

Index

Unit	Chapter	Particulars	Page No.
1	1	Introduction to Business Statistics: <ul style="list-style-type: none"> ● Statistics in Business, ● Types of data – Nominal, Ordinal, Interval, Ratio. ● Types of variables – Dependent, independent, moderating, intervening, extraneous. Discrete / continuous. 	3
	2	Charts and Graphs.	9
	3	Descriptive Statistics: <ul style="list-style-type: none"> ● Measure of central tendency – mean, median, quartile, mode (for Group and ungrouped data) ● Measure of variability – Range, interquartile range, standard deviation, variance, coefficient of variation, (for Group and ungrouped data) ● Measures of shape – kurtosis, skewness, boxplot. 	18
	4	Probability: <ul style="list-style-type: none"> ● Introduction to probability ● Theories of probability – Classical, Relative frequency and subjective. ● Laws of probability – addition, multiplication. ● Inverse Probability. ● Revision of probability: BAYES' RULE 	42
2	1	Probability Distribution: Discrete distribution – Binomial, Poisson.	66
		Probability Distribution: Continuous distribution – Uniform, normal.	83



	2	Hypothesis testing: <ul style="list-style-type: none">● Types of hypothesis – research, statistical, substantive.● Null and alternative hypothesis.● One-tailed & Two-tailed test.● Types of Error – Type I & Type II.● Level of significance.● Steps of hypothesis testing.	98
3	1	Parametric Tests: Uni-variate tests: z-test	122
	2	T-test	128
	3	Levene's F-test	130
	4	Bi-variate tests: T-test – Paired	133
		T-independent	138
		Simple Linear Regression,	144
		One Way ANOVA	153
4	1	Non-Parametric Tests: Uni-variate tests: Chi-square goodness of fit for uniform distribution	160
	2	Mann-Whitney U test	180
	3	Wilcoxon Sign Paired Rank Test	190
	4	Kruskal-Wallis	200
	5	Friedman's test	208
	6	Spearman's Rank Correlation	213



	7	Multivariate analysis: Overview of Multiple Regression, Factor Analysis, Multidimensional scaling, Discriminant analysis. (theoretical concepts only)	241
--	----------	--	------------

❖ Introduction to Statistics

India is the second largest country in the world, with more than a billion people. Three quarters of the people live in rural areas, yet the rural market accounts for only about one third of total national product sales. However, because of free-market reforms in the 1990s and a strong agricultural output, India's rural market has become more open for trade in consumer goods. Although India's urban market seems to be saturated, markets in rural India are relatively untapped, offering enormous potential. Because of these factors, many U.S. firms, such as Microsoft, General Electric, Kellogg's, and others, have entered the Indian market.

Presently, rural India can be described as poor and semi-illiterate. More than 65% of the people in rural India earn less than \$574 annually, and 23% earn between \$574 and \$1,146. Sixty-six percent of the women are illiterate, as are 38% of the men. These rates are about double those for urban Indians. Seventy-seven percent of the households in rural India use wood as the cooking fuel, 39% have electricity, 18% have piped water, and 7% have flush toilets.

Nevertheless, conditions are changing and companies are moving into this relatively untapped market. For example, by the late 1990s, Colgate-Palmolive planned to increase its rural marketing budget to five times that of 1991. Colgate-Palmolive India's goal is that more than half its revenue by the year 2003 comes from rural India, which presently accounts for only about 30% of business.

Marketing to rural India is a challenging task and requires some nontraditional approaches because the illiteracy rates are high and only about one-third of the households have a television. One such technique is the use of video vans in which half-hour infomercials are carried through the countryside. A video van cruises into a small hamlet with speakers playing a popular movie melody. As shoppers congregate to the van, a marketer opens the door and plays a video on a screen with scenarios depicting the need for a particular product. After the video is completed, free samples are distributed. Hindustan Lever Ltd., India's leading consumer-products company, estimates that the cost per contact of such marketing is about four times the cost to city dwellers. However, the rural market for personal care products is growing about three times faster than city markets, which makes such marketing efforts more viable. Other companies use direct door-to-door campaigns to promote products to rural India. In addition, the advent of satellite television to rural homes and villages in India opens up some new avenues for advertising and marketing to this population segment.



Statistics available from the first half of the 1990s shed some light on the potential market of rural India. Toothpaste consumption in rural India doubled from 8,825 metric tons in 1990 to 17,023 in 1994. The annual per capita consumption for toothpaste is still only 30 grams per person in rural India compared to 160 grams in urban India and 400 grams in the United States. Thus, the potential for much growth is there. Sales of other products have been growing rapidly in this emerging market. The sales of laundry detergent increased from 272,540 metric tons in 1990 to 422,741 metric tons in 1994. Toilet soap went from 158,919 metric tons in 1990 to 231,084 metric tons in 1994. Shampoo increased in sales nearly fourfold from 497,000 liters in 1990 to 2,116,000 liters in 1994.

Rural India is a huge untapped market for businesses. Some evidence indicates that rural Indian consumers are buying products in increasing numbers. However, annual income statistics show a limited purchasing capacity. The dilemma facing companies is whether to enter this marketplace, and if so, to what extent and how.

Virtually every area of business uses statistics in decision making. Here are some recent examples:

■ According to a TNS Retail Forward ShopperScape survey, the average amount spent by a shopper on electronics in a three-month period is \$629 at Circuit City, \$504 at Best Buy, \$246 at Wal-Mart, \$172 at Target, and \$120 at RadioShack.

