}{n} = \frac{99+77+94+\ldots+100+113+83}{12} = \frac{1{,}170}{12} = 97.5$$

Properties of the Arithmetic Mean

- Every set of interval-level and ratio-level data has a mean.
- All the values (observations) are included in computing the mean.
- A set of data (observations) has a unique mean.
- The mean is affected by unusually large or small data values (observations).
- The arithmetic mean is the only measure of central tendency where the sum of the deviations of each value from the mean is zero.



Weighted Mean

- The weighted mean of a set of numbers X_1, X_2, ..., X_n, with corresponding weights w_1, w_2, ...,w_n, is computed from the following formula:

WEIGHTED MEAN

$$\bar{X}_w = \frac{w_1X_1 + w_2X_2 + w_3X_3 + \cdots + w_nX_n}{w_1 + w_2 + w_3 + \cdots + w_n} \quad [3\text{–}3]$$

EXAMPLE – Weighted Mean

The Carter Construction Company pays its hourly employees $16.50, $19.00, or $25.00 per hour. There are 26 hourly employees, 14 of which are paid at the $16.50 rate, 10 at the $19.00 rate, and 2 at the $25.00 rate. What is the mean hourly rate paid the 26 employees?

$$\bar{X}_w = \frac{14(\$16.50) + 10(\$19.00) + 2(\$26.00)}{14 + 10 + 2}$$

$$= \frac{\$471.00}{26} = \$18.1154$$

Week 5

The Median

- The Median is the midpoint of the values after they have been ordered from the smallest to the largest.
  - There are as many values above the median as below it in the data array.
  - For an even set of values, the median will be the arithmetic average of the two middle numbers.

Properties of the Median

- There is a unique median for each data set.
- It is not affected by extremely large or small values and is therefore a valuable measure of central tendency when such values occur.
- It can be computed for ratio-level, interval-level, and ordinal-level data.
- It can be computed for an open-ended frequency distribution if the median does not lie in an open-ended class.

EXAMPLES - Median

The ages for a sample of five college students are:

21, 25, 19, 20, 22

Arranging the data in ascending order gives:

19, 20, 21, 22, 25.

Thus the median is 21.

The heights of four basketball players, in inches, are:

76, 73, 80, 75

Arranging the data in ascending order gives:

73, 75, 76, 80.

Thus the median is 75.5

The Mode

- The mode is the value of the observation that appears most frequently.

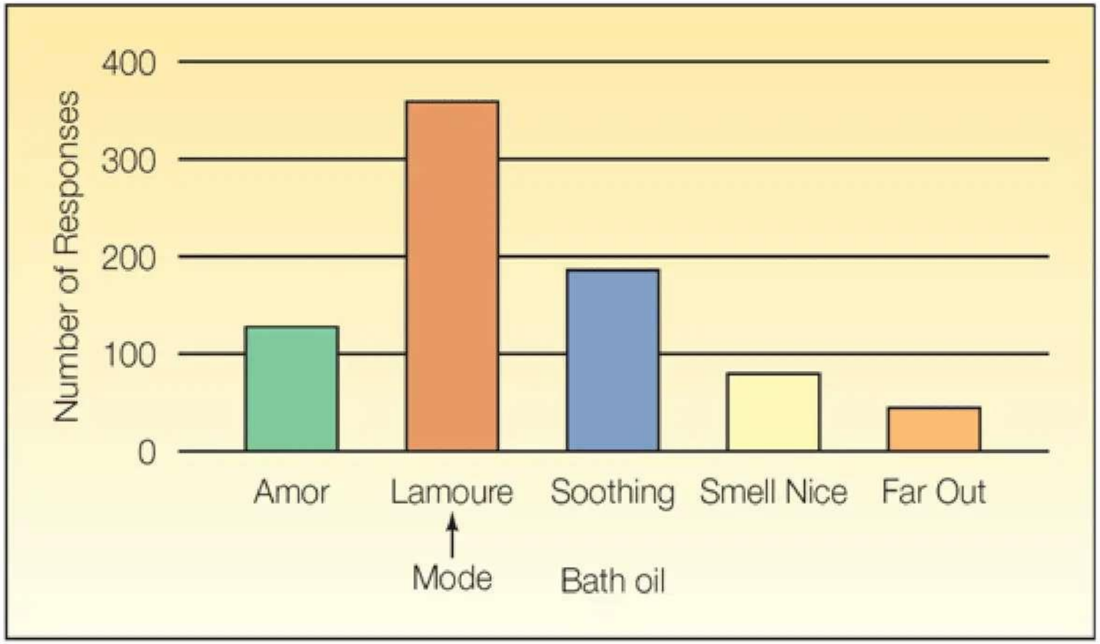


CHART 3–1 Number of Respondents Favoring Various Bath Oils

Example - Mode

The annual salaries of quality-control managers in selected states are shown below. What is the modal annual salary?

State	Salary	State	Salary	State	Salary
Arizona	$35,000	Illinois	$58,000	Ohio	$50,000
California	49,100	Louisiana	60,000	Tennessee	60,000
Colorado	60,000	Maryland	60,000	Texas	71,400
Florida	60,000	Massachusetts	40,000	West Virginia	60,000
Idaho	40,000	New Jersey	65,000	Wyoming	55,000

A perusal of the salaries reveals that the annual salary of $60,000 appears more often (six times) than any other salary. The mode is, therefore, $60,000.

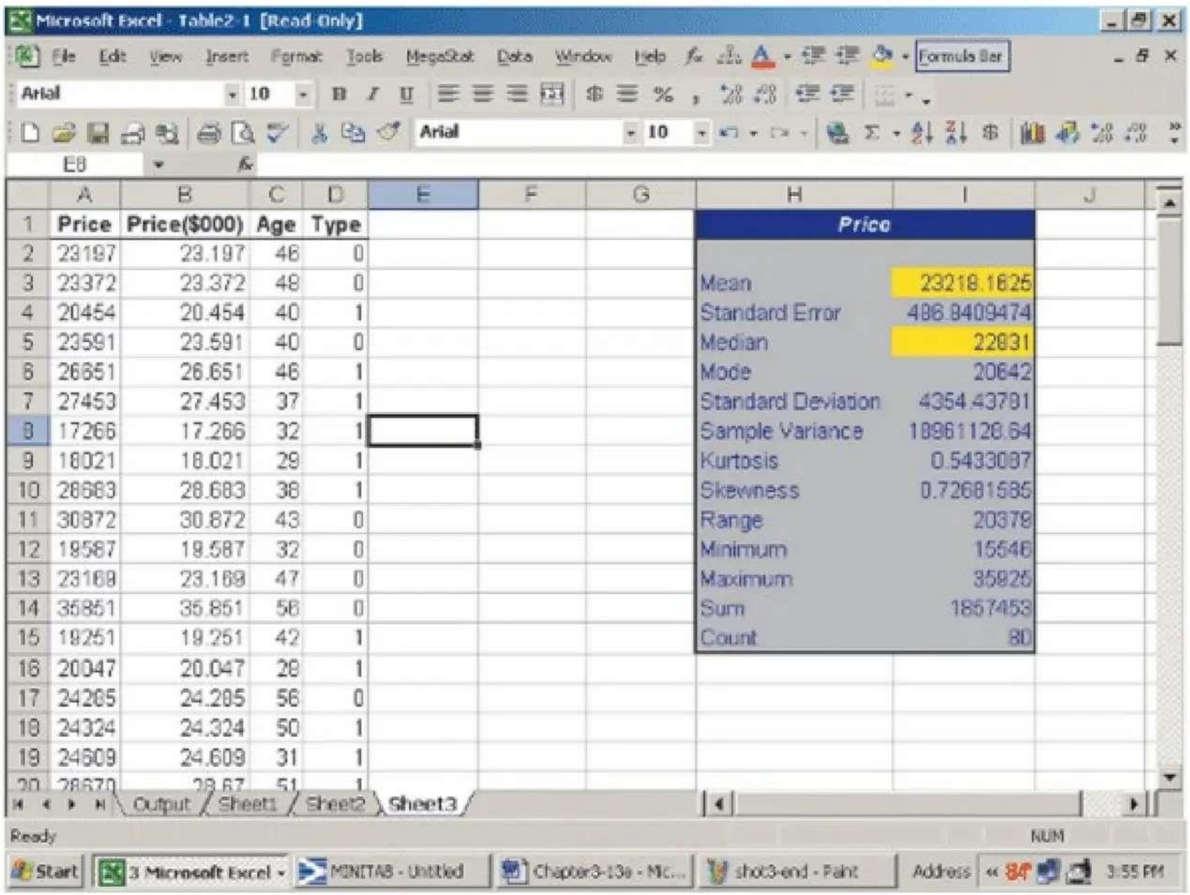
Mean, Median, Mode Using Excel

Table 2–4 in Chapter 2 shows the prices of the 80 vehicles sold last month at Whitner Autoplex in Raytown, Missouri. Determine the mean and the median selling price. The mean and the median selling prices are reported in the following Excel output. There are 80 vehicles in the study. So the calculations with a calculator would be tedious and prone to error.

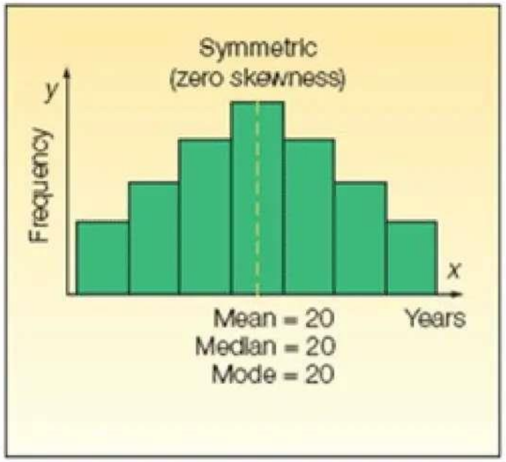
TABLE 2–4 Prices of Vehicles Sold Last Month at Whitner Autoplex

$23,197	$23,372	$20,454	$23,591	$26,651	$27,453	$17,266
18,021	28,683	30,872	19,587	23,169	35,851	19,251
20,047	24,285	24,324	24,609	28,670	15,546 (Lowest)	15,935
19,873	25,251	25,277	28,034	24,533	27,443	19,889
20,004	17,357	20,155	19,688	23,657	26,613	20,895
20,203	23,765	25,783	26,661	32,277	20,642	21,981
24,052	25,799	15,794	18,263	35,925 (Highest)	17,399	17,968
20,356	21,442	21,722	19,331	22,817	19,766	20,633
20,962	22,845	26,285	27,896	29,076	32,492	18,890
21,740	22,374	24,571	25,449	28,337	20,642	23,613
24,220	30,655	22,442	17,891	20,818	26,237	20,445
21,556	21,639	24,296				

Mean, Median, Mode Using Excel

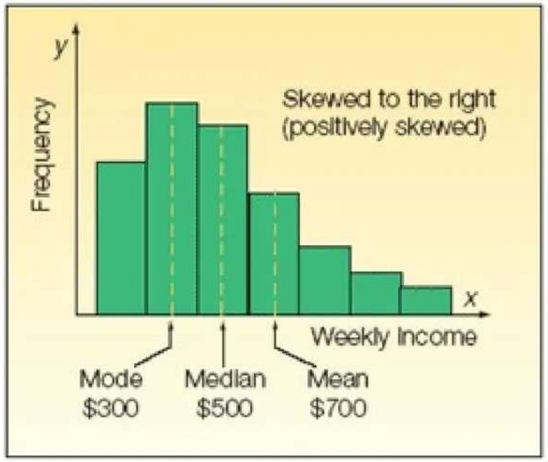


The Relative Positions of the Mean, Median and the Mode



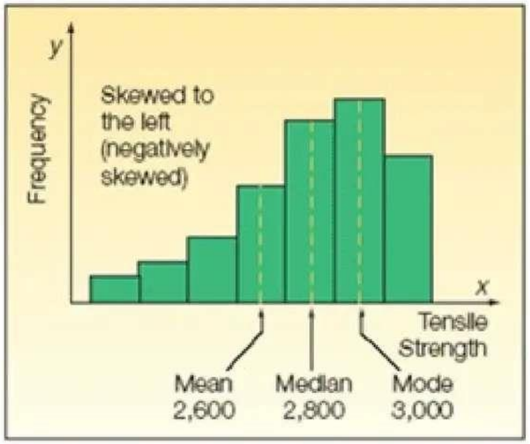
zero skewness

mode = median = mean



positive skewness

mode < median < mean



negative skewness

mode > median > mean

The Geometric Mean

- Useful in finding the average change of percentages, ratios, indexes, or growth rates over time.
- It has a wide application in business and economics because we are often interested in finding the percentage changes in sales, salaries, or economic figures, such as the GDP, which compound or build on each other.
- The geometric mean will always be less than or equal to the arithmetic mean.
- The geometric mean of a set of *n* positive numbers is defined as the *n*th root of the product of *n* values.
- The formula for the geometric mean is written:

GEOMETRIC MEAN $$GM = \sqrt[n]{(X_1)(X_2)\cdots(X_n)}$$ **[3–4]**

EXAMPLE – Geometric Mean

Suppose you receive a 5 percent increase in salary this year and a 15 percent increase next year. The average annual percent increase is 9.886, not 10.0. Why is this so? We begin by calculating the geometric mean.

$$GM = \sqrt{(1.05)(1.15)} = 1.09886$$

EXAMPLE – Geometric Mean (2)

The return on investment earned by Atkins construction Company for four successive years was: 30 percent, 20 percent, 40 percent, and 200 percent. What is the geometric mean rate of return on investment?

$$GM = \sqrt[4]{(1.3)(1.2)(0.6)(3.0)} = \sqrt[4]{2.808} = 1.294$$

Dispersion

Why Study Dispersion?

- A measure of location, such as the mean or the median, only describes the center of the data. It is valuable from that standpoint, but it does not tell us anything about the spread of the data.
- For example, if your nature guide told you that the river ahead averaged 3 feet in depth, would you want to wade across on foot without additional information? Probably not. You would want to know something about the variation in the depth.
- A second reason for studying the dispersion in a set of data is to compare the spread in two or more distributions.

Samples of Dispersions

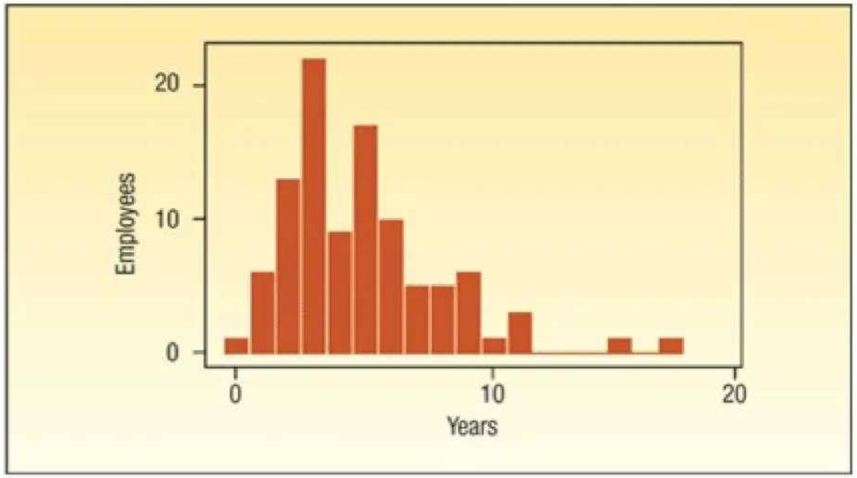


CHART 3–5 Histogram of Years of Employment at Hammond Iron Works, Inc.

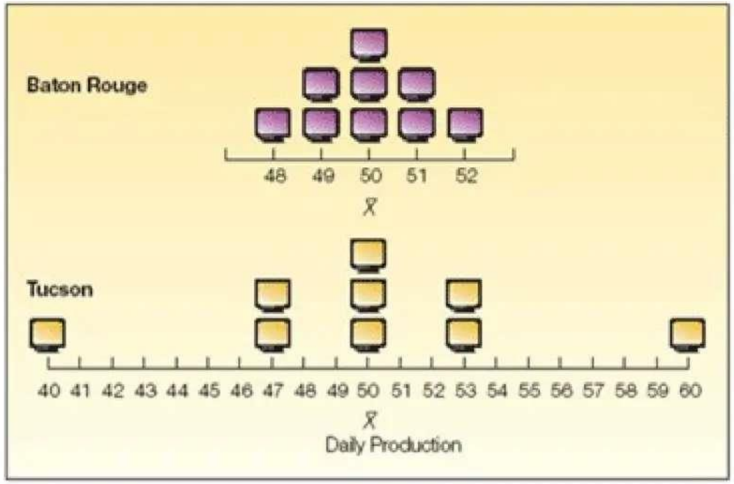


CHART 3–6 Hourly Production of Computer Monitors at the Baton Rouge and Tucson Plants

Measures of Dispersion

- Range

RANGE $$\text{Range} = \text{Largest value} - \text{Smallest value} \qquad [3\text{–}6]$$

- Mean Deviation

MEAN DEVIATION $$MD = \frac{\Sigma|X - \bar{X}|}{n} \qquad [3\text{–}7]$$

- Variance and Standard Deviation

POPULATION VARIANCE $$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N} \qquad [3\text{–}8]$$

POPULATION STANDARD DEVIATION $$\sigma = \sqrt{\frac{\Sigma(X - \mu)^2}{N}} \qquad [3\text{–}9]$$

EXAMPLE – Range

The number of cappuccinos sold at the Starbucks location in the Orange Country Airport between 4 and 7 p.m. for a sample of 5 days last year were 20, 40, 50, 60, and 80. Determine the range and mean deviation for the number of cappuccinos sold.

$$\text{Range} = \text{Largest} - \text{Smallest value}$$
$$= 80 - 20 = 60$$

Week 6

EXAMPLE – Mean Deviation

The number of cappuccinos sold at the Starbucks location in the Orange Country Airport between 4 and 7 p.m. for a sample of 5 days last year were 20, 40, 50, 60, and 80. Determine the mean deviation for the number of cappuccinos sold.

Number of Cappuccinos Sold Daily	$(X - \bar{X})$	Absolute Deviation
20	$(20 - 50) = -30$	30
40	$(40 - 50) = -10$	10
50	$(50 - 50) = 0$	0
60	$(60 - 50) = 10$	10
80	$(80 - 50) = 30$	30
		Total 80

$$MD = \frac{\Sigma|X - \bar{X}|}{n} = \frac{80}{5} = 16$$

EXAMPLE – Variance and Standard Deviation

The number of traffic citations issued during the last five months in Beaufort County, South Carolina, is 38, 26, 13, 41, and 22. What is the population variance?

Number (X)	$X - \mu$	$(X - \mu)^2$
38	+10	100
26	−2	4
13	−15	225
41	+13	169
22	−6	36
140	0*	534

$$\mu = \frac{\Sigma X}{N} = \frac{140}{5} = 28$$

$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N} = \frac{534}{5} = 106.8$$

EXAMPLE – Sample Variance

The hourly wages for a sample of part-time employees at Home Depot are: $12, $20, $16, $18, and $19. What is the sample variance?

SAMPLE VARIANCE $$s^2 = \frac{\Sigma(X - \overline{X})^2}{n - 1}$$ [3–10]

Hourly Wage (X)	$X - \overline{X}$	$(X - \overline{X})^2$
$12	−$5	25
20	3	9
16	−1	1
18	1	1
19	2	4
$85	0	40

$$s^2 = \frac{\Sigma(X - \overline{X})^2}{n - 1} = \frac{40}{5 - 1}$$

$$= 10 \text{ in dollars squared}$$

Chebyshev's Theorem

The arithmetic mean biweekly amount contributed by the Dupree Paint employees to the company's profit-sharing plan is \$51.54, and the standard deviation is \$7.51. At least what percent of the contributions lie within plus 3.5 standard deviations and minus 3.5 standard deviations of the mean?

CHEBYSHEV'S THEOREM For any set of observations (sample or population), the proportion of the values that lie within k standard deviations of the mean is at least $1 - 1/k^2$, where k is any constant greater than 1.

$$1 - \frac{1}{k^2} = 1 - \frac{1}{(3.5)^2} = 1 - \frac{1}{12.25} = 0.92$$

The Empirical Rule

EMPIRICAL RULE For a symmetrical, bell-shaped frequency distribution, approximately 68 percent of the observations will lie within plus and minus one standard deviation of the mean; about 95 percent of the observations will lie within plus and minus two standard deviations of the mean; and practically all (99.7 percent) will lie within plus and minus three standard deviations of the mean.