■ A survey of 1465 workers by Hotjobs reports that 55% of workers believe that the quality of their work is perceived the same when they work remotely as when they are physically in the office.

■ A survey of 477 executives by the Association of Executive Search Consultants determined that 48% of men and 67% of women say they are more likely to negotiate for less travel compared with five years ago.

■ A survey of 1007 adults by RBC Capital Markets showed that 37% of adults would be willing to drive 5 to 10 miles to save 20 cents on a gallon of gas.

■ A Deloitte Retail “Green” survey of 1080 adults revealed that 54% agreed that plastic, non-compostable shopping bags should be banned.



■ A recent Household Economic Survey by Statistic New Zealand determined that the average weekly household net expenditure in New Zealand was \$956 and that households in the Wellington region averaged \$120 weekly on recreation and culture. In addition, 75% of all households were satisfied or very satisfied with their material standard of living.

■ The Experience's Life After College survey of 320 recent college graduates showed that 58% moved back home after college. Thirty-two percent then remained at home for more than a year.

You can see from these few examples that there is a wide variety of uses and applications of statistics in business. Note that in most of these examples, business researchers have conducted a study and provided us rich and interesting information.

In this text we will examine several types of graphs for depicting data as we study ways to arrange or structure data into forms that are both meaningful and useful to decision makers. We will learn about techniques for sampling from a population that allow studies of the business world to be conducted more inexpensively and in a more timely manner. We will explore various ways to forecast future values and examine techniques for predicting trends. This text also includes many statistical tools for testing hypotheses and for estimating population values. These and many other exciting statistics and statistical techniques await us on this journey through business statistics.

Data Measurements

In statistics, level of measurement is a classification that relates the values that are assigned to variables with each other. In other words, level of measurement is used to describe information within the values. Psychologist Stanley Smith is known for developing four levels of measurement: nominal, ordinal, interval, and ratio.

Four common levels of data measurement follow.

1. Nominal
2. Ordinal
3. Interval
4. Ratio

1. Nominal scales

Nominal scales contain the least amount of information. In nominal scales, the numbers assigned to each variable or observation are only used to classify the variable or observation. For example, a fund manager may choose to assign the number 1 to small-cap stocks, the number 2 to corporate bonds, the number 3 to derivatives, and so on.



2. Ordinal scales

Ordinal scales present more information than nominal scales and are, therefore, a higher level of measurement. In ordinal scales, there is an ordered relationship between the variable's observations. For example, a list of 500 managers of mutual funds may be ranked by assigning the number 1 to the best-performing manager, the number 2 to the second best-performing manager, and so on.

With this type of measurement, one can conclude that the number 1-ranked mutual fund manager performed better than the number 2-ranked mutual fund manager.

3. Interval scales

4. Interval scales present more information than ordinal scales in that they provide assurance that the differences between values are equal. In other words, interval scales are ordinal scales but with equivalent scale values from low to high intervals.

For example, temperature measurement is an example of an interval scale: 60°C is colder than 65°C, and the temperature difference is the same as the difference between 50°C and 55°C. In other words, the difference of 5°C in both intervals shares the same interpretation and meaning.

Consider why the ordinal scale example is not an interval scale: A fund manager ranked 1 probably did not outperform the fund manager ranked 2 by the exact same amount that a fund manager ranked 6 outperformed a fund manager ranked 7. Ordinal scales provide a relative ranking, but there is no assurance that the differences between the scale values are the same.

A drawback in interval scales is that they do not have a true zero point. Zero does not represent an absence of something in an interval scale. Consider that the temperature -0°C does not represent the absence of temperature. For this reason, interval-scale-based ratios fail to provide some insights – for example, 50°C is not twice as hot as 25°C.

4. Ratio scales

Ratio scales are the most informative scales. Ratio scales provide rankings, assure equal differences between scale values, and have a true zero point. In essence, a ratio scale can be thought of as nominal, ordinal, and interval scales combined as one.

For example, the measurement of money is an example of a ratio scale. An individual with \$0 has an absence of money. With a true zero point, it would be correct to say that someone with \$100 has twice as much money as someone with \$50.

Types of Variables

❖ Independent Variables

The independent variable is the one that is computed in research to view the impact of dependent variables. It is also called as resultant variables, predictor or experimental variables. For example, A manager asks 100 employees to complete a project. He should know the capacity of the individual employee. He wants to know the reason behind smart guys and failure guys. The first reason is that some will be working hard for day and night to complete the project within the estimated time, and the other one is that some guys are born intelligent and smarter than others. The variable which is similar to an independent variable is called a covariate variable but is impacted by the dependent variable but not as common as a variable of interest.

❖ Dependent Variables

The dependent variable is also called a criterion variable which is applied in non-experimental circumstances. The dependent variable has relied on the independent variable. From the above-mentioned example, the project's productivity or completion is the main criteria that are dependent on estimated time and IQ. Here, the independent variables are IQ and estimated time, which may or may not reflect in an employee's productivity. So the extension of estimated time or enhancing the IQ of a person doesn't make any sense in employee's productivity as it is not predictable.

Hence, the managers' focus is to work on the independent variables such as allotted time and IQ that leads to certain changes in employee's productivity that are the dependent variables. So both the variables are connected in some measures. The variables which get affected by other variables in econometrics is termed as endogenous variables. A hidden variable impacts the relationship between the dependent and independent variable called lurking variables. When an independent variable is not impacted by any other variables and is restricted to a certain extent are called an explanatory variable.

❖ Intervening variables

An intervening variable, sometimes called a mediator variable, is a theoretical variable the researcher uses to explain a cause or connection between other study variables—usually dependent and independent ones. They are associations instead of observations. For example, if wealth is the independent variable, and a long life span is a dependent variable, the researcher might hypothesize that access to quality healthcare is the intervening variable that links wealth and life span.

❖ Moderating variables

A moderating or moderator variable changes the relationship between dependent and independent variables by strengthening or weakening the intervening variable's effect. For example, in a study looking at the relationship between economic status (independent variable) and how frequently people get physical exams from a doctor (dependent variable), age is a moderating variable. That relationship might be weaker in younger individuals and stronger in older individuals.

❖ Extraneous variables

Extraneous variables are factors that affect the dependent variable but that the researcher did not originally consider when designing the experiment. These unwanted variables can unintentionally change a study's results or how a researcher interprets those results. Take, for example, a study assessing whether private tutoring or online courses are more effective at improving students' Spanish test scores. Extraneous variables that might unintentionally influence the outcome include parental support, prior knowledge of a foreign language or socioeconomic status.

❖ Discrete

Quantitative discrete variables are variables for which the values it can take are **countable** and have a **finite number of possibilities**. The values are often (but not always) integers. Here are some examples of discrete variables:

- Number of children per family
- Number of students in a class
- Number of citizens of a country

Even if it would take a long time to count the citizens of a large country, it is still technically doable. Moreover, for all examples, the number of possibilities is **finite**. Whatever the number of children in a family, it will never be 3.58 or 7.912 so the number of possibilities is a finite number and thus countable.

❖ Continuous

On the other hand, **quantitative continuous** variables are variables for which the values are **not countable** and have an **infinite number of possibilities**. For example:

- Age
- Weight
- Height

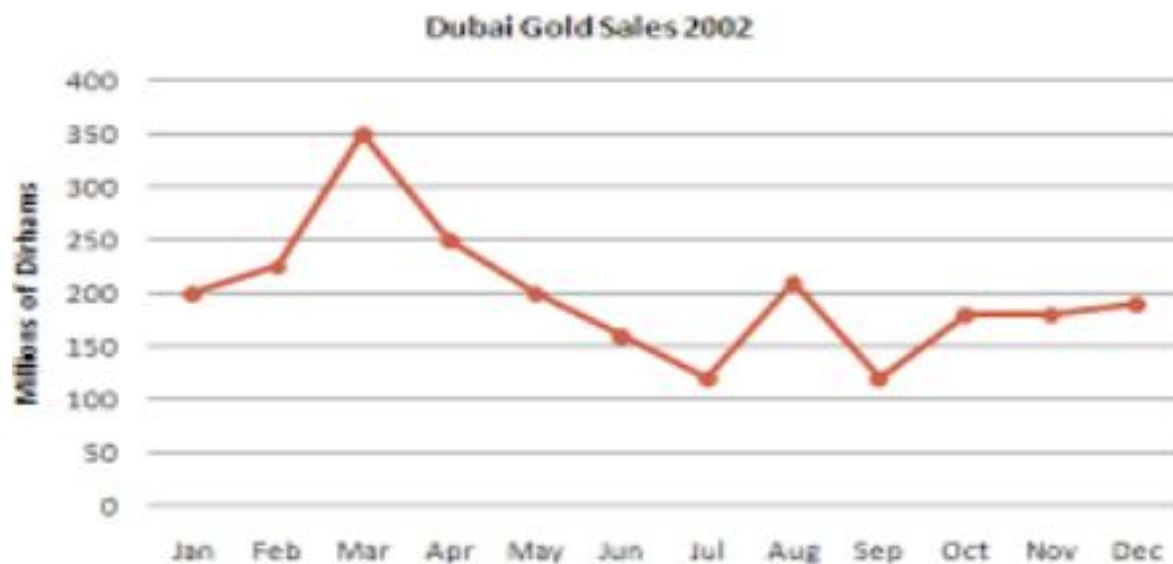
For simplicity, we usually referred to years, kilograms (or pounds) and centimeters (or feet and inches) for age, weight and height respectively. However, a 28-year-old man could actually be 28 years, 7 months, 16 days, 3 hours, 4 minutes, 5 seconds, 31 milliseconds, 9 nanoseconds old.

For all measurements, we usually stop at a standard level of granularity, but nothing (except our measurement tools) prevents us from going deeper, leading to an **infinite number of potential values**. The fact that the values can take an infinite number of possibilities makes it uncountable.

❖ Types of Diagrams

● (a) Line Diagrams

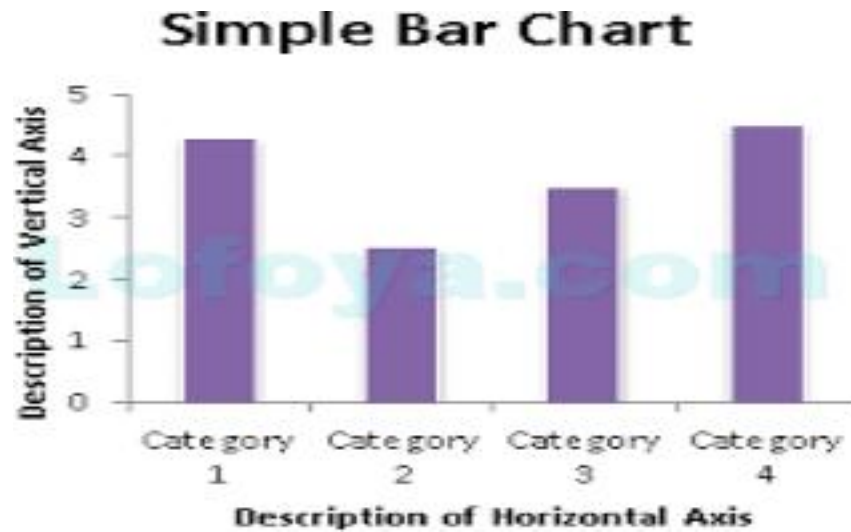
In these diagrams only line is drawn to represent one variable. These lines may be vertical or horizontal. The lines are drawn such that their length is the proportion to value of the terms or items so that comparison may be done easily