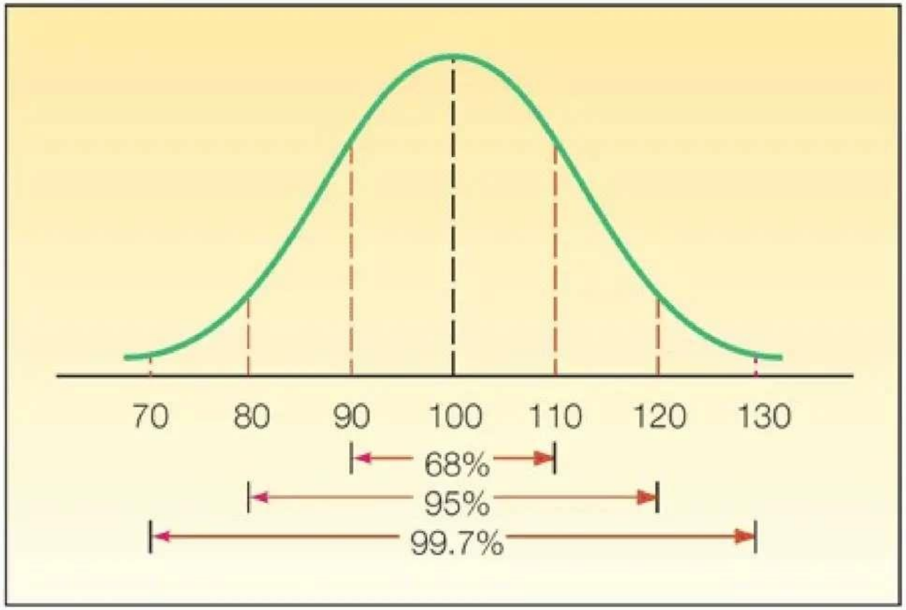


CHART 3–7 A Symmetrical, Bell-Shaped Curve Showing the Relationships between the Standard Deviation and the Observations

The Arithmetic Mean of Grouped Data

ARITHMETIC MEAN OF GROUPED DATA $$\overline{X} = \frac{\Sigma fM}{n}$$ [3–12]

where:

- $\overline{X}$ is the designation for the sample mean.
- M is the midpoint of each class.
- f is the frequency in each class.
- fM is the frequency in each class times the midpoint of the class.
- ΣfM is the sum of these products.
- n is the total number of frequencies.

The Arithmetic Mean of Grouped Data - Example

Recall in Chapter 2, we constructed a frequency distribution for the vehicle selling prices. The information is repeated below. Determine the arithmetic mean vehicle selling price.

Selling Prices ($ thousands)	Frequency
15 up to 18	8
18 up to 21	23
21 up to 24	17
24 up to 27	18
27 up to 30	8
30 up to 33	4
33 up to 36	2
Total	80

The Arithmetic Mean of Grouped Data - Example

Selling Price ($ thousands)	Frequency (f)	Midpoint (M)	fM
15 up to 18	8	$16.5	$ 132.0
18 up to 21	23	19.5	448.5
21 up to 24	17	22.5	382.5
24 up to 27	18	25.5	459.0
27 up to 30	8	28.5	228.0
30 up to 33	4	31.5	126.0
33 up to 36	2	34.5	69.0
Total	80		$1,845.0

Solving for the arithmetic mean using formula (3–12), we get:

$$\overline{X} = \frac{\Sigma fM}{n} = \frac{\$1,845}{80} = \$23.1 \text{ (thousands)}$$

Standard Deviation of Grouped Data

STANDARD DEVIATION, GROUPED DATA

$$s = \sqrt{\frac{\Sigma f(M - \overline{X})^2}{n - 1}} \qquad [3\text{–}13]$$

where:

- s is the symbol for the sample standard deviation.
- M is the midpoint of the class.
- f is the class frequency.
- n is the number of observations in the sample.
- $\overline{X}$ is the designation for the sample mean.

Standard Deviation of Grouped Data - Example

Refer to the frequency distribution for the Whitner Autoplex data used earlier. Compute the standard deviation of the vehicle selling prices

Selling Price ($ thousands)	Frequency (f)	Midpoint (M)	$(M - \bar{X})$	$(M - \bar{X})^2$	$f(M - \bar{X})^2$
15 up to 18	8	16.5	−6.6	43.56	348.48
18 up to 21	23	19.5	−3.6	12.96	298.08
21 up to 24	17	22.5	−0.6	0.36	6.12
24 up to 27	18	25.5	2.4	5.76	103.68
27 up to 30	8	28.5	5.4	29.16	233.28
30 up to 33	4	31.5	8.4	70.56	282.24
33 up to 36	2	34.5	11.4	129.96	259.92
	80				1,531.80

$$s = \sqrt{\frac{\Sigma f(M - \bar{X})^2}{n - 1}} = \sqrt{\frac{1531.8}{80 - 1}} = 4.403.$$

Week 7

CORRELATION & REGRESSION

Birinder Singh, Assistant Professor, PCTE

CORRELATION

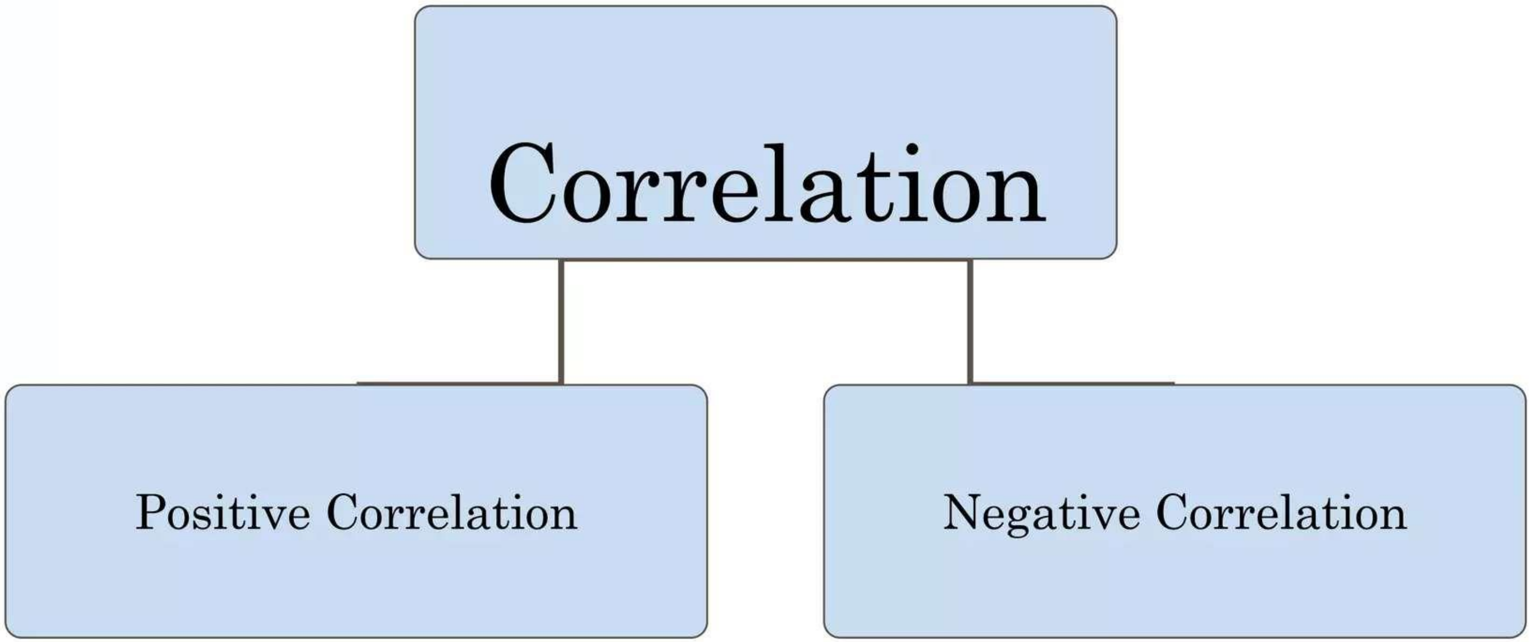
- Correlation is a statistical tool that helps to measure and analyze the degree of relationship between two variables.
- Correlation analysis deals with the association between two or more variables.

Birinder Singh, Assistant Professor, PCTE

CORRELATION

- The degree of relationship between the variables under consideration is measure through the correlation analysis.
- The measure of correlation called the correlation coefficient .
- The degree of relationship is expressed by coefficient which range from correlation
(-1 ≤ r ≥ +1)
- The direction of change is indicated by a sign.
- The correlation analysis enable us to have an idea about the degree & direction of the relationship between the two variables under study.

Types of Correlation - Type I



Types of Correlation Type I

- **Positive Correlation:** The correlation is said to be positive correlation if the values of two variables changing with same direction.

  Ex. Pub. Exp. & Sales, Height & Weight.

- **Negative Correlation:** The correlation is said to be negative correlation when the values of variables change with opposite direction.

  Ex. Price & Quantity demanded.

Direction of the Correlation

- **Positive relationship** – Variables change in the same direction.
  - As X is increasing, Y is increasing
  - As X is decreasing, Y is decreasing
  - E.g., As height increases, so does weight.

Indicated by sign; (+) or (-).

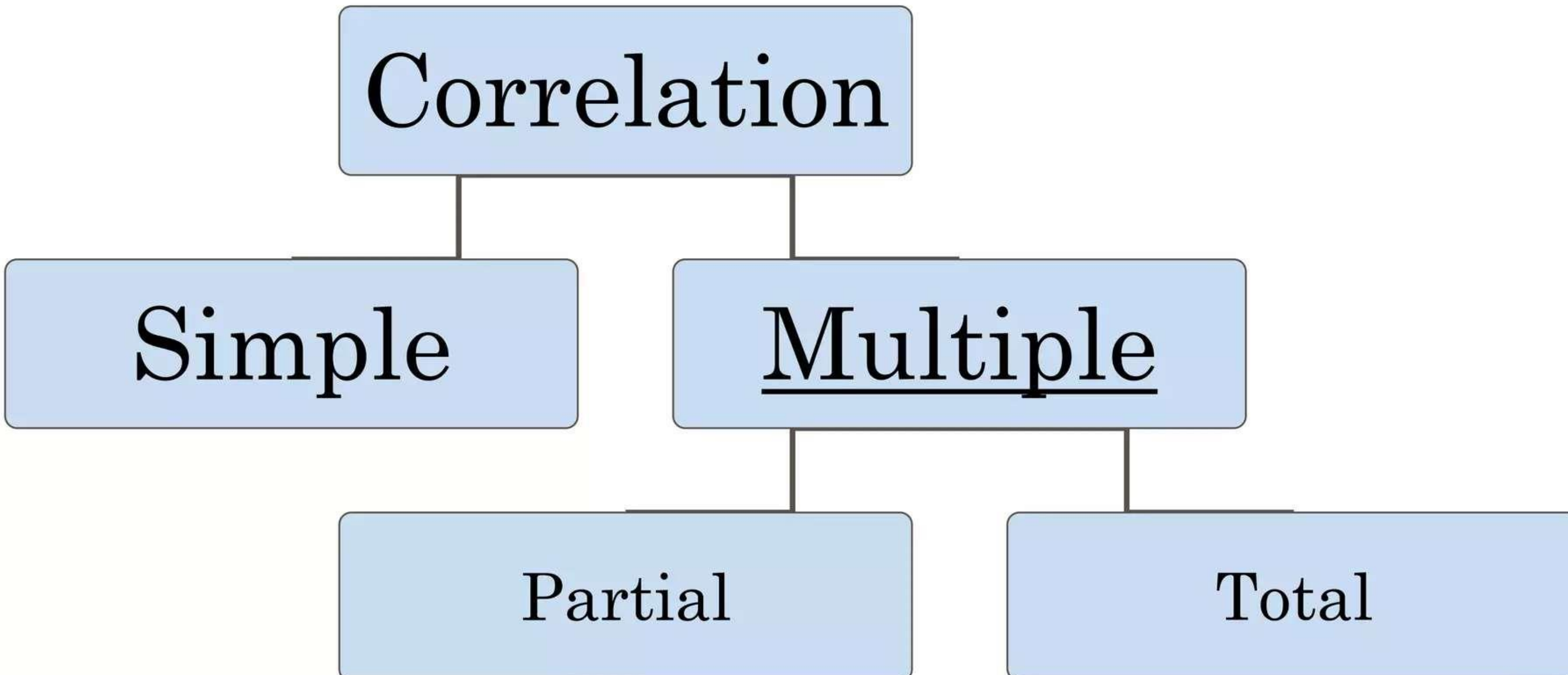
- **Negative relationship** – Variables change in opposite directions.
  - As X is increasing, Y is decreasing
  - As X is decreasing, Y is increasing
  - E.g., As TV time increases, grades decrease

EXAMPLES

Positive Correlation	Negative Correlation
Water consumption and temperature.	Alcohol consumption and driving ability.
Study time and grades.	Price & quantity demanded

Birinder Singh, Assistant Professor, PCTE

Types of Correlation Type II

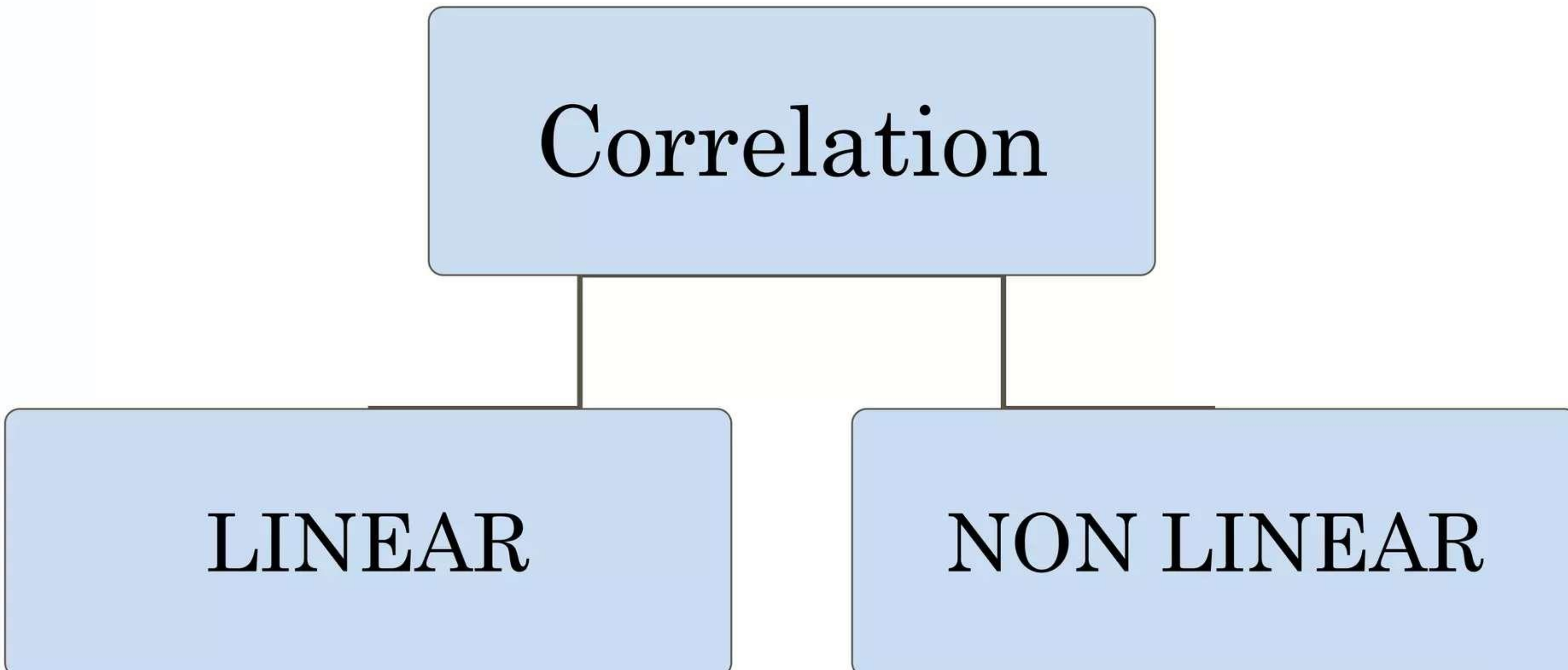


Week 8

Types of Correlation Type II

- **Simple correlation:** Under simple correlation problem there are only two variables are studied.
- **Multiple Correlation:** Under Multiple Correlation three or more than three variables are studied. Ex. $Q_d = f(P, P_C, P_S, t, y)$
- **Partial correlation:** analysis recognizes more than two variables but considers only two variables keeping the other constant.
- **Total correlation:** is based on all the relevant variables, which is normally not feasible.

Types of Correlation
Type III



Types of Correlation Type III

- **Linear correlation:** Correlation is said to be linear when the amount of change in one variable tends to bear a constant ratio to the amount of change in the other. The graph of the variables having a linear relationship will form a straight line.

  Ex X = 1, 2, 3, 4, 5, 6, 7, 8,

  Y = 5, 7, 9, 11, 13, 15, 17, 19,

  $Y = 3 + 2x$

- **Non Linear correlation:** The correlation would be non linear if the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable.

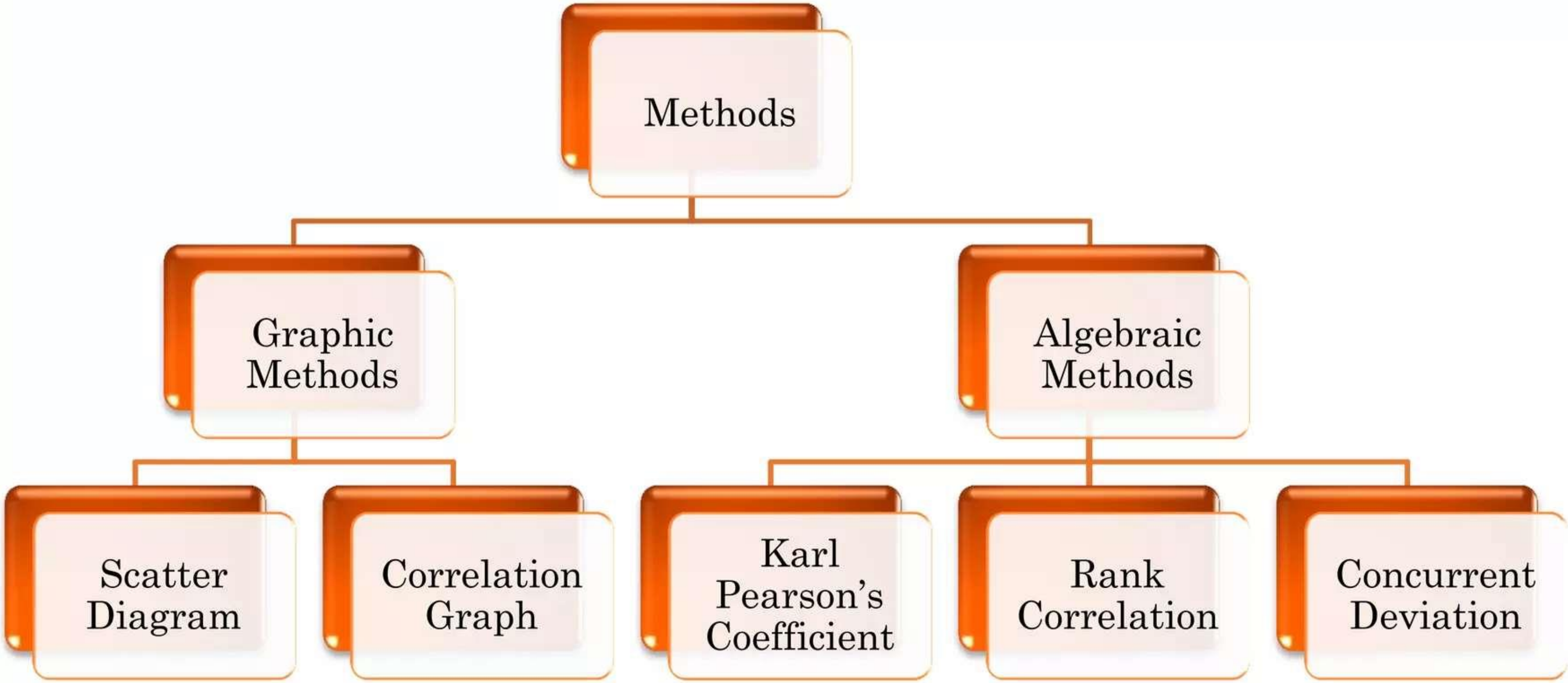
Correlation & Causation

- Causation means cause & effect relation.
- Correlation denotes the interdependency among the variables for correlating two phenomenon, it is essential that the two phenomenon should have cause-effect relationship,& if such relationship does not exist then the two phenomenon can not be correlated.
- If two variables vary in such a way that movement in one are accompanied by movement in other, these variables are called cause and effect relationship.
- Causation always implies correlation but correlation does not necessarily implies causation.