● (b) Simple Bar Diagram

Like line diagrams these figures are also used where only single dimension i.e. length can present the data. Procedure is almost the same, only one thickness of lines is measured. These can also be drawn either vertically or horizontally. Breadth of these lines or bars should be equal.

Similarly distance between these bars should be equal. The breadth and distance between them should be taken according to space available on the paper.

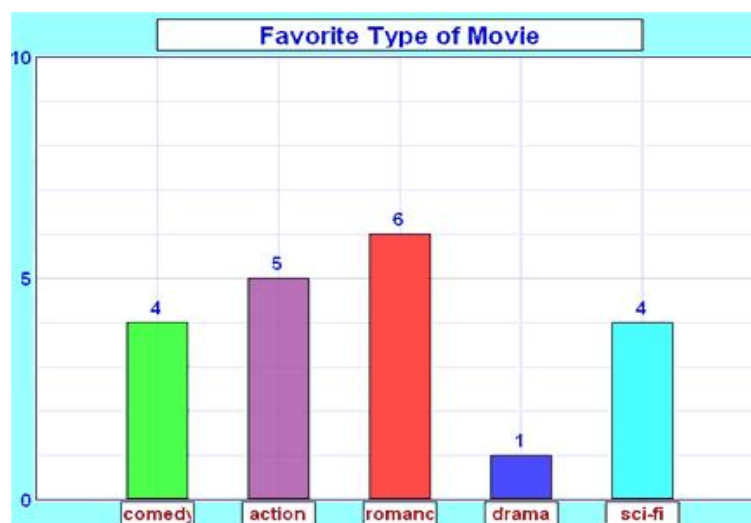


Imagine you just did a survey of your friends to find which kind of movie they liked best:

Table: Favorite Type of Movie

Comedy	Action	Romance	Drama	SciFi
4	5	6	1	4

We can show that on a bar graph like this:



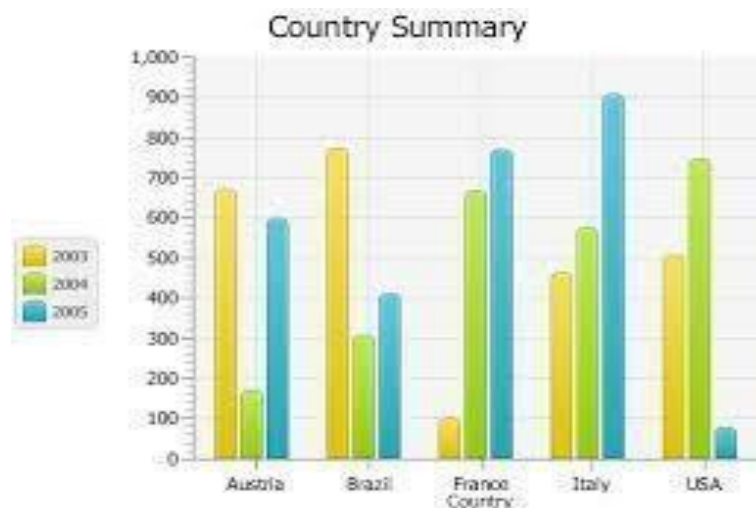
It is a really good way to show relative sizes: we can see which types of movie are most liked, and which are least liked, at a glance.

We can use bar graphs to show the relative sizes of many things, such as what type of car people have, how many customers a shop has on different days and so on.

● (c) Multiple Bar Diagrams

The diagram is used, when we have to make comparison between more than two variables. The number of variables may be 2, 3 or 4 or more. In case of 2 variables, pair of bars is drawn.

Similarly, in case of 3 variables, we draw triple bars. The bars are drawn on the same proportionate basis as in case of simple bars. The same shade is given to the same item.



Draw a multiple bar chart to represent the import and export of Canada (values in \$) for the years 1991 to 1995

Years	Imports	Exports
1991	7930	4260
1992	8850	5225
1993	9780	6150
1994	11720	7340
1995	12150	8145

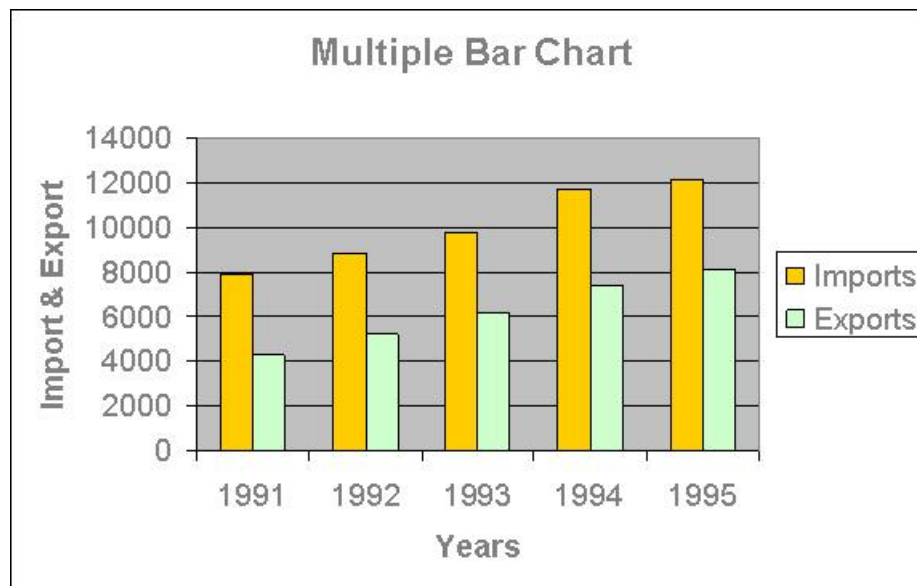


chart showing the import and export of Canada from 1991 – 1995.

Advantages

The chief advantages of a bar diagram can be outlined as under:

1. It is very simple to draw and read as well.
2. It is the only form of diagram which can represent a large number of data on a piece of paper.
3. It can be drawn both vertically and horizontally.
4. It gives a better look and facilitates comparison.

Disadvantages

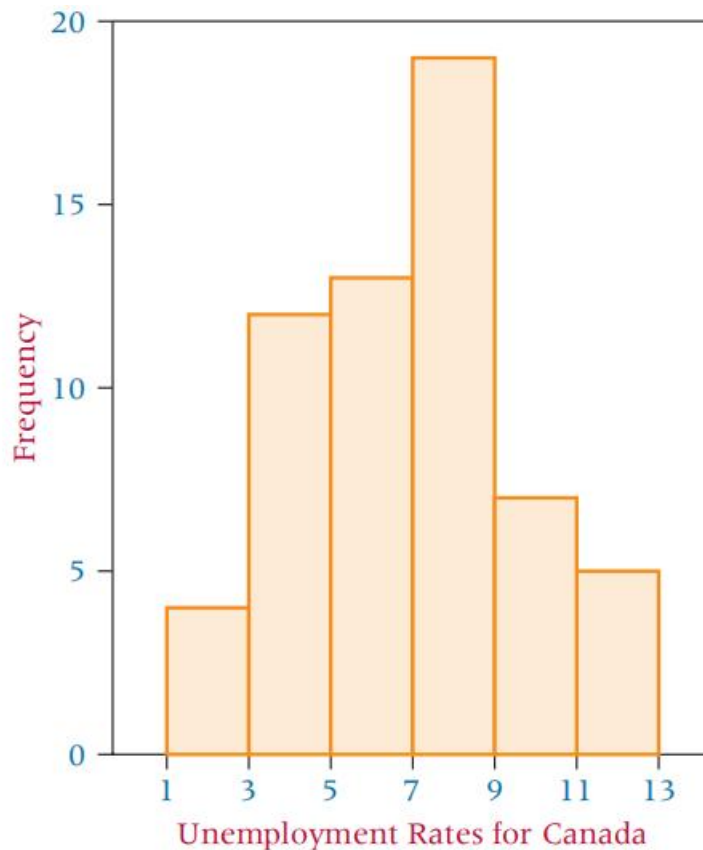
1. It cannot exhibit a large number of aspects of the data.
2. The width of the bars are fixed arbitrarily by a drawer.

● Histograms

One of the more widely used types of graphs for quantitative data is the **histogram**. A histogram is a series of contiguous bars or rectangles that represent the frequency of data in given class intervals. If the class intervals used along the horizontal axis are equal, then the height of the bars represent the frequency of values in a given class interval. If the class intervals are unequal, then the areas of the bars (rectangles) can be used for relative comparisons of class frequencies. Construction of a histogram involves labeling the x-axis