Degree of Correlation

- Perfect Correlation
- High Degree of Correlation
- Moderate Degree of Correlation
- Low Degree of Correlation
- No Correlation

METHODS OF STUDYING CORRELATION



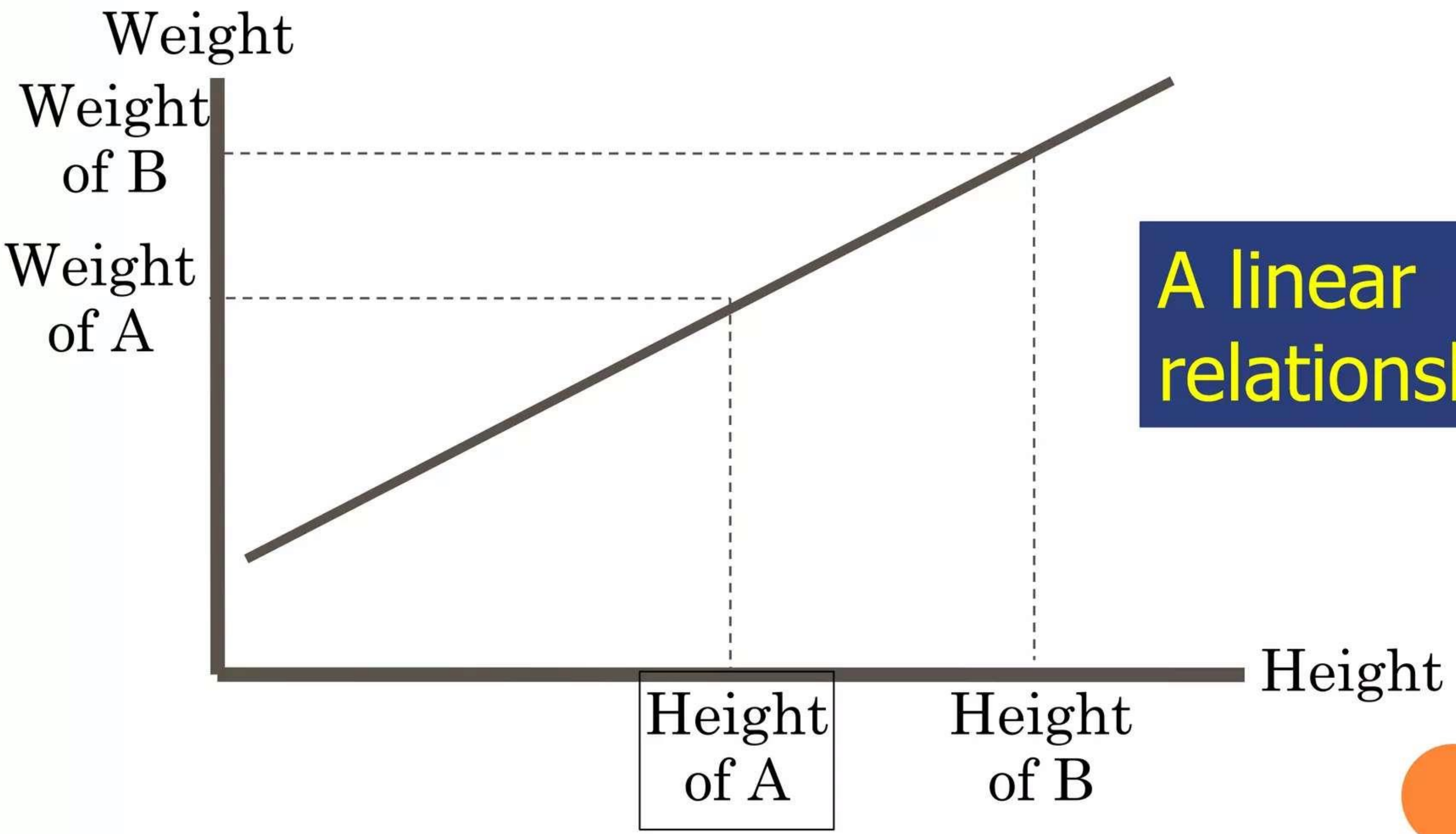
Birinder Singh, Assistant Professor, PCTE

Week 9

SCATTER DIAGRAM METHOD

- Scatter Diagram is a graph of observed plotted points where each points represents the values of X & Y as a coordinate.
- It portrays the relationship between these two variables graphically.