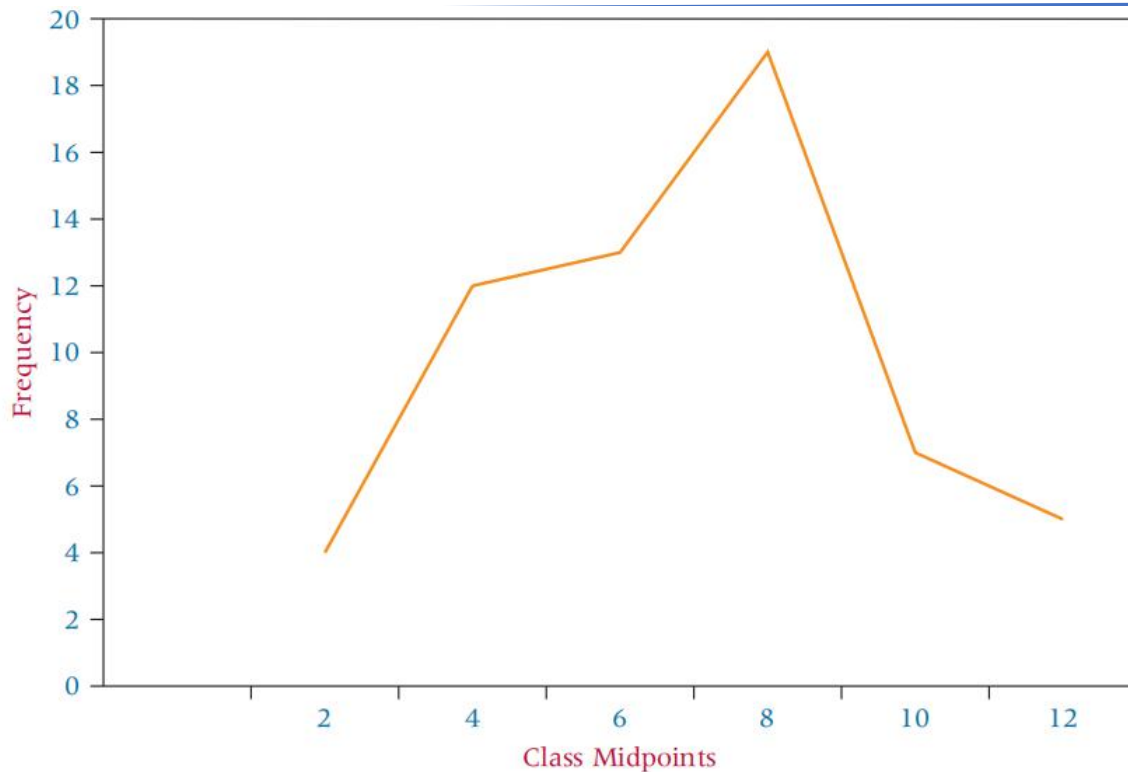
(abscissa) with the class endpoints and the y-axis (ordinate) with the frequencies, drawing a horizontal line segment from class endpoint to class endpoint at each frequency value, and connecting each line segment vertically from the frequency value to the x-axis to form a series of rectangles (bars).

A histogram is a useful tool for differentiating the frequencies of class intervals. A quick glance at a histogram reveals which class intervals produce the highest frequency totals.



● Frequency Polygons

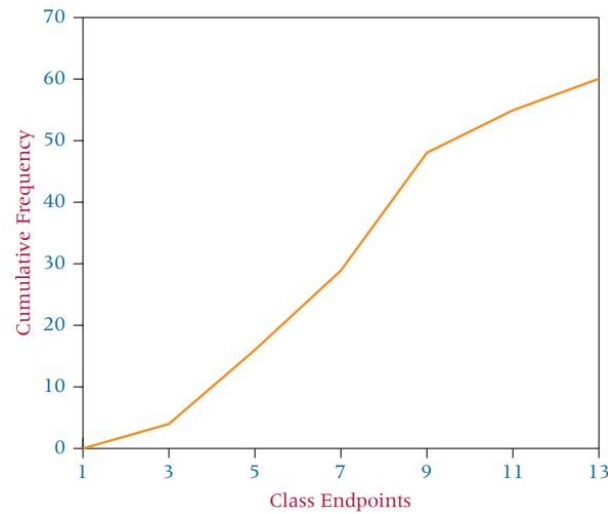
A **frequency polygon**, like the histogram, is a graphical display of class frequencies. However, instead of using bars or rectangles like a histogram, in a frequency polygon each class frequency is plotted as a dot at the class midpoint, and the dots are connected by a series of line segments. Construction of a frequency polygon begins by scaling class midpoints along the horizontal axis and the frequency scale along the vertical axis. A dot is plotted for the associated frequency value at each class midpoint. Connecting these midpoint dots completes the graph.



● Ogives

An **ogive** (o-jive) is a cumulative frequency polygon. Construction begins by labeling the x-axis with the class endpoints and the y-axis with the frequencies. However, the use of cumulative frequency values requires that the scale along the y-axis be great enough to include the frequency total. A dot of zero frequency is plotted at the beginning of the first class, and construction proceeds by marking a dot at the end of each class interval for the cumulative value. Connecting the dots then completes the ogive.

Ogives are most useful when the decision maker wants to see running totals. For example, if a comptroller is interested in controlling costs, an ogive could depict cumulative costs over a fiscal year.

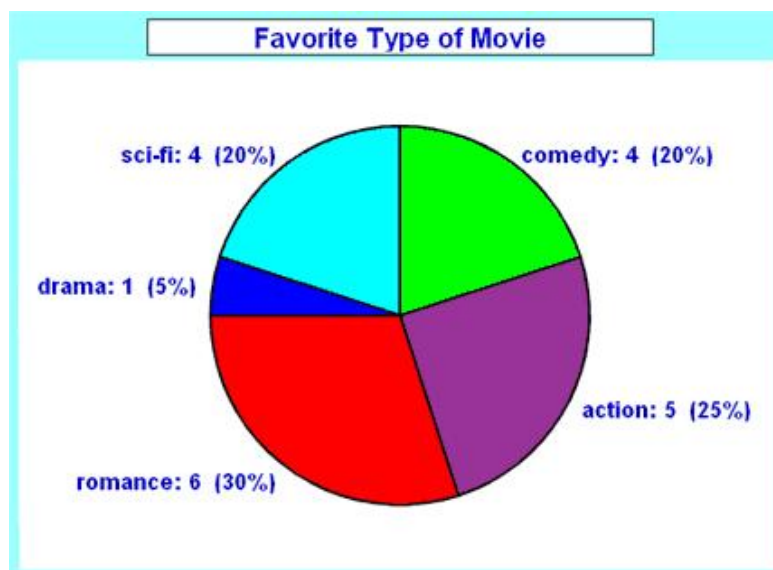


- **Pie Chart:** a special chart that uses "pie slices" to show relative sizes of data.