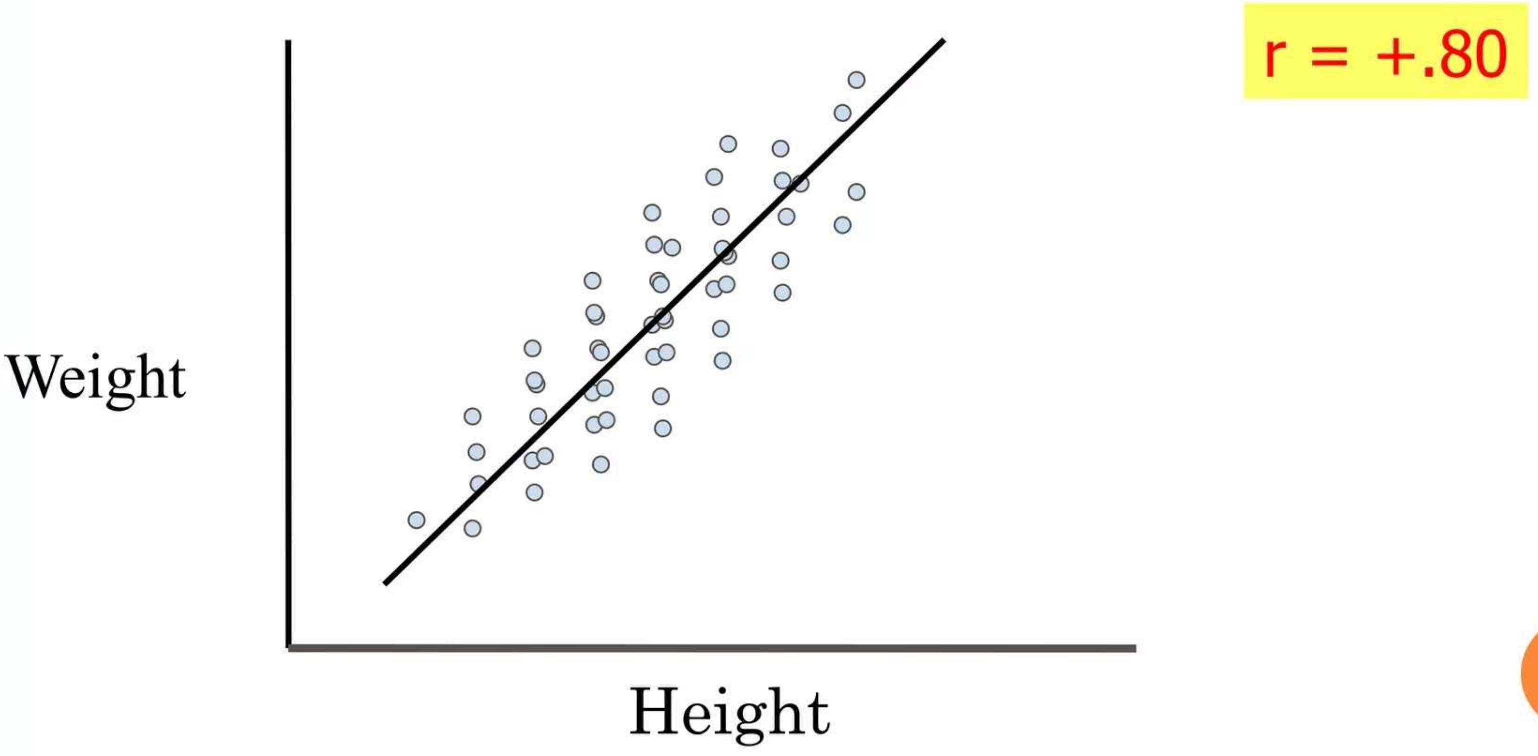
A perfect positive correlation



A linear relationship

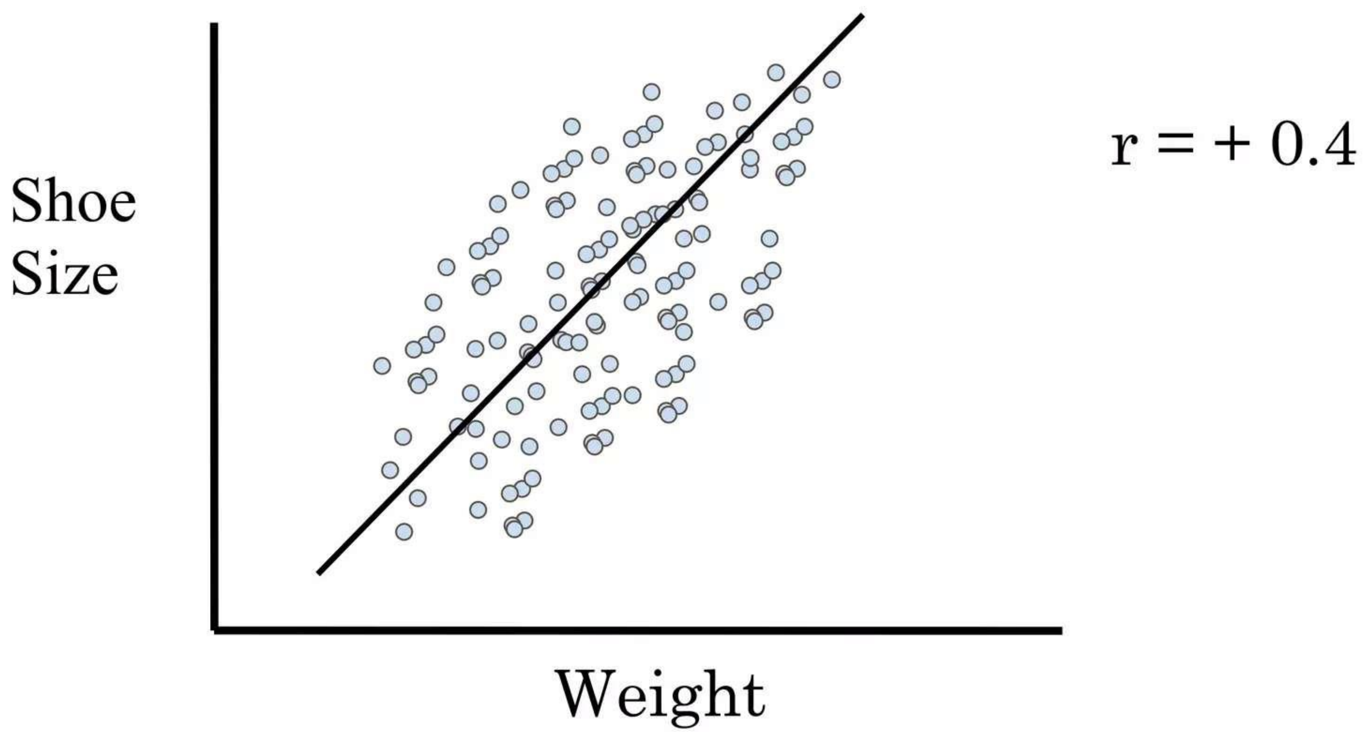
High Degree of positive correlation

- Positive relationship



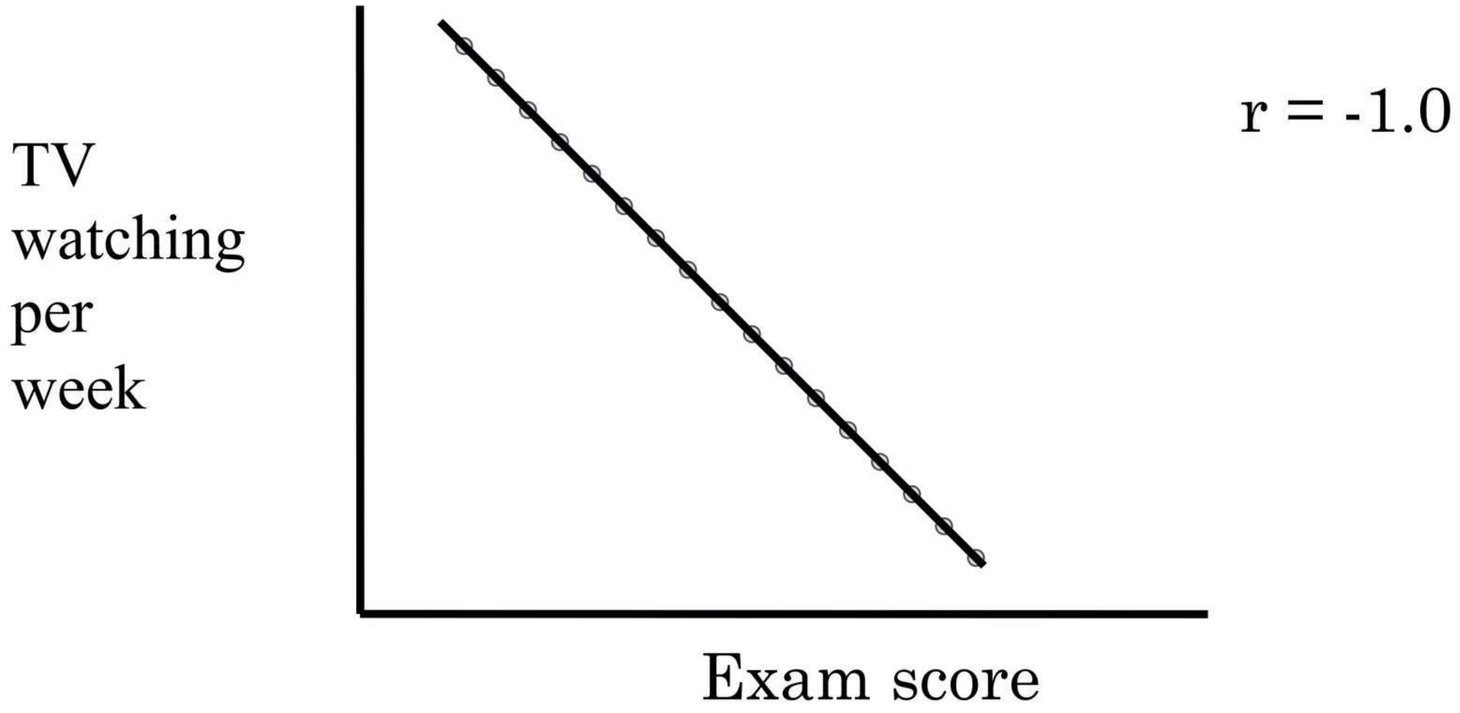
Degree of correlation

- **Moderate Positive Correlation**



Degree of correlation

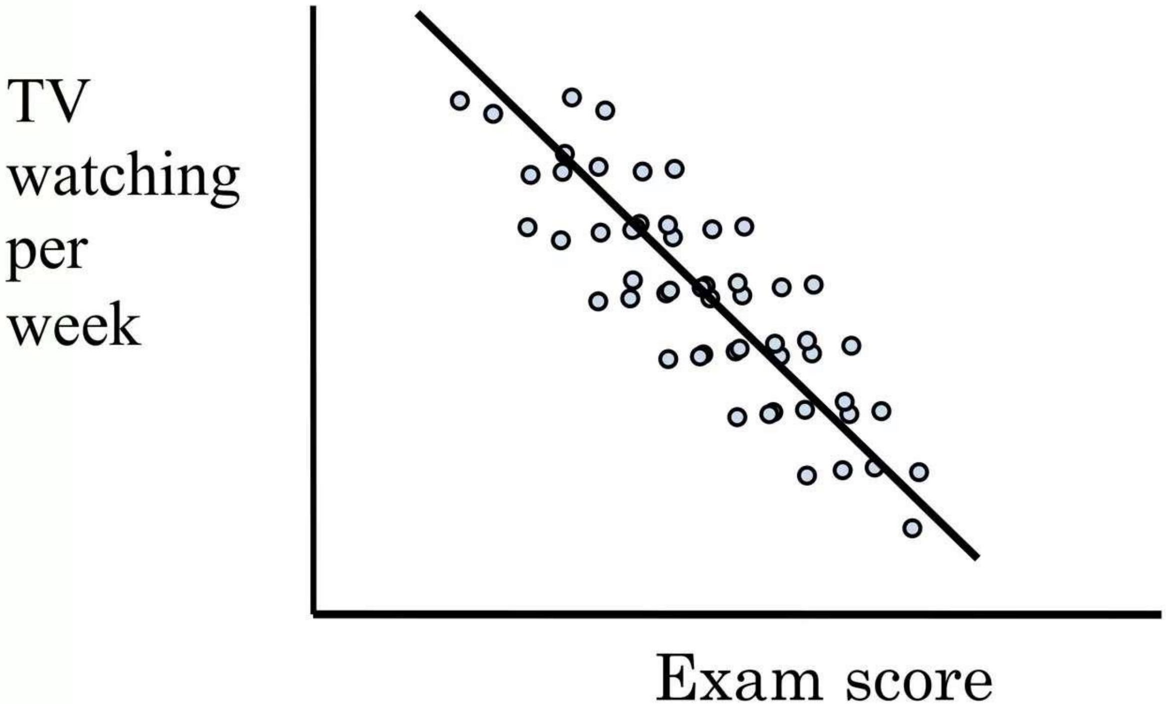
- **Perfect Negative Correlation**



Degree of correlation

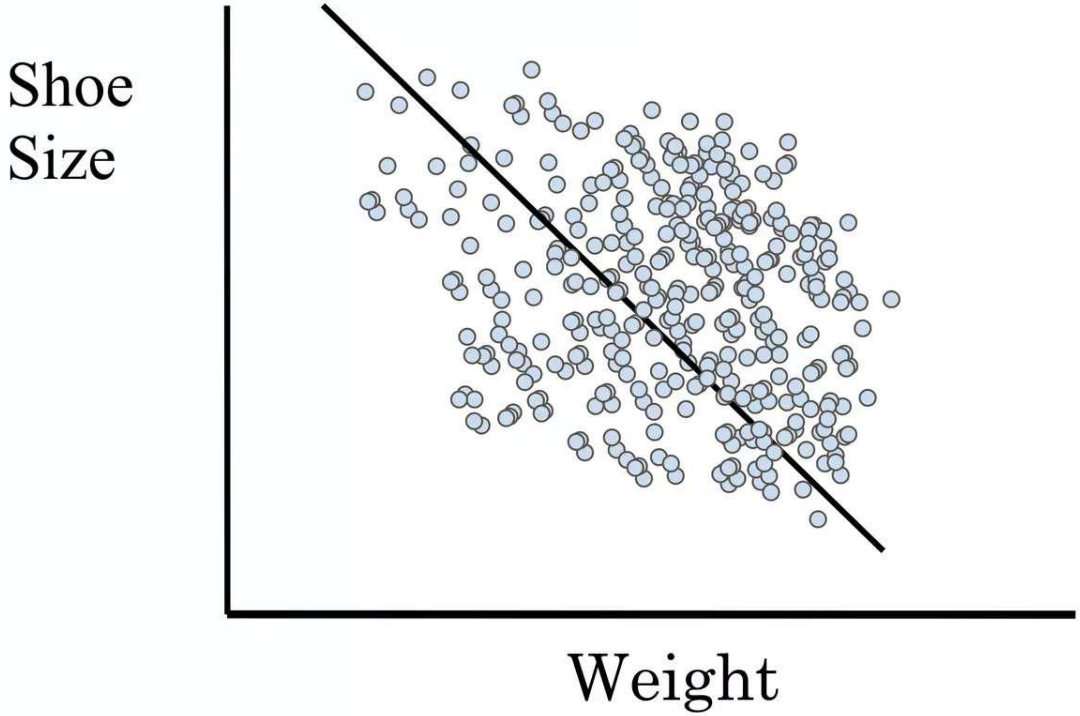
- **Moderate Negative Correlation**

r = -.80



Degree of correlation

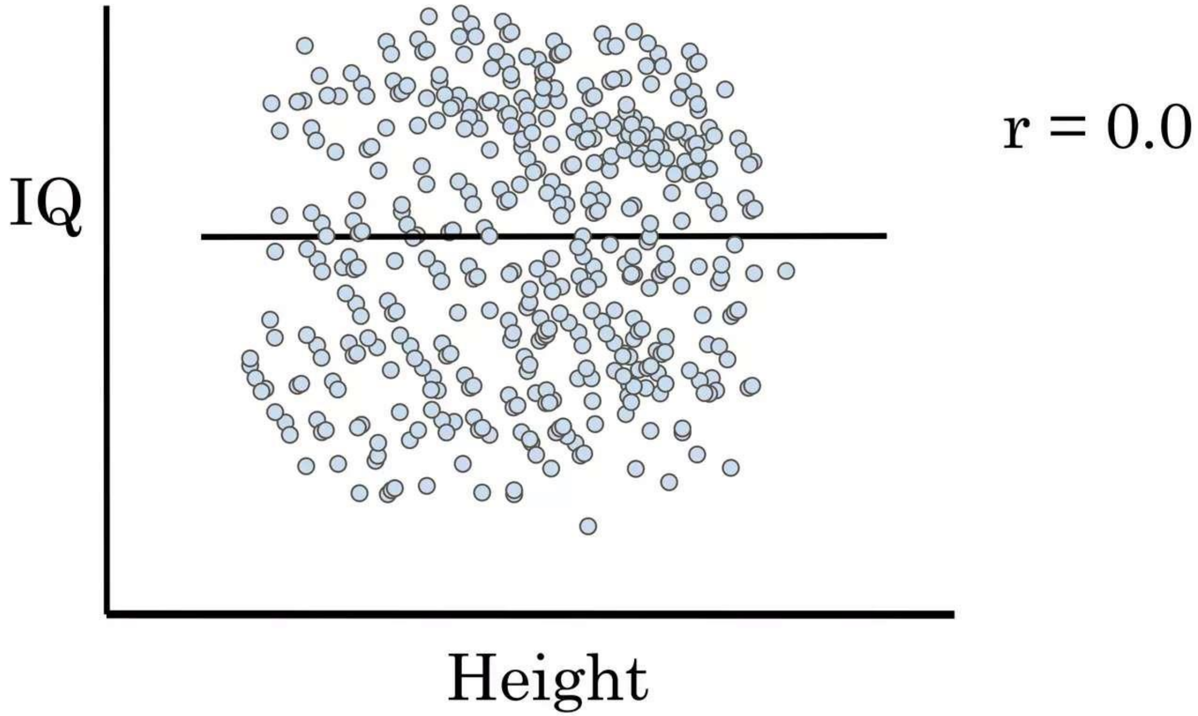
- **Weak negative Correlation**



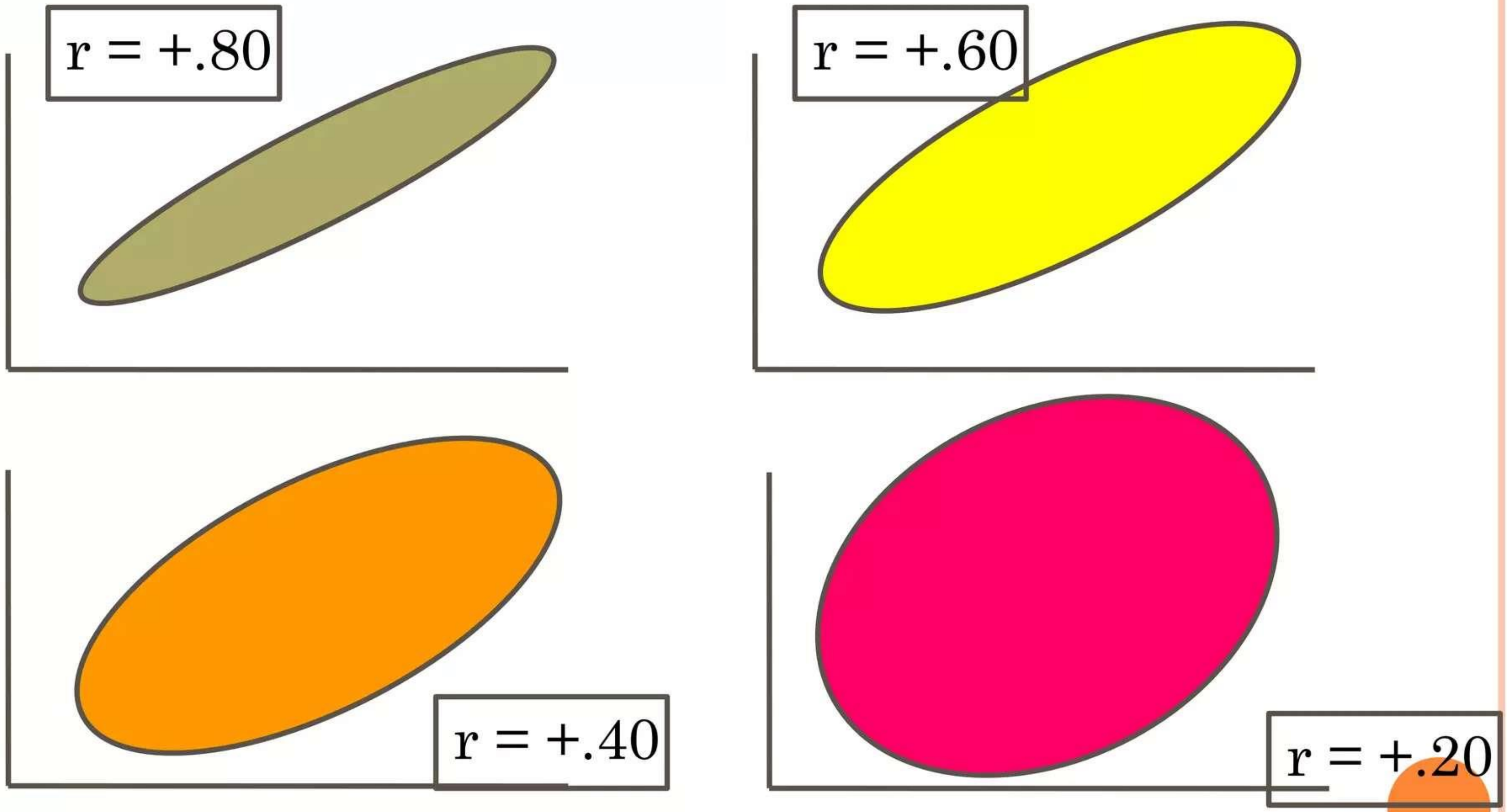
r = - 0.2

Degree of correlation

- **No Correlation (horizontal line)**



DEGREE OF CORRELATION (R)



Direction of the Relationship

- **Positive relationship** – Variables change in the same direction.
  - As X is increasing, Y is increasing
  - As X is decreasing, Y is decreasing
  - E.g., As height increases, so does weight.

Indicated by sign; (+) or (-).

- **Negative relationship** – Variables change in opposite directions.
  - As X is increasing, Y is decreasing
  - As X is decreasing, Y is increasing
  - E.g., As TV time increases, grades decrease

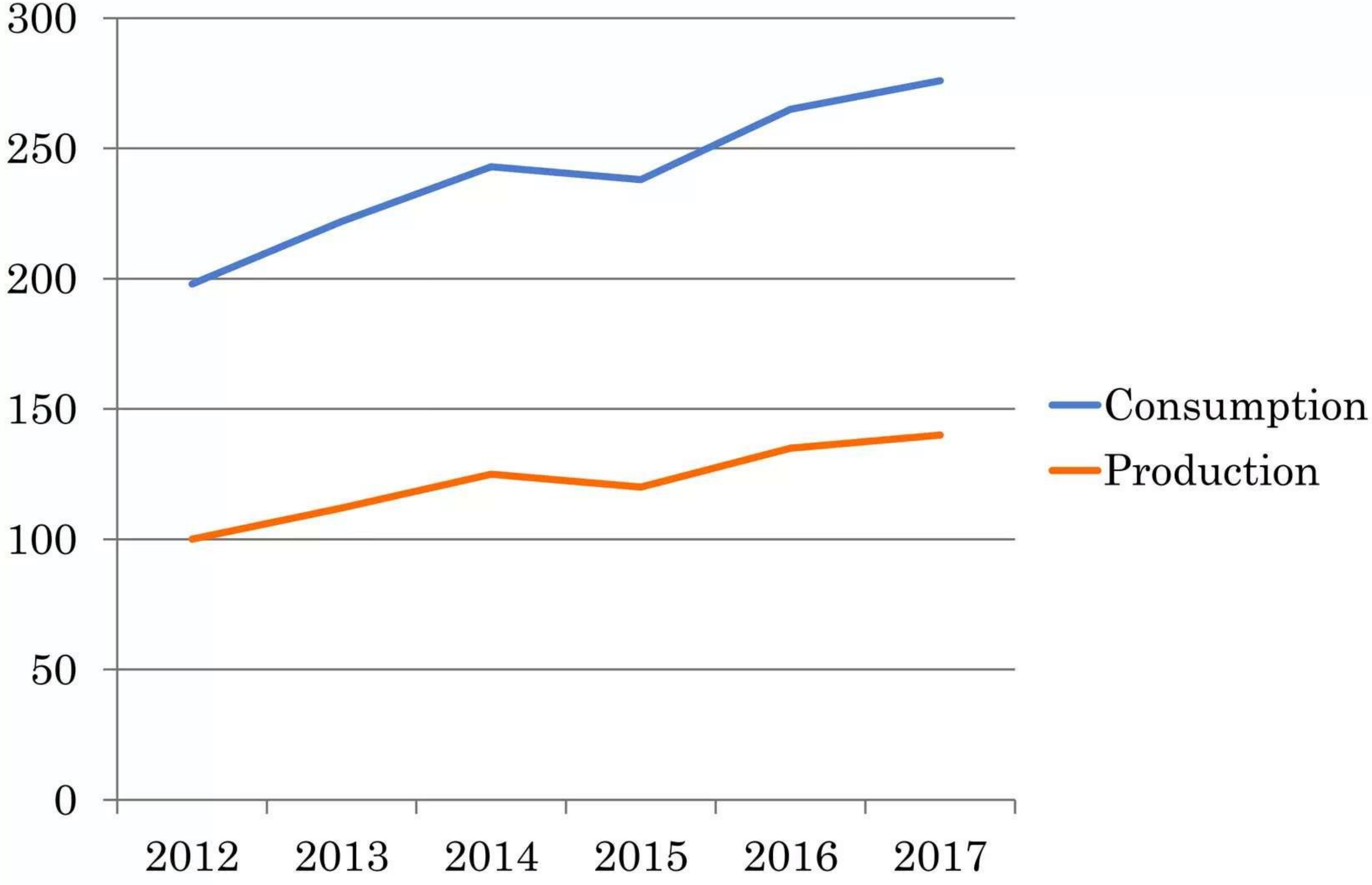
ADVANTAGES OF SCATTER DIAGRAM

- Simple & Non Mathematical method
- Not influenced by the size of extreme item
- First step in investing the relationship between two variables

DISADVANTAGE OF SCATTER DIAGRAM

Can not adopt the an exact degree of correlation

Correlation Graph



Birinder Singh, Assistant Professor, PCTE

Week 10

Karl Pearson's Coefficient of Correlation

- It is quantitative method of measuring correlation
- This method has been given by Karl Pearson
- It's the best method

Calculation of Coefficient of Correlation – Actual Mean Method

- Formula used is:
  - $r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}}$ where $x = X - \bar{X}$; $y = Y - \bar{Y}$

Q1: Find Karl Pearson's coefficient of correlation:

X	2	3	4	5	6	7	8
Y	4	7	8	9	10	14	18

Ans: 0.96

Q2: Find Karl Pearson's coefficient of correlation:

	X- Series	Y-series
No. of items	15	15
AM	25	18
Squares of deviations from mean	136	138

Summation of product of deviations of X & Y series from their respective arithmetic means = 122 Ans: 0.89

Practice Problems - Correlation

Q3: Find Karl Pearson's coefficient of correlation:

X	6	2	10	4	8
Y	9	11	?	8	7

Arithmetic Means of X & Y are 6 & 8 respectively.   Ans: – 0.92

Q4: Find the number of items as per the given data:

$r = 0.5, \Sigma xy = 120, \sigma_y = 8, \Sigma x^2 = 90$

where x & y are deviations from arithmetic means

Ans: 10

Q5: Find r:

$\Sigma X = 250, \Sigma Y = 300, \Sigma(X - 25)^2 = 480, \Sigma(Y - 30)^2 = 600$

$\Sigma(X - 25)(Y - 30) = 150$, $N = 10$ Ans: 0.28

Birinder Singh, Assistant Professor, PCTE

Calculation of Coefficient of Correlation – Assumed Mean Method

- Formula used is:
  - r = $\frac{N.\Sigma dxdy - \Sigma dx.\Sigma dy}{\sqrt{N.\Sigma dx^2 - (\Sigma dx)^2}\sqrt{N.\Sigma dy^2 - (\Sigma dy)^2}}$

Q6:Find r:

X	10	12	18	16	15	19	18	17
Y	30	35	45	44	42	48	47	46

Ans: 0.98

Q7: Find r, when deviations of two series from assumed mean are as follows: *Ans: 0.895*

Dx	+5	-4	-2	+20	-10	0	+3	0	-15	-5
Dy	+5	-12	-7	+25	-10	-3	0	+2	-9	-15

Birinder Singh, Assistant Professor, PCTE

Calculation of Coefficient of Correlation – Actual Data Method

- Formula used is:
  - $$r = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{\sqrt{N.\Sigma X^2 - (\Sigma X)^2}\sqrt{N.\Sigma Y^2 - (\Sigma Y)^2}}$$

Q8:Find r:

X	10	12	18	16	15	19	18	17
Y	30	35	45	44	42	48	47	46

Ans: 0.98

Q9: Calculate product moment correlation coefficient from the following data: *Ans: 0.996*

X	-5	-10	-15	-20	-25	-30
Y	50	40	30	20	10	5

Birinder Singh, Assistant Professor, PCTE

Important Typical Problems

Q10: Calculate the coefficient of correlation from the following data and interpret the result: *Ans: 0.76*

$N = 10, \ \Sigma XY = 8425, \ \bar{X} = 28.5, \ \bar{Y} = 28.0, \ \sigma x = 10.5, \ \sigma y = 5.6$

Q11: Following results were obtained from an analysis:

$N = 12, \ \Sigma XY = 334, \ \Sigma X = 30, \ \Sigma Y = 5, \ \Sigma X^2 = 670, \ \Sigma Y^2 = 285$

Later on it was discovered that one pair of values ($X = 11, Y = 4$) were wrongly copied. The correct value of the pair was ($X = 10, Y = 14$). Find the correct value of correlation coefficient. *Ans: 0.774*

Week 11

VARIANCE – COVARIANCE METHOD

- This method of determining correlation coefficient is based on covariance.
- $r = \frac{Cov\ (X,Y)}{\sqrt{Var(X)\ Var\ (Y)}} = \frac{Cov\ (X,Y)}{\sigma_x . \sigma_y}$

  where $\text{Cov (X, Y)} = \frac{\Sigma xy}{N} = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{N} = \frac{\Sigma XY}{N} - \bar{X}\bar{Y}$
- Another Way of calculating $r = \frac{\Sigma xy}{N.\ \sigma_x . \sigma_y .}$

Q12: For two series X & Y, Cov(X,Y) = 15, Var(X)=36, Var (Y)=25. Find r. Ans: 0.5

Q13: Find r when N = 30, $\bar{X} = 40$, $\bar{Y} = 50$, $\sigma_x = 6$, $\sigma_y = 7$, $\Sigma xy = 360$

Ans: 0.286

Q14: For two series X & Y, Cov(X,Y) = 25, Var(X)=36, r = 0.6. Find σ_y. Ans: 6.94

Birinder Singh, Assistant Professor, PCTE

Calculation of Correlation Coefficient – Grouped Data

- Formula used is:
  - $$r = \frac{N.\Sigma f dxdy - \Sigma f dx.\Sigma f dy}{\sqrt{N.\Sigma f dx^2 - (\Sigma f dx)^2}\sqrt{N.\Sigma f dy^2 - (\Sigma f dy)^2}}$$

Q15: Calculate Karl Pearson's coefficient of correlation:

X / Y	10-25	25-40	40-55
0-20	10	4	6
20-40	5	40	9
40-60	3	8	15

Ans: 0.33

Birinder Singh, Assistant Professor, PCTE

Properties of Coefficient of Correlation

- Karl Pearson's coefficient of correlation lies between -1 & 1, i.e. $-1 \leq r \leq +1$
- If the scale of a series is changed or the origin is shifted, there is no effect on the value of 'r'.
- 'r' is the geometric mean of the regression coefficients b_{yx} & b_{xy}, i.e. $r = \sqrt{b_{xy} . byx}$
- If X & Y are independent variables, then coefficient of correlation is zero but the converse is not necessarily true.
- 'r' is a pure number and is independent of the units of measurement.
- The coefficient of correlation between the two variables x & y is symmetric. i.e. $r_{yx} = r_{xy}$

Birinder Singh, Assistant Professor, PCTE

Week 12

Probable Error & Standard Error

- Probable Error is used to test the reliability of Karl Pearson's correlation coefficient.
- Probable Error (P.E.) = 0.6745 x $\frac{1-r^2}{\sqrt{N}}$
- Probable Error is used to interpret the value of the correlation coefficient as per the following:
  - If $|r|$ > 6 P.E., then 'r' is significant.
  - If $|r|$ < 6 P.E., then 'r' is insignificant. It means that there is no evidence of the existence of correlation in both the series.
- Probable Error also determines the upper & lower limits within which the correlation of randomly selected sample from the same universe will fall.
  - Upper Limit = r + P.E.
  - Lowe Limit = r – P.E.

Practice Problem – Probable Error

Q16: Find Karl Pearson's coefficient of correlation from the following data:

X	9	28	45	60	70	50
Y	100	60	50	40	33	57

Also calculate probable error and check whether it is significant or not. Ans: – 0.94, 0.032

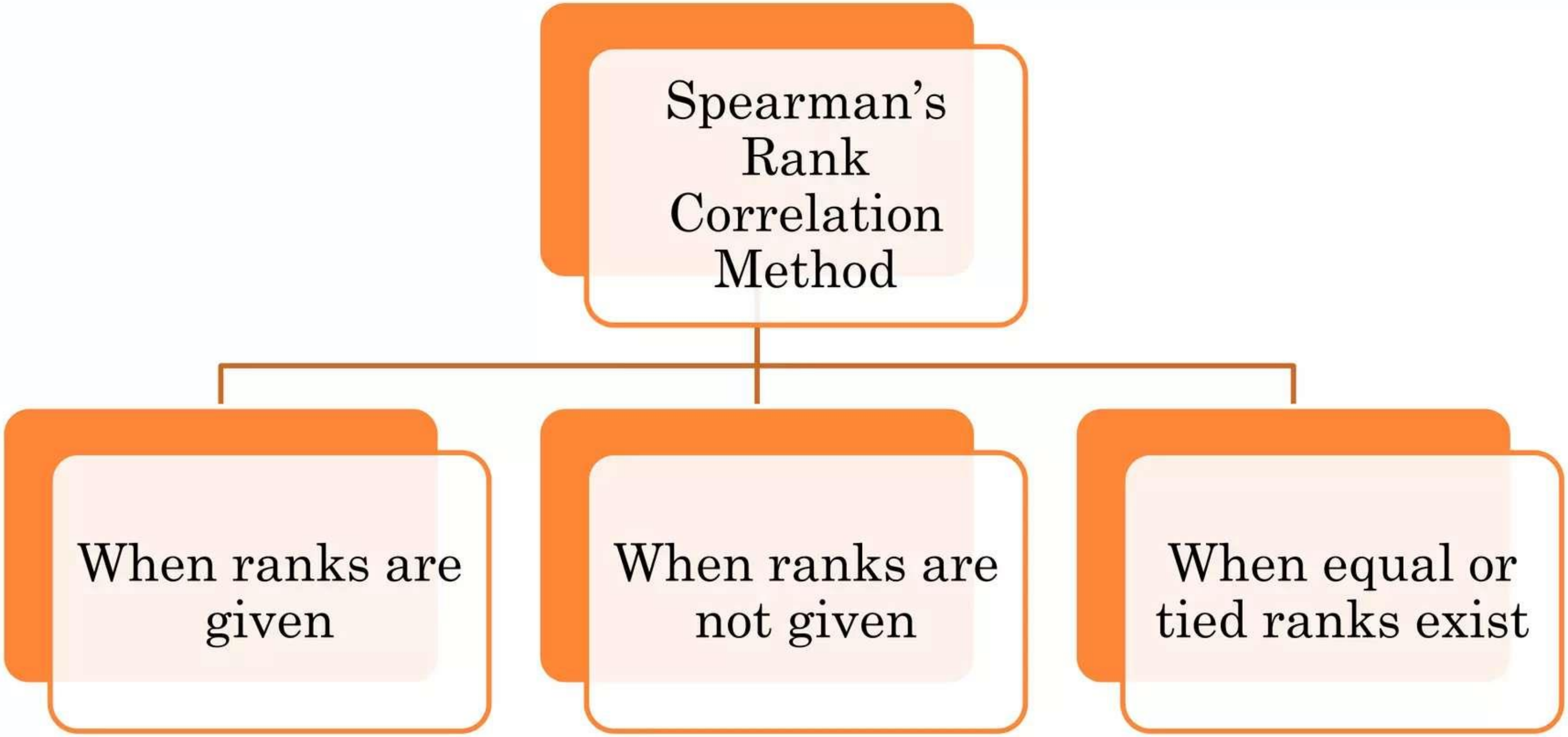
Q17: A student calculates the value of r as 0.7 when N = 5. He concludes that r is highly significant. Comment. Ans: Insignificant

SPEARMAN'S RANK CORRELATION METHOD

- Given by Prof. Spearman in 1904
- By this method, correlation between qualitative aspects like intelligence, honesty, beauty etc. can be calculated.
- These variables can be assigned ranks but their quantitative measurement is not possible.
- It is denoted by $\mathbf{R = 1 - \frac{6\,\Sigma D^2}{N\,(N^2 - 1)}}$
  - R = Rank correlation coefficient
  - D = Difference between two ranks ($R_1 - R_2$)
  - N = Number of pair of observations
- As in case of r, $-1 \leq R \leq 1$
- *The sum total of Rank Difference is always equal to zero. i.e. $\Sigma D = 0$.*

Birinder Singh, Assistant Professor, PCTE

Three Cases


Spearman's Rank Correlation Method
When ranks are given
When ranks are not given
When equal or tied ranks exist

PRACTICE PROBLEMS – RANK CORRELATION (WHEN RANKS ARE GIVEN)

Q18: In a fancy dress competition, two judges accorded the following ranks to eight participants:

Judge X	8	7	6	3	2	1	5	4
Judge Y	7	5	4	1	3	2	6	8

Calculate the coefficient of rank correlation. Ans: .62

Q19: Ten competitors in a beauty contest are ranked by three judges X, Y, Z:

X	1	6	5	10	3	2	4	9	7	8
Y	3	5	8	4	7	10	2	1	6	9
Z	6	4	9	8	1	2	3	10	5	7

Use the rank correlation coefficient to determine which pair of judges has the nearest approach to common tastes in beauty.

Ans: X & Z

Birinder Singh, Assistant Professor, PCTE

Practice Problems – Rank Correlation (When Ranks are Not Given)

Q20: Find out the coefficient of Rank Correlation between X & Y:

X	15	17	14	13	11	12	16	18	10	9
Y	18	12	4	6	7	9	3	10	2	5

Ans: 0.48

Practice Problems – Rank Correlation (When Ranks are Equal or Tied)

- When two or more items have equal values in a series, so common ranks i.e. average of the ranks are assigned to equal values.
- Here $\mathbf{R = 1} - \dfrac{6\left[\Sigma D^2 + \frac{m^3 - m}{12} + \frac{m^3 - m}{12} + \ldots\ldots\ldots\ldots\right]}{N\,(N^2 - 1)}$
  - m = No. of items of equal ranks
  - The correction factor of $\frac{m^3 - m}{12}$ is added to ΣD^2 for such number of times as the cases of equal ranks in the question

Practice Problems – Rank Correlation (When Ranks are Equal or Tied)

Q21: Calculate R:

X	15	10	20	28	12	10	16	18
Y	16	14	10	12	11	15	18	10

Ans: – 0.37

Q22: Calculate Rank Correlation:

X	40	50	60	60	80	50	70	60
Y	80	120	160	170	130	200	210	130

Ans: 0.43

Birinder Singh, Assistant Professor, PCTE

Important Typical Problems – Rank Correlation

Q23: Calculate Rank Correlation from the following data:

Ans: 0.64

Serial No.	1	2	3	4	5	6	7	8	9	10
Rank Difference	-2	?	-1	+3	+2	0	-4	+3	+3	-2

Q24: The coefficient of rank correlation of marks obtained by 10 students in English & Math was found to be 0.5. It was later discovered that the difference in the ranks in two subjects was wrongly taken as 3 instead of 7. Find the correct rank correlation.

Ans: 0.26

Q25: The rank correlation coefficient between marks obtained by some students in English & Math is found to be 0.8. If the total of squares of rank differences is 33, find the number of students.

Ans: 10

Week 13

Concurrent Deviation Method

- Correlation is determined on the basis of direction of the deviations.
- **Under this method,** the direction of deviations are assigned (+) or (-) or (0) signs.
- If the value is more than its preceding value, then its deviation is assigned (+) sign.
- If the value is less than its preceding value, then its deviation is assigned (-) sign.
- If the value is equal to its preceding value, then its deviation is assigned (0) sign.
- The deviations dx & dy are multiplied to get dxdy. Product of similar signs will be (+) and for opposite signs will be (-).
- Summing the positive dxdy signs, their number is counted. It is called *CONCURRENT DEVIATIONS*. It is denoted by *C*.
- ***Formula used:*** $r_c = \pm\sqrt{\pm\left[\frac{2C-n}{n}\right]}$ ***where*** r_c ***= Correlation of CD, C = No. of Concurrent Deviations, n = N – 1.***

Practice Problems – Coefficient of Concurrent Deviations

Q26: Find the Coefficient of Concurrent Deviation from the following data:

Year	2001	2002	2003	2004	2005	2006	2007
Demand	150	154	160	172	160	165	180
Price	200	180	170	160	190	180	172

Ans: – 1

Q27: Find the Coefficient of Concurrent Deviation from the following data:

X	112	125	126	118	118	121	125	125	131	135
Y	106	102	102	104	98	96	97	97	95	90

Ans: – 0.75

Birinder Singh, Assistant Professor, PCTE

Coefficient of Determination (CoD)

- CoD is used for the interpretation of coefficient of correlation and comparing the two or more correlation coefficients.
- ***It is the square of the coefficient of correlation i.e. r^2.***
- It explains the percentage variation in the dependent variable Y that can be explained in terms of the independent variable X.
- If $r = 0.8$, $r^2 = 0.64$, it implies that 64% of the total variations in Y occurs due to X. The remaining 34% variation occurs due to external factors.
- So, CoD $= r^2 = \frac{Explained\ Variance}{Total\ Variance}$
- Coefficient of Non Determination$= K^2 = 1 - r^2 = \frac{Unexplained\ Variance}{Total\ Variance}$
- Coefficient of Alienation $= \sqrt{1 - r^2}$

Practice Problems – CoD

Q28: The coefficient of correlation between consumption expenditure (C) and disposable income (Y) in a study was found to be +0.8. What percentage of variation in C are explained by variation in Y? Ans: 64%

Birinder Singh, Assistant Professor, PCTE

CLASS TEST

Q1: In a fancy dress competition, two judges accorded the following ranks to eight participants:

Judge X	8	7	6	3	2	1	5	4
Judge Y	7	5	4	1	3	2	6	8

Calculate the coefficient of rank correlation.

Q2: Following results were obtained from an analysis:

$N = 12, \quad \Sigma XY = 334, \quad \Sigma X = 30, \quad \Sigma Y = 5, \quad \Sigma X^2 = 670, \quad \Sigma Y^2 = 285$

Later on it was discovered that one pair of values ($X = 11, Y = 4$) were wrongly copied. The correct value of the pair was ($X = 10, Y = 14$). Find the correct value of correlation coefficient.

Birinder Singh, Assistant Professor, PCTE

- Median is the number present in the middle when the numbers in a set of data are arranged in ascending or descending order. If the number of numbers in a data set is even, then the median is the mean of the two middle numbers.
- Mode is the value that occurs most frequently in a set of data.

Week 14

Statistics for Business and Economics

7th Edition

Chapter 9

Hypothesis Testing: Single Population

Chapter Goals

After completing this chapter, you should be able to:

- Formulate null and alternative hypotheses for applications involving
  - a single population mean from a normal distribution
  - a single population proportion (large samples)
  - the variance of a normal distribution
- Formulate a decision rule for testing a hypothesis
- Know how to use the critical value and p-value approaches to test the null hypothesis (for both mean and proportion problems)
- Know what Type I and Type II errors are
- Assess the power of a test

What is a Hypothesis?

- A hypothesis is a claim (assumption) about a population parameter:

  - population mean

Example: The mean monthly cell phone bill of this city is μ = $42

  - population proportion

Example: The proportion of adults in this city with cell phones is $p = .68$

The Null Hypothesis, H_0

- States the assumption (numerical) to be tested

> Example: The average number of TV sets in U.S. Homes is equal to three ($H_0 : \mu = 3$)

- Is always about a population parameter, not about a sample statistic

$$H_0 : \mu = 3$$

$$H_0 : \bar{X} = 3$$

The Null Hypothesis, H_0

(continued)

- Begin with the assumption that the null hypothesis is true
  - Similar to the notion of innocent until proven guilty

- Refers to the status quo
- Always contains "=" , "≤" or "≥" sign
- May or may not be rejected

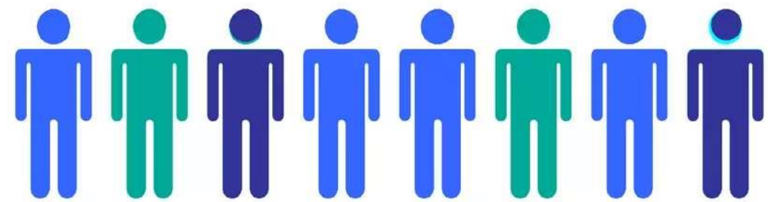
The Alternative Hypothesis, H_1

- Is the opposite of the null hypothesis
  - e.g., The average number of TV sets in U.S. homes is not equal to 3 ($H_1: \mu \neq 3$)
- Challenges the status quo
- Never contains the "=" , "≤" or "≥" sign
- May or may not be supported
- Is generally the hypothesis that the researcher is trying to support

Week 15

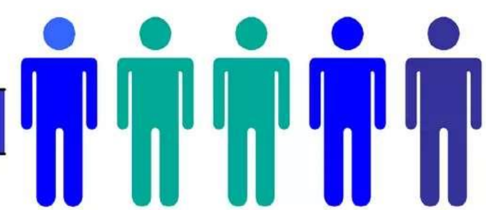
Hypothesis Testing Process

Claim: **the population mean age is 50.** **(Null Hypothesis: H_0: μ = 50)**



Population

Now select a random sample



Sample

Suppose the sample mean age is 20: $\bar{X} = 20$

Is $\bar{X} = 20$ likely if $\mu = 50$?

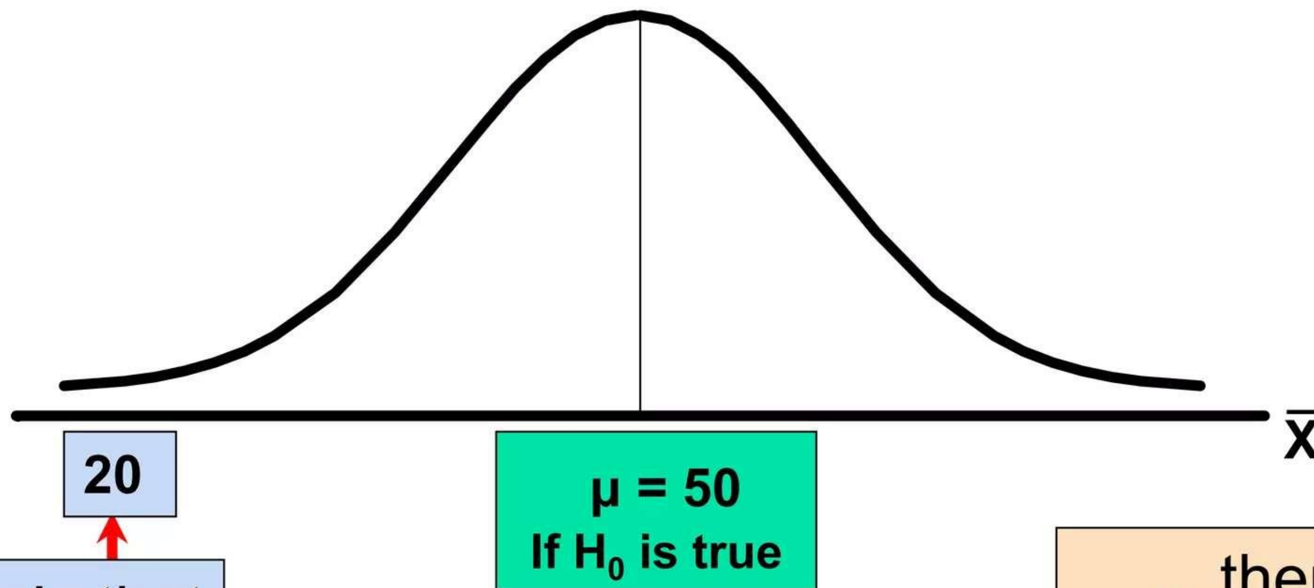
If not likely,

REJECT

Null Hypothesis

Reason for Rejecting H_0

Sampling Distribution of $\bar{X}$



If it is unlikely that we would get a sample mean of this value ...

... if in fact this were the population mean…

... then we reject the null hypothesis that μ = 50.

Level of Significance, α

- **Defines the unlikely values of the sample statistic if the null hypothesis is true**
  - Defines rejection region of the sampling distribution
- Is designated by α , (level of significance)
  - Typical values are .01, .05, or .10
- Is selected by the researcher at the beginning
- Provides the critical value(s) of the test

Level of Significance and the Rejection Region

Level of significance = α

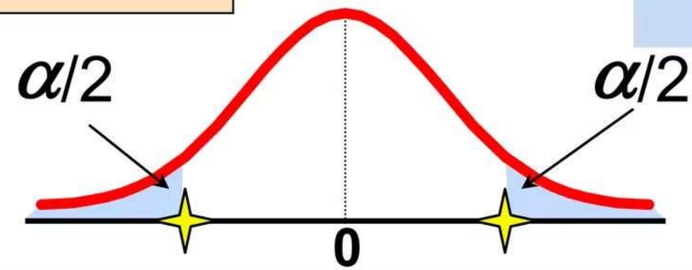
Represents critical value

Rejection region is shaded

$H_0: \mu = 3$

$H_1: \mu \neq 3$

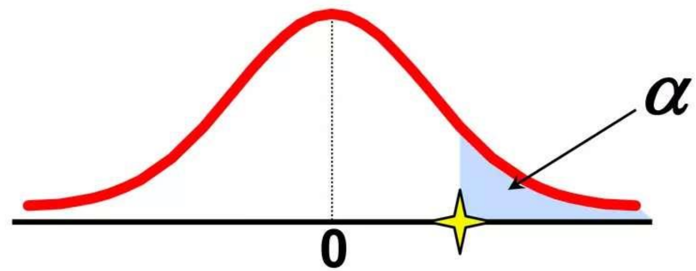
Two-tail test



$H_0: \mu \leq 3$

$H_1: \mu > 3$

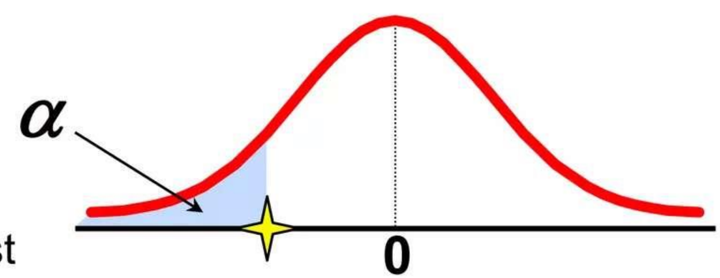
Upper-tail test



$H_0: \mu \geq 3$

$H_1: \mu < 3$

Lower-tail test



Errors in Making Decisions

- **Type I Error**
  - Reject a true null hypothesis
  - Considered a serious type of error

The probability of Type I Error is α

- Called level of significance of the test
- Set by researcher in advance

Errors in Making Decisions

(continued)

- **Type II Error**
  - Fail to reject a false null hypothesis

The probability of Type II Error is β

Outcomes and Probabilities

Possible Hypothesis Test Outcomes

	Actual Situation	
Decision	H_0 True	H_0 False
Do Not Reject H_0	**No Error** **(1 - α)**	**Type II Error** **(β)**
Reject H_0	**Type I Error** **(α)**	**No Error** **(1 - β)**

Key:
Outcome
(Probability)

Type I & II Error Relationship

- Type I and Type II errors can not happen at the same time
  - Type I error can only occur if H_0 is true
  - Type II error can only occur if H_0 is false

If Type I error probability (α) ↑, then

Type II error probability (β) ↓

Factors Affecting Type II Error

- All else equal,
  - β ↑ when the difference between hypothesized parameter and its true value ↓
  - β ↑ when α ↓
  - β ↑ when σ ↑
  - β ↑ when n ↓

Power of the Test

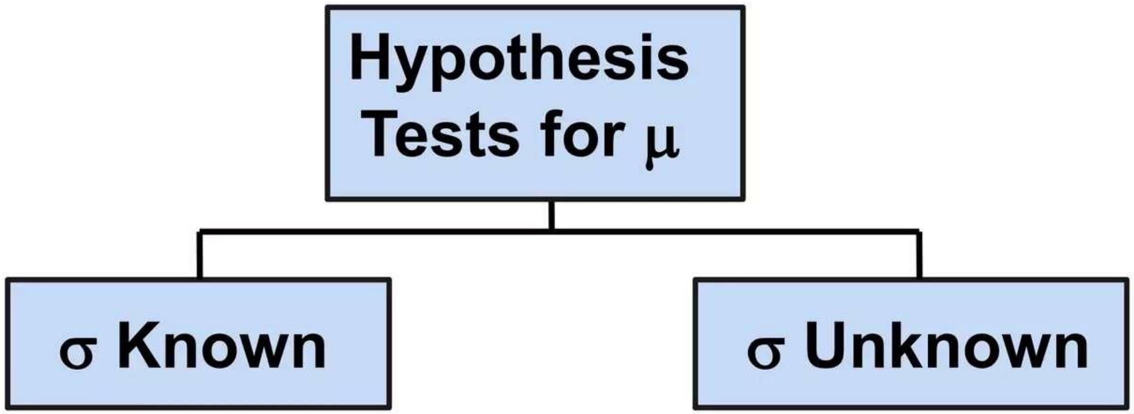
- The power of a test is the probability of rejecting a null hypothesis that is false

- i.e., Power = P(Reject H_0 | H_1 is true)

  - Power of the test increases as the sample size increases

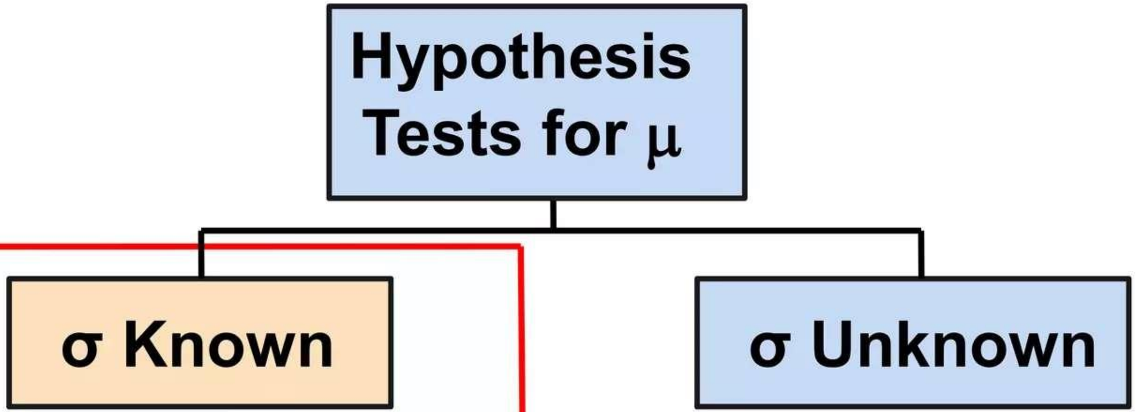
Week 16

Hypothesis Tests for the Mean



Test of Hypothesis for the Mean (σ Known)

- Convert sample result ($\bar{x}$) to a z value



Consider the test

$$H_0: \mu = \mu_0$$
$$H_1: \mu > \mu_0$$

(Assume the population is normal)

The decision rule is:

$$\text{Reject } H_0 \text{ if } \quad z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > z_\alpha$$

Decision Rule

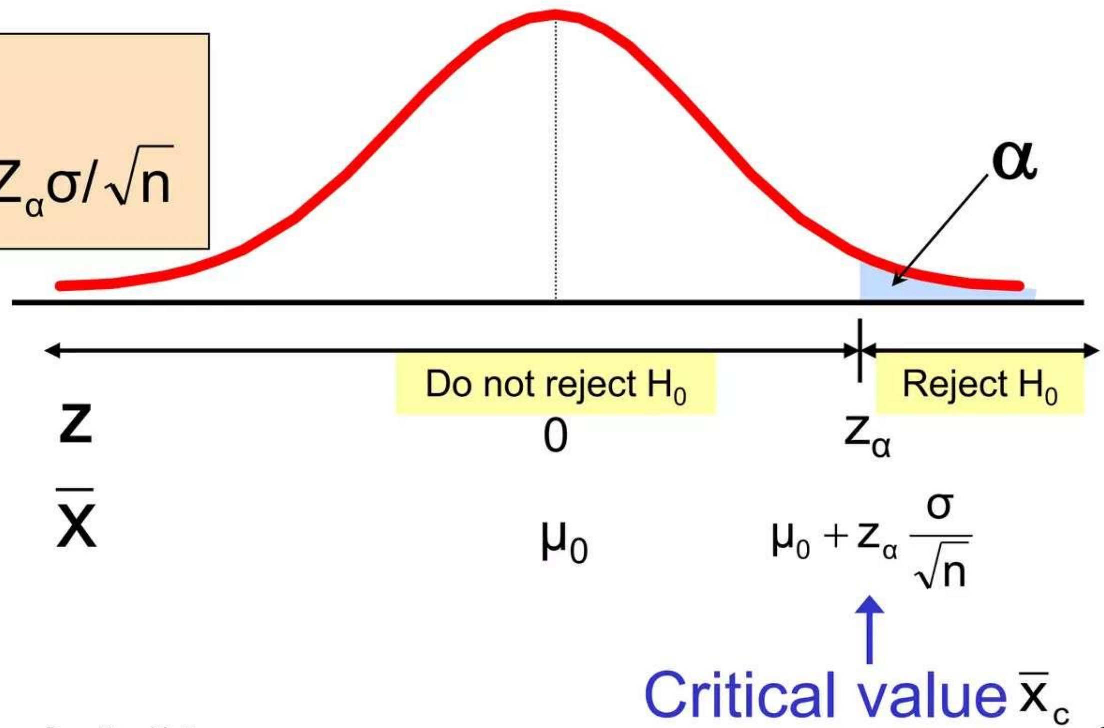
$H_0: \mu = \mu_0$

$H_1: \mu > \mu_0$

Reject H_0 if $z = \dfrac{\bar{x} - \mu_0}{\dfrac{\sigma}{\sqrt{n}}} > z_\alpha$

Alternate rule:

Reject H_0 if $\bar{x} > \mu_0 + Z_\alpha \sigma / \sqrt{n}$



p-Value Approach to Testing

- p-value: Probability of obtaining a test statistic more extreme (≤ or ≥) than the observed sample value given H_0 is true
  - Also called observed level of significance
  - Smallest value of α for which H_0 can be rejected

p-Value Approach to Testing

(continued)