Imagine you survey your friends to find the kind of movie they like best:

Table: Favorite Type of Movie

Comedy	Action	Romance	Drama	SciFi
4	5	6	1	4



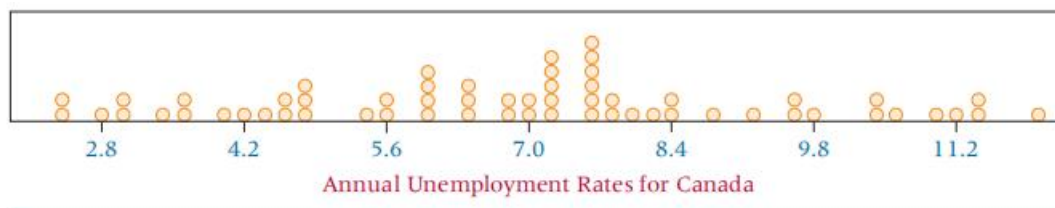
You can show the data by this Pie Chart:

It is a really good way to show relative sizes: it is easy to see which movie types are most liked, and which are least liked, at a glance.

● Stem-and-Leaf Plots

Another way to organize raw data into groups besides using a frequency distribution is a **stem-and-leaf plot**. This technique is simple and provides a unique view of the data. A stem-and-leaf plot is constructed by separating the digits for each number of the data into two groups, a stem and a leaf. The leftmost digits are the stem and consist of the higher valued digits. The rightmost digits are the leaves and contain the lower values. If a set of data has only two digits, the stem is the value on the left and the leaf is the value on the right. For example, if 34 is one of the numbers, the stem is 3 and the leaf is 4. For numbers with more than two digits, division of stem and leaf is a matter of researcher preference.

● Dot Plots



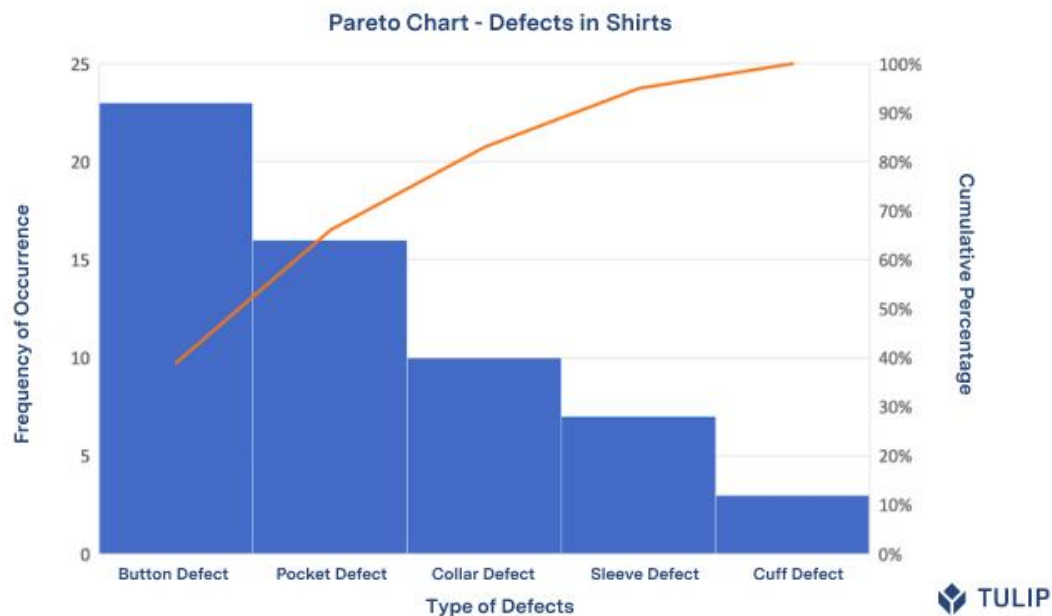
A relatively simple statistical chart that is generally used to display continuous, quantitative data is the **dot plot**. In a dot plot, each data value is plotted along the horizontal axis and is represented on the chart by a dot. If multiple data points have the same values, the dots will stack up vertically. If there are a large number of close points, it may not be possible to display all of the data values along the horizontal axis. Dot plots can be especially useful for observing the overall shape of the distribution of data points along with identifying data values or intervals for which there are groupings and gaps in the data.

● Pareto Charts

A Pareto Chart is a graph that indicates the frequency of defects, as well as their cumulative impact. Pareto Charts are useful to find the defects to prioritize in order to observe the greatest overall improvement.

To expand on this definition, let's break a Pareto Chart into its components.

- 1) A Pareto Chart is a combination of a bar graph and a line graph. Notice the presence of both bars and a line on the Pareto Chart below.
- 2) Each bar usually represents a type of defect or problem. The height of the bar represents any important unit of measure — often the frequency of occurrence or cost.
- 3) The bars are presented in descending order (from tallest to shortest). Therefore, you can see which defects are more frequent at a glance.
- 4) The line represents the cumulative percentage of defects.



Let's look at the table of data for the Pareto Chart above to understand what cumulative percentage is.

TYPE OF DEFECT	FREQUENCY OF DEFECT	% OF TOTAL	CUMULATIVE %
Button Defect	23	39.0	39.0

For Collar the % of simply	Pocket Defect	16	27.1	66.1	Defects, Total is
	Collar Defect	10	16.9	83.1	
	Cuff Defect	7	11.9	11.9	
	Sleeve Defect	3	5.1	16.9	
	Total	59	-	-	

$(10/59)*100$.