- Convert sample result (e.g., $\bar{x}$) to test statistic (e.g., z statistic)
- Obtain the p-value
  - For an upper tail test:

$$\text{p-value} = P\left(z > \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}, \text{ given that } H_0 \text{ is true}\right)$$

$$= P\left(z > \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \mid \mu = \mu_0\right)$$

- Decision rule: compare the p-value to α
  - If p-value $< \alpha$, reject H_0
  - If p-value $\geq \alpha$, do not reject H_0

Example: Upper-Tail Z Test for Mean (σ Known)

A phone industry manager thinks that customer monthly cell phone bill have increased, and now average over \$52 per month. The company wishes to test this claim. (Assume $\sigma = 10$ is known)

Form hypothesis test:

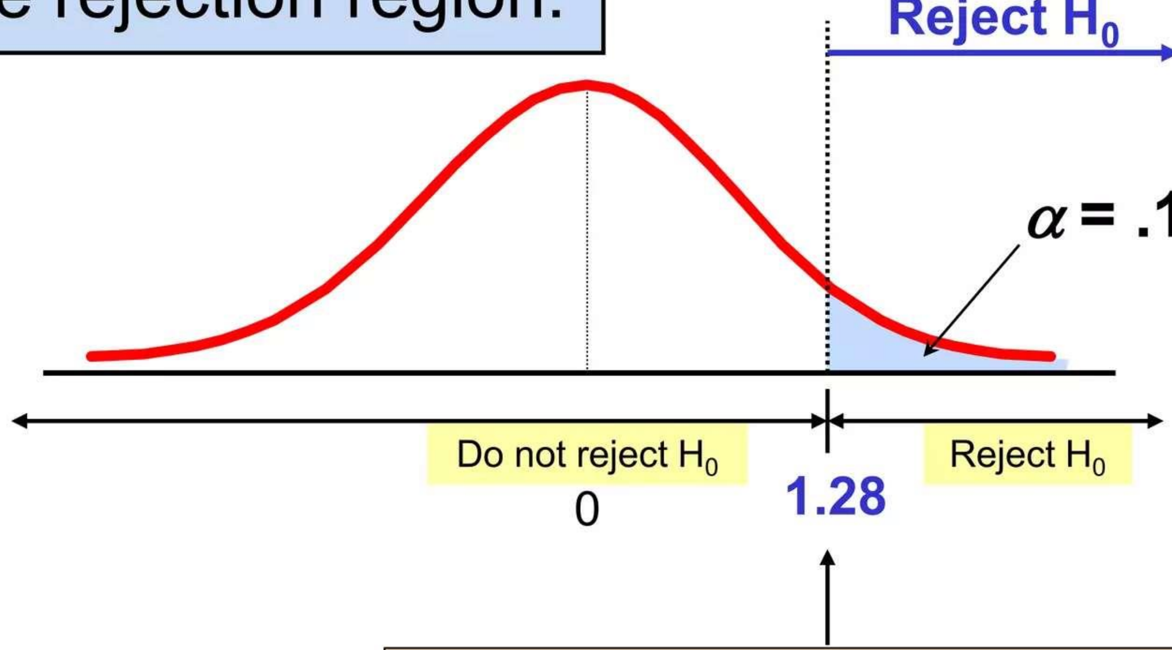
H_0: $\mu \le 52$	the average is not over \$52 per month
H_1: $\mu > 52$	the average is greater than \$52 per month (i.e., sufficient evidence exists to support the manager's claim)

Example: Find Rejection Region

(continued)

- Suppose that α = .10 is chosen for this test

Find the rejection region:



Reject H_0 if $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > 1.28$

Example: Sample Results

(continued)

Obtain sample and compute the test statistic

Suppose a sample is taken with the following results: $n = 64$, $\bar{x} = 53.1$ ($\sigma = 10$ was assumed known)

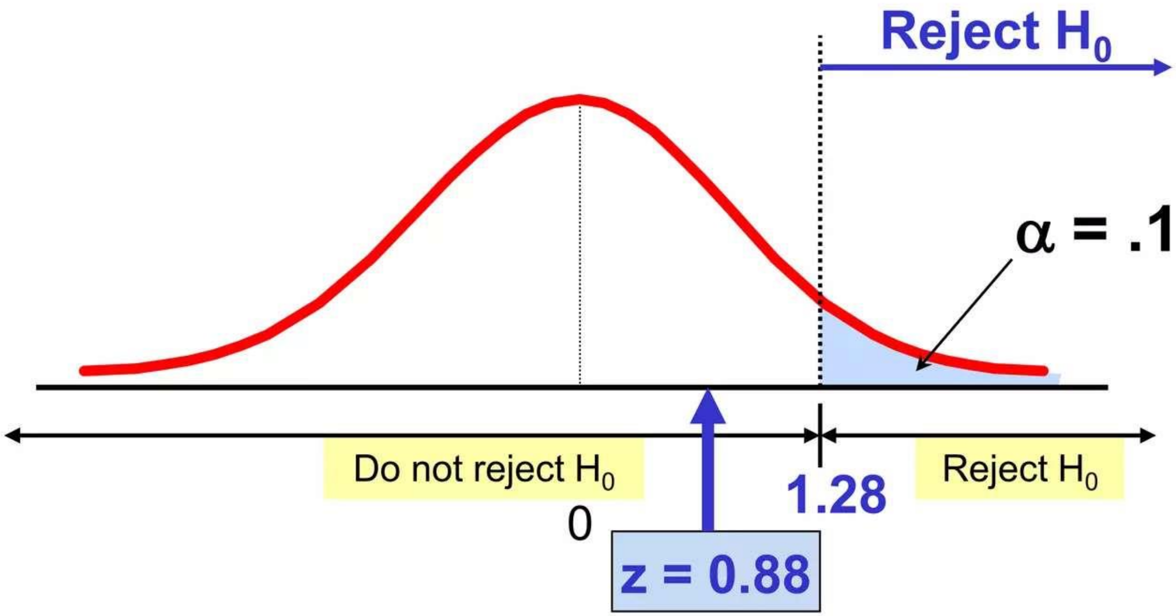
- Using the sample results,

$$z = \frac{\bar{x} - \mu_0}{\dfrac{\sigma}{\sqrt{n}}} = \frac{53.1 - 52}{\dfrac{10}{\sqrt{64}}} = 0.88$$

Example: Decision

(continued)

Reach a decision and interpret the result:



Do not reject H_0 since z = 0.88 < 1.28

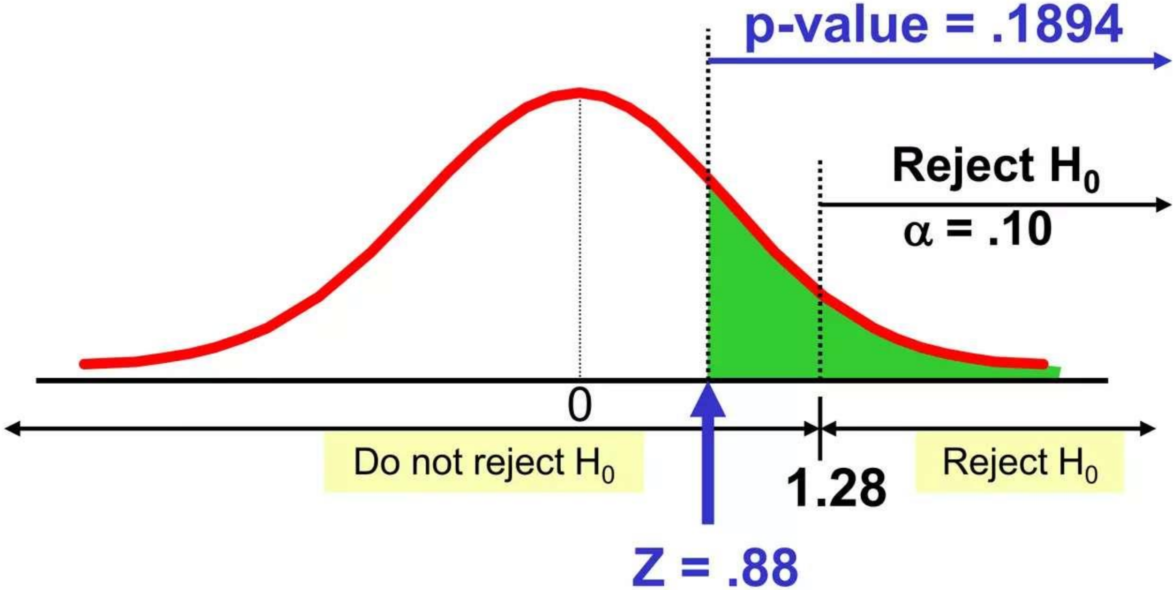
i.e.: there is not sufficient evidence that the mean bill is over $52

Example: p-Value Solution

(continued)

Calculate the p-value and compare to α

(assuming that μ = 52.0)



$$P(\bar{x} \geq 53.1 | \mu = 52.0)$$

$$= P\left(z \geq \frac{53.1 - 52.0}{10/\sqrt{64}}\right)$$

$$= P(z \geq 0.88) = 1 - .8106$$

$$= .1894$$

Do not reject H_0 since p-value = .1894 > α = .10

One-Tail Tests

- In many cases, the alternative hypothesis focuses on one particular direction

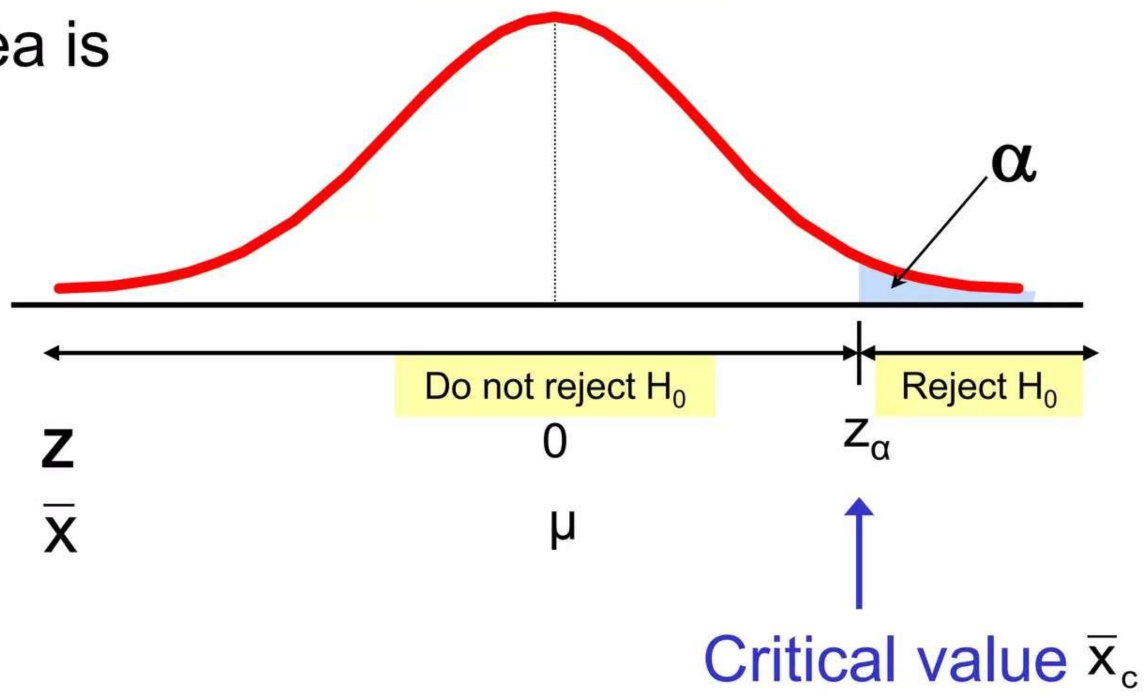
$H_0: \mu \leq 3$ $H_1: \mu > 3$	→ This is an upper-tail test since the alternative hypothesis is focused on the upper tail above the mean of 3
$H_0: \mu \geq 3$ $H_1: \mu < 3$	→ This is a lower-tail test since the alternative hypothesis is focused on the lower tail below the mean of 3

Upper-Tail Tests

- There is only one critical value, since the rejection area is in only one tail

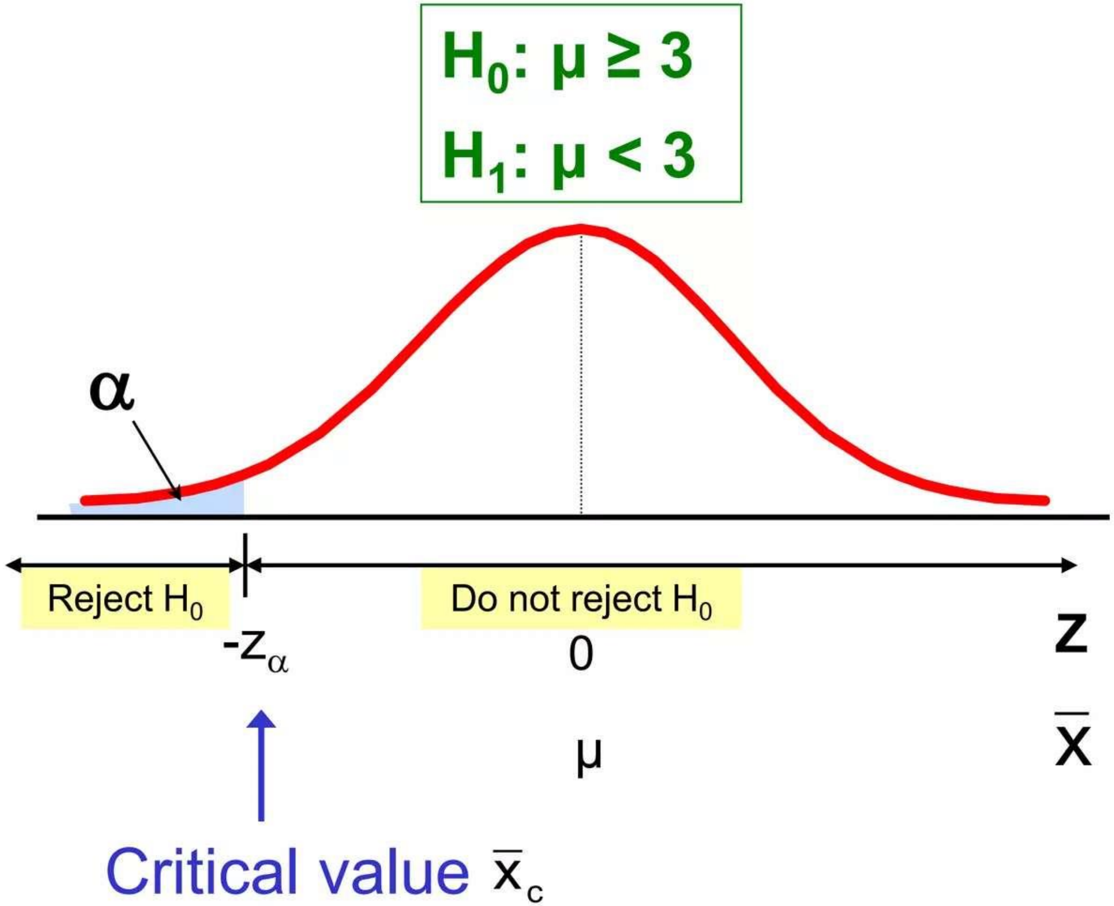
H_0: μ ≤ 3

H_1: μ > 3



Lower-Tail Tests

- There is only one critical value, since the rejection area is in only one tail

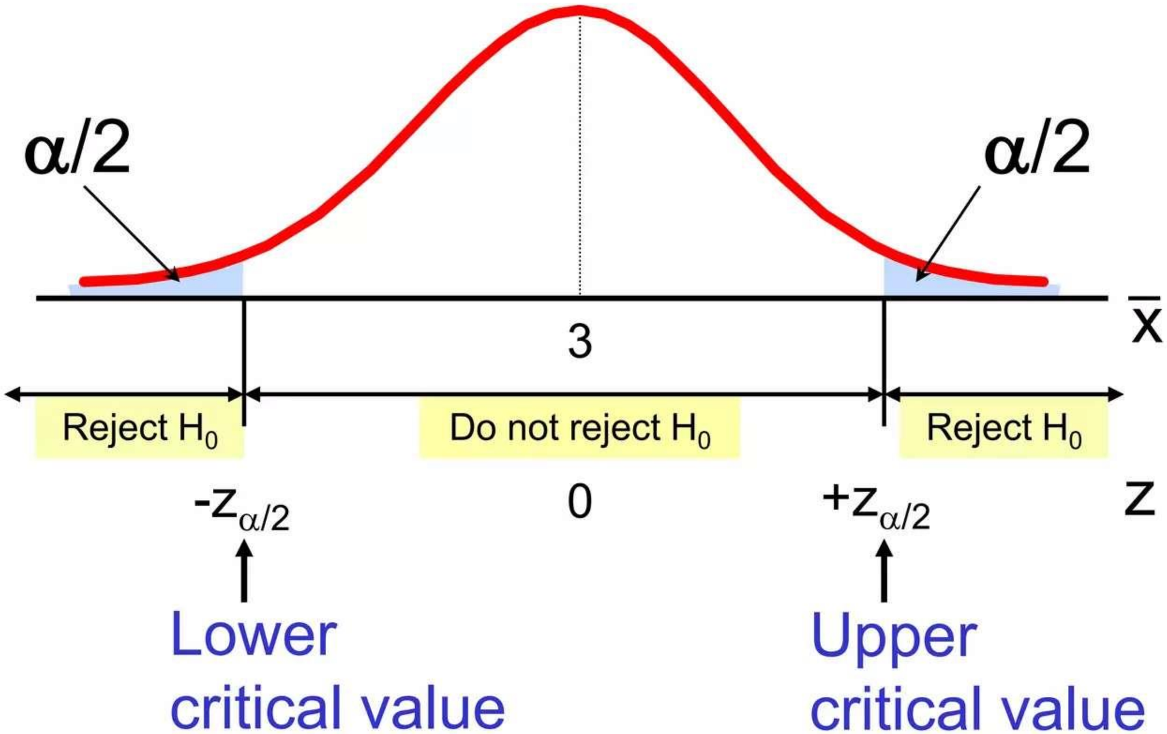


Two-Tail Tests

- In some settings, the alternative hypothesis does not specify a unique direction

H_0: $\mu = 3$
H_1: $\mu \neq 3$

- There are two critical values, defining the two regions of rejection



Hypothesis Testing Example

Test the claim that the true mean # of TV sets in US homes is equal to 3.
(Assume $\sigma = 0.8$)

- State the appropriate null and alternative hypotheses
  - $H_0: \mu = 3$, $H_1: \mu \neq 3$ (This is a two tailed test)
- Specify the desired level of significance
  - Suppose that $\alpha = .05$ is chosen for this test
- Choose a sample size
  - Suppose a sample of size n = 100 is selecte[d]

Hypothesis Testing Example

(continued)

- Determine the appropriate technique
  - σ is known so this is a z test
- Set up the critical values
  - For $\alpha = .05$ the critical z values are ±1.96
- Collect the data and compute the test statistic
  - Suppose the sample results are

  $n = 100$, $\bar{x} = 2.84$ ($\sigma = 0.8$ is assumed known)

So the test statistic is:

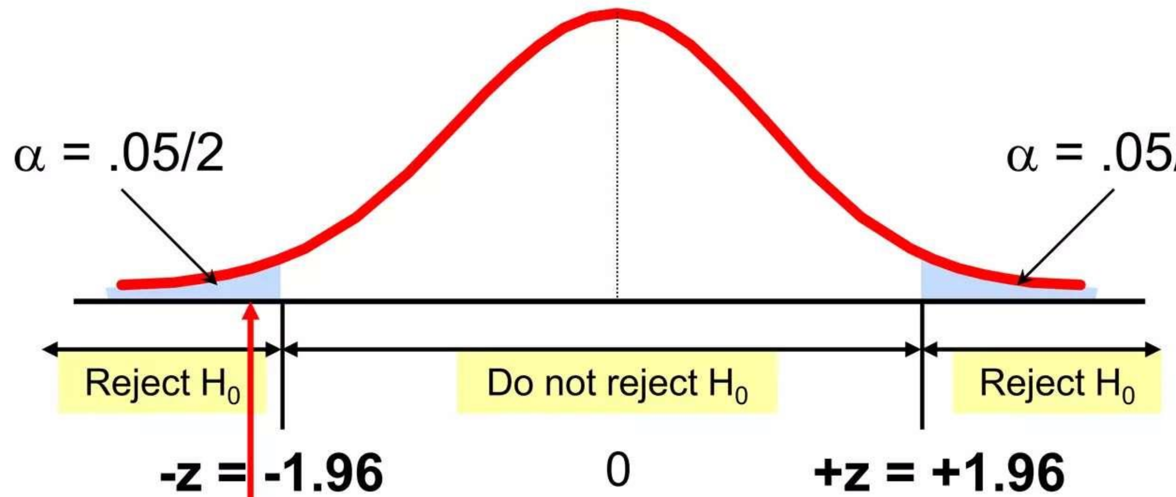
$$z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{2.84 - 3}{\frac{0.8}{\sqrt{100}}} = \frac{-.16}{.08} = -2.0$$

Hypothesis Testing Example

(continued)

- Is the test statistic in the rejection region?

Reject H_0 if $z < -1.96$ or $z > 1.96$; otherwise do not reject H_0

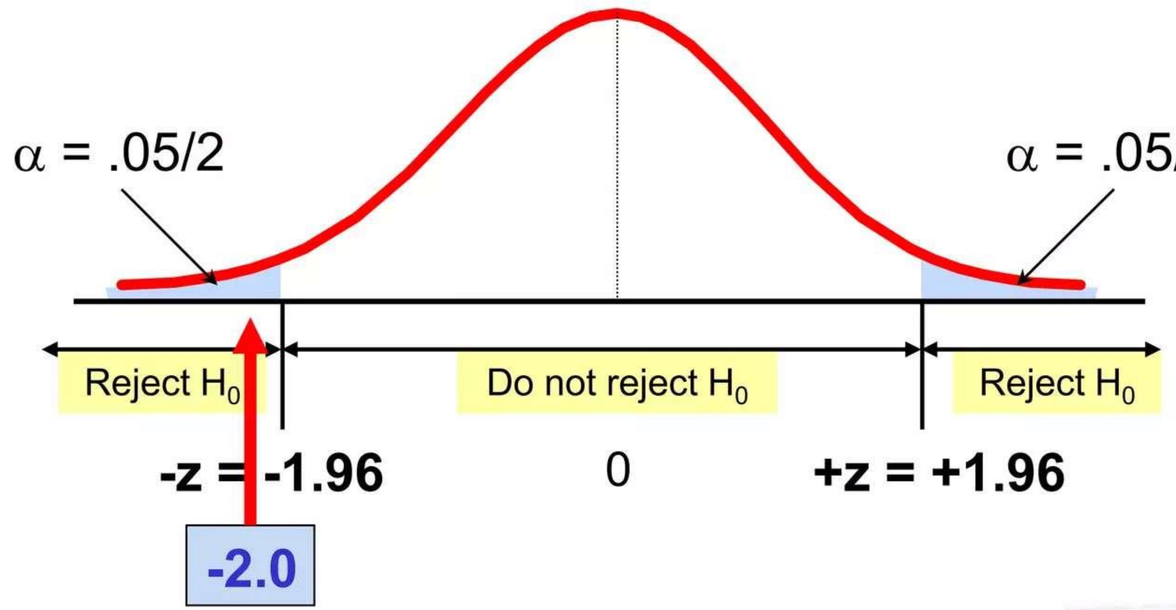


Here, $z = -2.0 < -1.96$, so the test statistic is in the rejection region

Hypothesis Testing Example

(continued)

- Reach a decision and interpret the result



Since z = -2.0 < -1.96, we <u>reject the null hypothesis</u> and conclude that there is sufficient evidence that the mean number of TVs in US homes is not equal to 3

Example: p-Value

- Example: How likely is it to see a sample mean of 2.84 (or something further from the mean, in either direction) if the true mean is μ = 3.0?

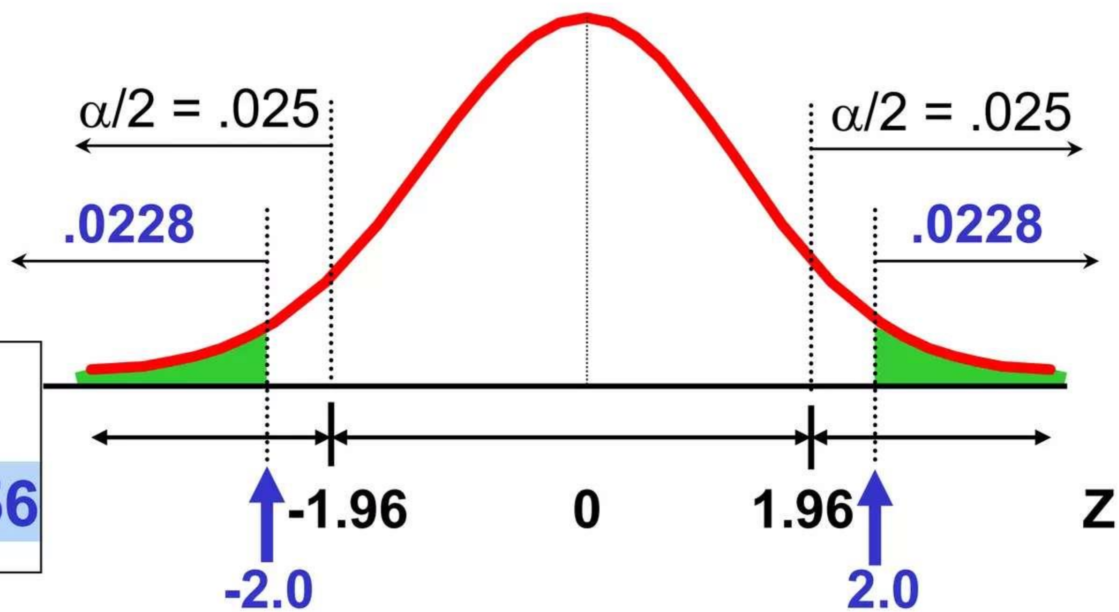
$\overline{x}$ = 2.84 is translated to a z score of z = -2.0

$P(z < -2.0) = .0228$

$P(z > 2.0) = .0228$

p-value

= .0228 + .0228 = .0456



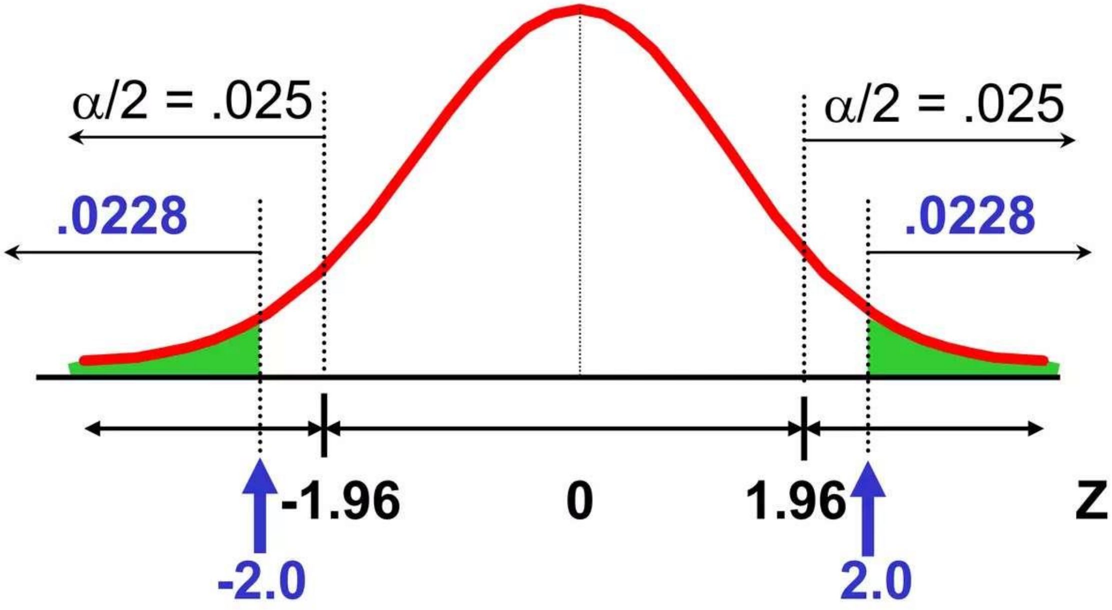
Example: p-Value

(continued)

- Compare the p-value with α
  - If p-value $<$ α , reject H_0
  - If p-value $\geq$ α , do not reject H_0

Here: p-value = .0456
α = .05

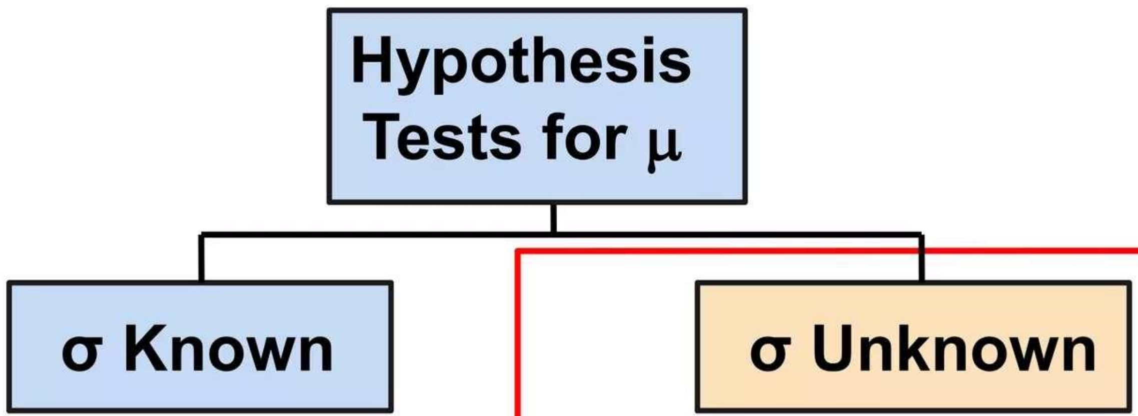
Since .0456 < .05, we reject the null hypothesis



t Test of Hypothesis for the Mean (σ Unknown)

9.3

- Convert sample result ($\bar{x}$) to a t test statistic



Consider the test

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu > \mu_0$$

(Assume the population is normal)

The decision rule is:

$$\text{Reject } H_0 \text{ if } \quad t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} > t_{n-1,\,\alpha}$$

t Test of Hypothesis for the Mean (σ Unknown)

(continued)

- For a two-tailed test:

Consider the test

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

(Assume the population is normal, and the population variance is unknown)

The decision rule is:

Reject H_0 if $t = \dfrac{\bar{x} - \mu_0}{\dfrac{s}{\sqrt{n}}} < -t_{n-1,\, \alpha/2}$ or if $t = \dfrac{\bar{x} - \mu_0}{\dfrac{s}{\sqrt{n}}} > t_{n-1,\, \alpha/2}$

Example: Two-Tail Test (σ Unknown)

The average cost of a hotel room in Chicago is said to be \$168 per night. A random sample of 25 hotels resulted in $\bar{x}$ = \$172.50 and s = \$15.40. Test at the α = 0.05 level.

(Assume the population distribution is normal)

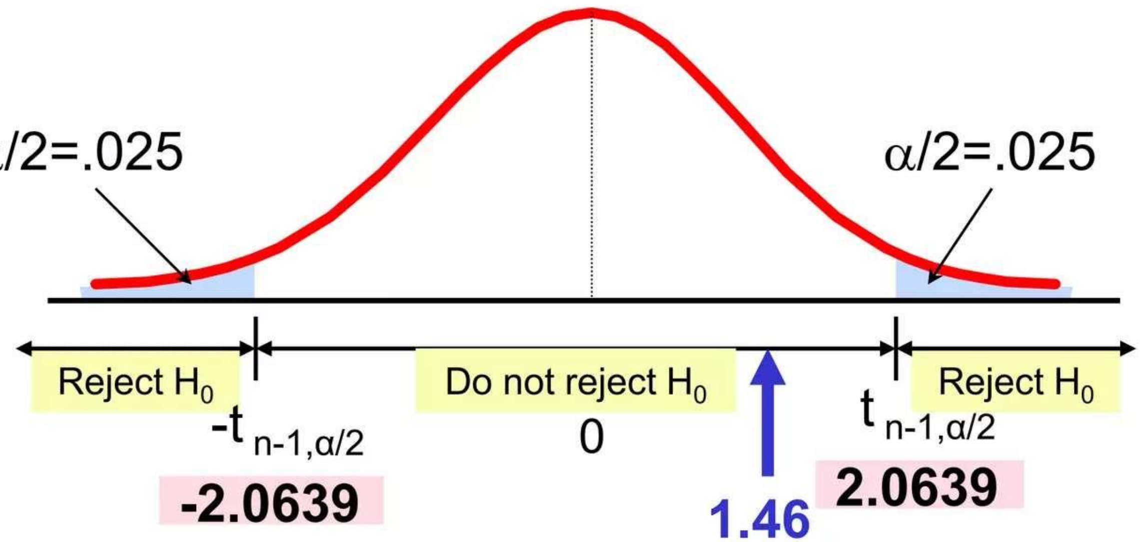
H_0: $\mu = 168$

H_1: $\mu \neq 168$

Week 17

Example Solution: Two-Tail Test

H_0: μ = 168
H_1: μ ≠ 168



- α = 0.05
- n = 25
- σ is unknown, so use a **t statistic**
- Critical Value:

$t_{24\ ,\ .025} = \pm\ 2.0639$

$$t_{n-1} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{172.50 - 168}{\frac{15.40}{\sqrt{25}}} = 1.46$$

Do not reject H_0: not sufficient evidence that true mean cost is different than $168

9.4

Tests of the Population Proportion

- Involves categorical variables
- Two possible outcomes
  - "Success" (a certain characteristic is present)
  - "Failure" (the characteristic is not present)
- Fraction or proportion of the population in the "success" category is denoted by P
- Assume sample size is large

Proportions

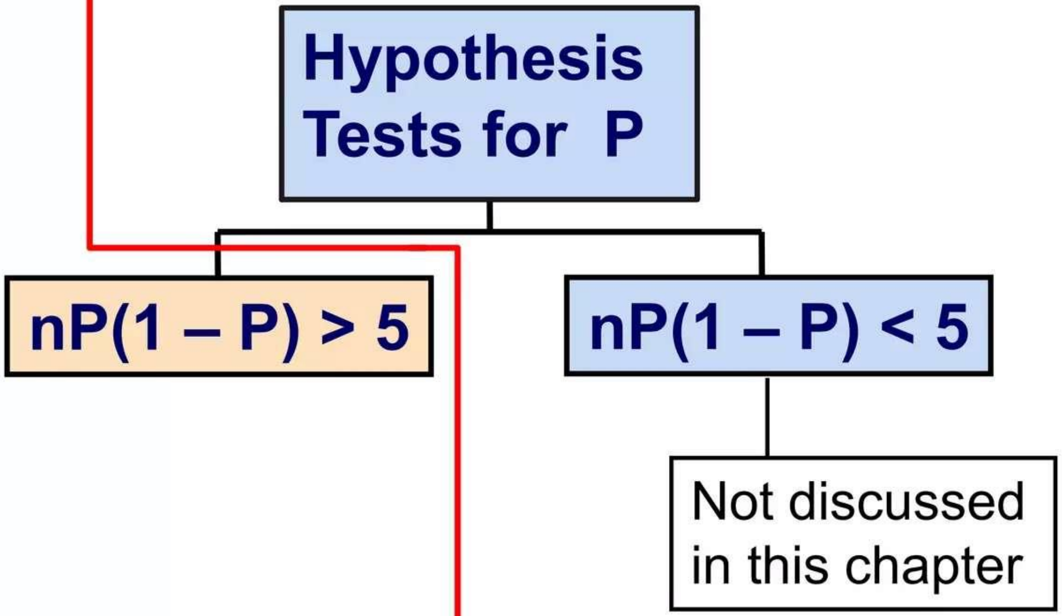
(continued)

- Sample proportion in the success category is denoted by $\hat{p}$
  - $$\hat{p} = \frac{x}{n} = \frac{\text{number of successes in sample}}{\text{sample size}}$$
- When $nP(1 – P) > 5$, $\hat{p}$ can be approximated by a normal distribution with mean and standard deviation
  - $$\mu_{\hat{p}} = P \qquad \sigma_{\hat{p}} = \sqrt{\frac{P(1-P)}{n}}$$

Hypothesis Tests for Proportions

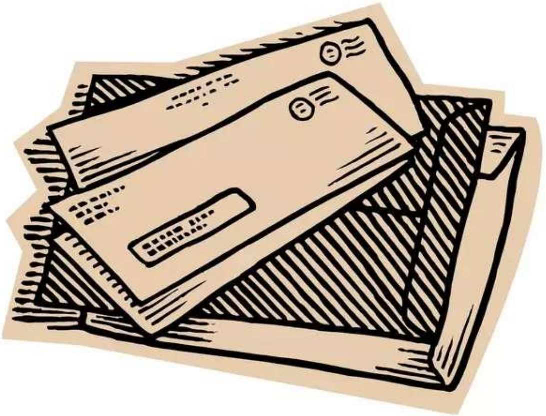
- The sampling distribution of $\hat{p}$ is approximately normal, so the test statistic is a z value:

$$z = \frac{\hat{p} - P_0}{\sqrt{\frac{P_0(1 - P_0)}{n}}}$$



Example: Z Test for Proportion

A marketing company claims that it receives 8% responses from its mailing. To test this claim, a random sample of 500 were surveyed with 25 responses. Test at the $\alpha = .05$ significance level.



Check:

Our approximation for P is

$\hat{p} = 25/500 = .05$

$nP(1 - P) = (500)(.05)(.95)$
$= 23.75 > 5$ ✓

Z Test for Proportion: Solution

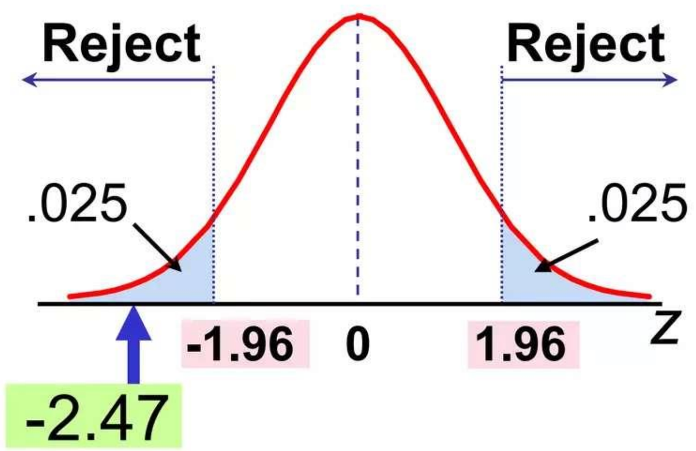
H_0: P = .08
H_1: P ≠ .08

α = .05
n = 500, $\hat{p}$ = .05

Test Statistic:

$$z = \frac{\hat{p} - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} = \frac{.05 - .08}{\sqrt{\frac{.08(1-.08)}{500}}} = -2.47$$

Critical Values: ± 1.96



Decision:

Reject H_0 at α = .05

Conclusion:

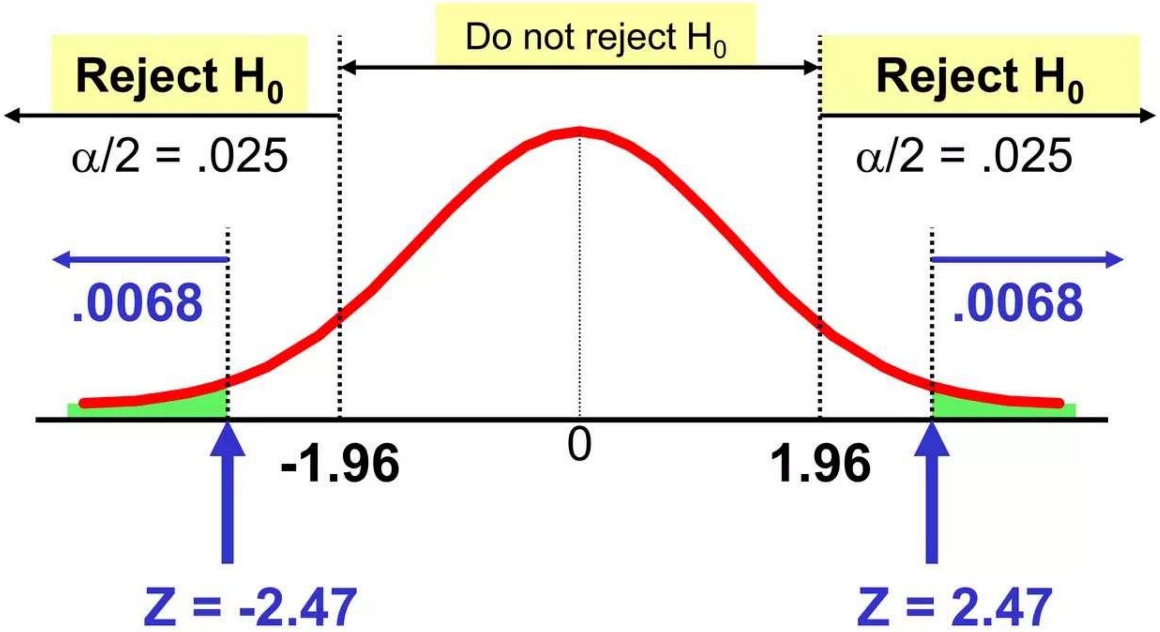
There is sufficient evidence to reject the company's claim of 8% response rate.

p-Value Solution

(continued)

Calculate the p-value and compare to α

(For a two sided test the p-value is always two sided)



p-value = .0136:

$$P(Z \le -2.47) + P(Z \ge 2.47)$$

$$= 2(.0068) = 0.0136$$

Reject H_0 since p-value = .0136 < α = .05

9.5

Power of the Test

- Recall the possible hypothesis test outcomes:

	Actual Situation	
Decision	H_0 True	H_0 False
Do Not Reject H_0	**No error** **(1 - α)**	**Type II Error** **(β)**
Reject H_0	**Type I Error** **(α)**	**No Error** **(1 - β)**

Key:
Outcome
(Probability)

- β denotes the probability of Type II Error
- $1 - \beta$ is defined as the power of the test

Power = $1 - \beta$ = the probability that a false null hypothesis is rejected

Type II Error

Assume the population is normal and the population variance is known. Consider the test

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu > \mu_0$$

The decision rule is:

$$\text{Reject } H_0 \text{ if } z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} > z_\alpha$$

or

$$\text{Reject } H_0 \text{ if } \bar{x} = \bar{x}_c > \mu_0 + Z_\alpha \sigma / \sqrt{n}$$

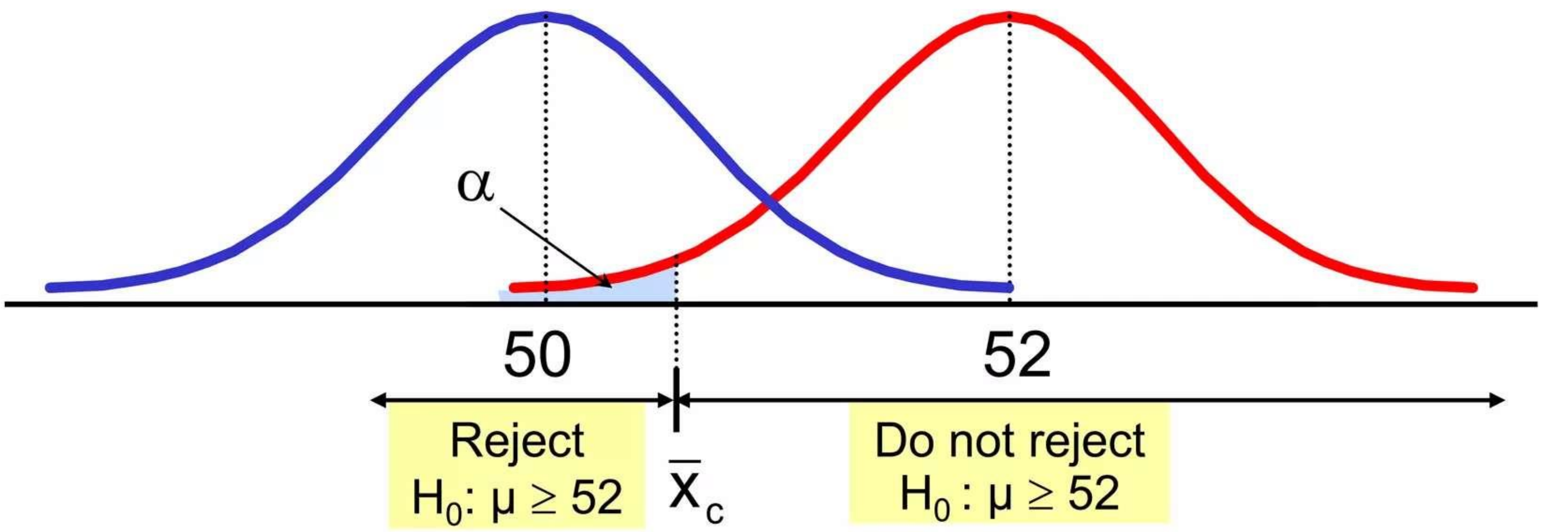
If the null hypothesis is false and the true mean is μ*, then the probability of type II error is

$$\beta = P(\bar{x} < \bar{x}_c \mid \mu = \mu^*) = P\left(z < \frac{\bar{x}_c - \mu^*}{\sigma / \sqrt{n}} \right)$$

Type II Error Example

- Type II error is the probability of failing to reject a false H_0

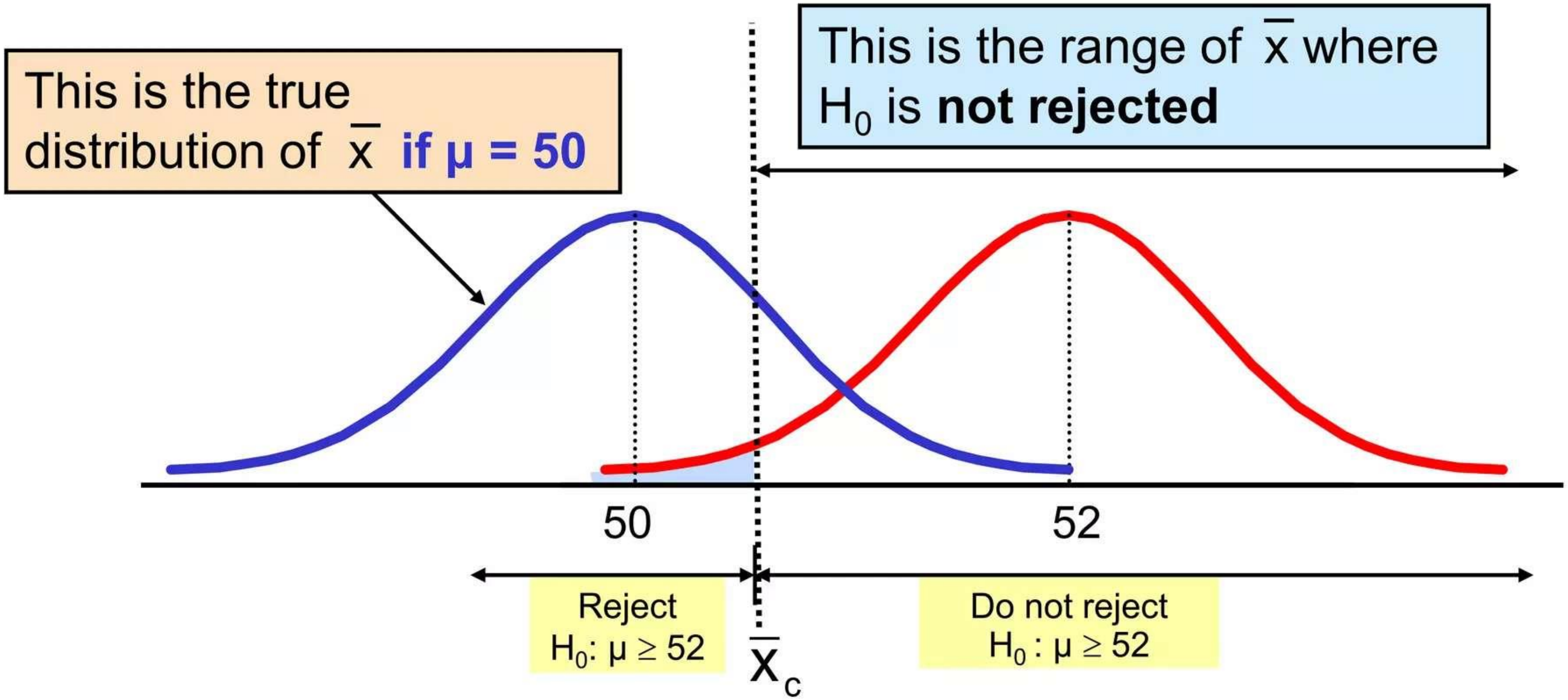
Suppose we fail to reject **H_0: μ ≥ 52** when in fact the true mean is **μ* = 50**



Type II Error Example

(continued)

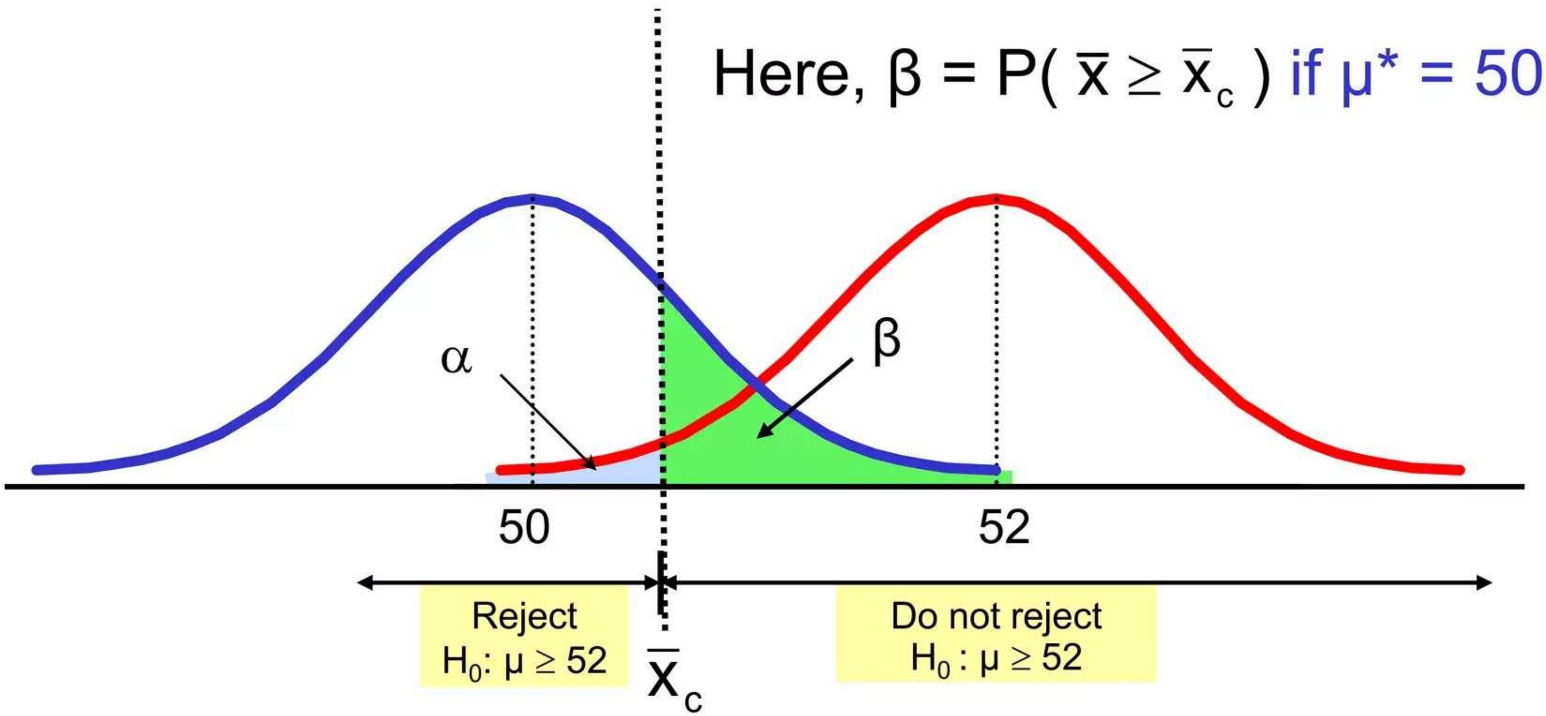
- Suppose we do not reject H_0: $\mu \geq 52$ when in fact the true mean is $\mu^* = 50$



Type II Error Example

(continued)

- Suppose we do not reject H_0: $\mu \geq 52$ when in fact the true mean is $\mu^* = 50$

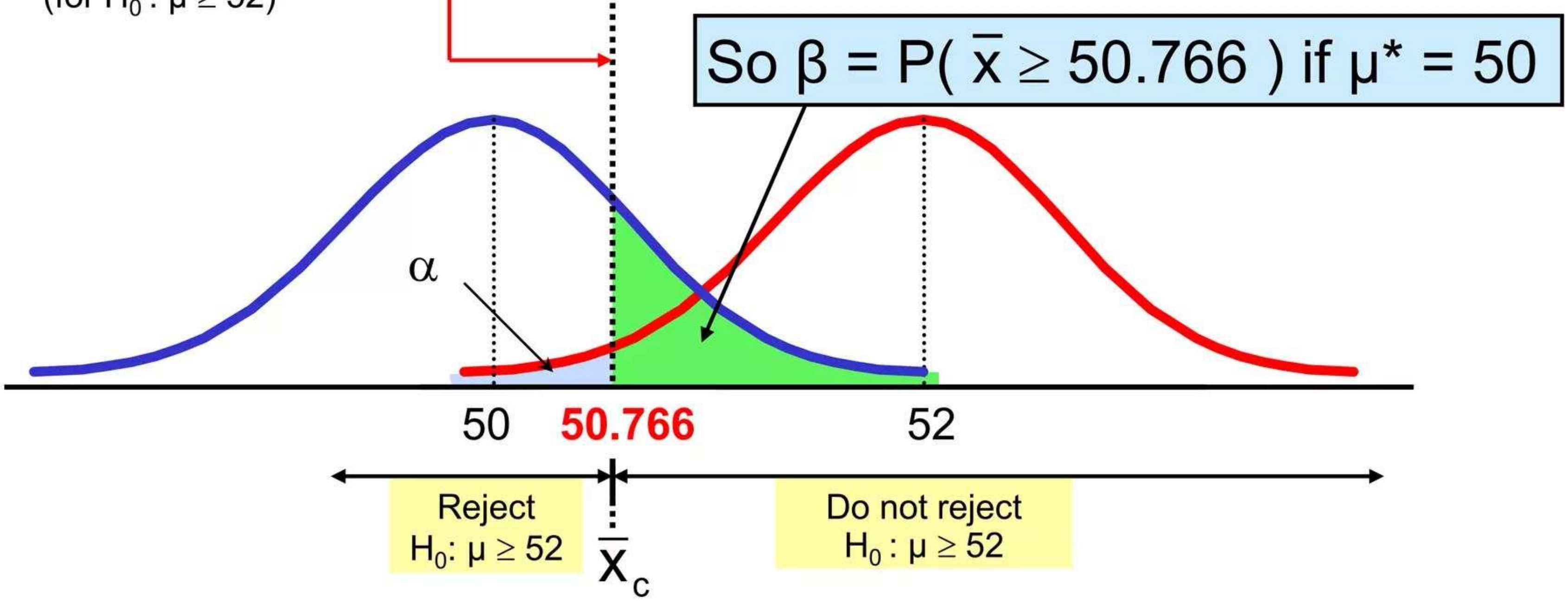


Calculating β

- Suppose n = 64 , σ = 6 , and α = .05

$$\bar{x}_c = \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}} = 52 - 1.645 \frac{6}{\sqrt{64}} = \boxed{50.766}$$

(for $H_0 : \mu \geq 52$)

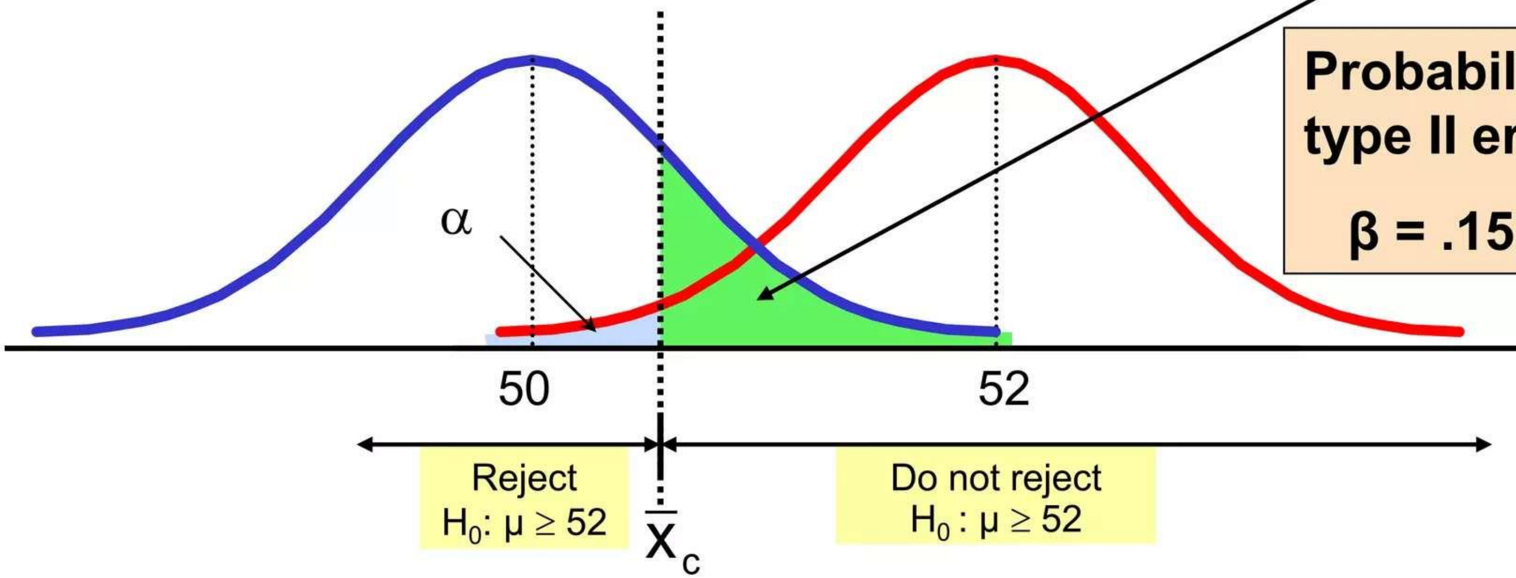


Calculating β

(continued)

- Suppose n = 64 , σ = 6 , and α = .05

$$P(\bar{x} \geq 50.766 | \mu^* = 50) = P\left(z \geq \frac{50.766 - 50}{6/\sqrt{64}}\right) = P(z \geq 1.02) = .5 - .3461 = .1539$$



Probability of type II error:

β = .1539

Power of the Test Example

If the true mean is $\mu^* = 50$,

- The probability of Type II Error = $\beta = 0.1539$
- The power of the test = $1 - \beta = 1 - 0.1539 = 0.8461$

Key:
Outcome
(Probability)

	Actual Situation	
Decision	H_0 True	H_0 False
Do Not Reject H_0	**No error** **$1 - \alpha = 0.95$**	**Type II Error** **$\beta = 0.1539$**
Reject H_0	**Type I Error** **$\alpha = 0.05$**	**No Error** **$1 - \beta = 0.8461$**

(The value of β and the power will be different for each μ^*)

Hypothesis Tests of one Population Variance

9.6

- Goal: Test hypotheses about the population variance, σ^2

  - If the population is normally distributed,

$$\chi^2_{n-1} = \frac{(n-1)s^2}{\sigma^2}$$

has a chi-square distribution with (n – 1) degrees of freedom

Hypothesis Tests of one Population Variance

(continued)

The test statistic for hypothesis tests about one population variance is

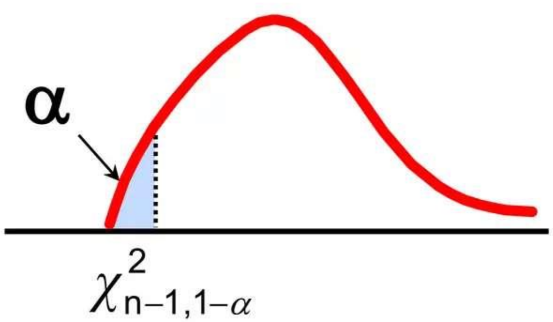
$$\chi^2_{n-1} = \frac{(n-1)s^2}{\sigma_0^2}$$

Decision Rules: Variance

Population variance

Lower-tail test:

$H_0: \sigma^2 \geq \sigma_0^2$
$H_1: \sigma^2 < \sigma_0^2$

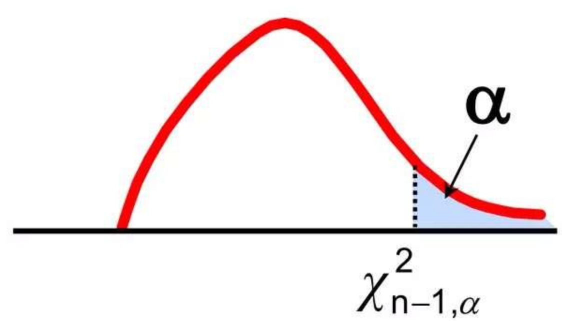


Reject H_0 if

$\chi^2_{n-1} < \chi^2_{n-1,1-\alpha}$

Upper-tail test:

$H_0: \sigma^2 \leq \sigma_0^2$
$H_1: \sigma^2 > \sigma_0^2$

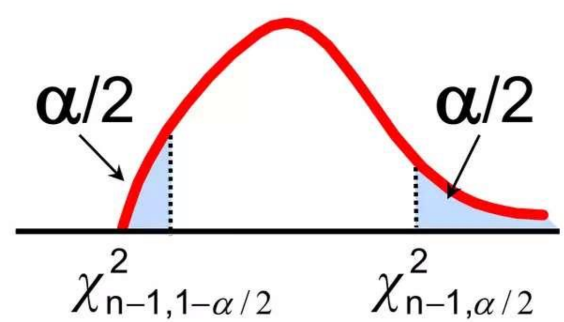


Reject H_0 if

$\chi^2_{n-1} > \chi^2_{n-1,\alpha}$

Two-tail test:

$H_0: \sigma^2 = \sigma_0^2$
$H_1: \sigma^2 \neq \sigma_0^2$



Reject H_0 if

$\chi^2_{n-1} > \chi^2_{n-1,\alpha/2}$

or

$\chi^2_{n-1} < \chi^2_{n-1,1-\alpha/2}$