The Cumulative % corresponds to the sum of all percentages previous to and including Collar Defects. In this case, this would be the sum of the percentages of Button Defects, Pocket Defects, and Collar Defects (39% + 27.1% + 16.9%).

The last cumulative percentage will always be 100%.

Cumulative percentages indicate what percentage of all defects can be removed if the most important types of defects are solved.

In the example above, solving just the two most important types of defects — Button Defects and Pocket Defects – will remove 66% of all defects.

In any Pareto Chart, for as long as the cumulative percentage line is steep, the types of defects have a significant cumulative effect. Therefore, it is worth finding the cause of these types of defects and solving them. When the cumulative percentage line starts to flatten, the types of defects do not deserve as much attention since solving them will not influence the outcome as much.

5) A Pareto Chart is a quality tool: it helps analyze and prioritize issue resolution.

The idea behind a Pareto Chart is that the few most significant defects make up most of the overall problem. We have already covered two ways the Pareto Charts help find the defects that have the most cumulative effect.

First, the first bars are always the tallest, indicating the most common sources of defects. Second, the cumulative percentage line indicates which defects to prioritize to get the most overall improvement.

❖ Measures of Central Tendency

❖ Mode

The **mode** is the most frequently occurring value in a set of data. For following data the mode is \$19.00 because the offer price that recurred the most times (four) was \$19.00. Organizing the data into an ordered array (an ordering of the numbers from smallest to largest) helps to locate the mode. The following is an ordered array of the values from following data.

7.00 11.00 14.25 15.00 15.00 15.50 19.00 19.00 19.00 19.00
21.00 22.00 23.00 24.00 25.00 27.00 27.00 28.00 34.22 43.25

This grouping makes it easier to see that 19.00 is the most frequently occurring number.

In the case of a tie for the most frequently occurring value, two modes are listed. Then the data are said to be **bimodal**. If a set of data is not exactly bimodal but contains two values that are more dominant than others, some researchers take the liberty of referring to the data set as bimodal even without an exact tie for the mode. Data sets with more than two modes are referred to as **multimodal**.

❖ Median

The **median** is the middle value in an ordered array of numbers. For an array with an odd number of terms, the median is the middle number. For an array with an even number of terms, the median is the average of the two middle numbers. The following steps are used to determine the median.

STEP 1. Arrange the observations in an ordered data array.

STEP 2. For an odd number of terms, find the middle term of the ordered array. It is the median.

STEP 3. For an even number of terms, find the average of the middle two terms. This average is the median.

Suppose a business researcher wants to determine the median for the following numbers.

15 11 14 3 21 17 22 16 19 16 5 7 19 8 9 20 4

The researcher arranges the numbers in an ordered array.

3 4 5 7 8 9 11 14 15 16 16 17 19 19 20 21 22

Because the array contains 17 terms (an odd number of terms), the median is the middle number, or 15.

If the number 22 is eliminated from the list, the array would contain only 16 terms.

3 4 5 7 8 9 11 14 15 16 16 17 19 19 20 21

Now, for an even number of terms, the statistician determines the median by averaging the two middle values, 14 and 15. The resulting median value is 14.5.

Another way to locate the median is by finding the term in an ordered array.

For example, if a data set contains 77 terms, the median is the 39th term. That is,

$$N+1/2 = 77+1/2 = 78/2 = 39^{\text{th}} \text{ term}$$

❖ Mean

The **arithmetic mean** is the average of a group of numbers and is computed by summing all numbers and dividing by the number of numbers. Because the arithmetic mean is so widely used, most statisticians refer to it simply as the mean.

POPULATION MEAN

$$\mu = \frac{\sum x}{N} = \frac{x_1 + x_2 + x_3 + \dots + x_N}{N}$$

SAMPLE MEAN

$$\bar{x} = \frac{\sum x}{n} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

Suppose a company has five departments with 24, 13, 19, 26, and 11 workers each. The population mean number of workers in each department is 18.6 workers. The computations follow.

24
13
19
11
26
93

$$\mu = \frac{\sum x}{N} = \frac{93}{5} = 18.6$$

❖ Percentiles

Percentiles are measures of central tendency that divide a group of data into 100 parts. There are 99 percentiles because it takes 99 dividers to separate a group of data into 100 parts.

Steps in Determining the Location of a Percentile

1. Organize the numbers into an ascending-order array.
2. Calculate the percentile location (i) by:

$$i = \frac{P}{100}(N)$$

where

3. Determine the location by either (a) or (b).
 - a. If i is a whole number, the P th percentile is the average of the value at the ith location and the value at the location.
 - b. If i is not a whole number, the Pth percentile value is located at the whole number part of

❖ Quartiles

Quartiles are measures of central tendency that divide a group of data into four subgroups or parts. The three quartiles are denoted as Q1, Q2, and Q3. The first quartile, Q1, separates the first, or lowest, one-fourth of the data from the upper three-fourths and is equal to the 25th percentile. The second quartile, Q2, separates the second quarter of the data from the third quarter. Q2 is located at the 50th percentile and equals the median of the data. The third quartile, Q3, divides the first three-quarters of the data from the last quarter and is equal to the value of the 75th percentile.

Illustration - 1

Determine the mode for the following numbers.

2 4 8 4 6 2 7 8 4 3 8 9 4 3 5

Mode

2, 2, 3, 3, 4, 4, 4, 4, 5, 6, 7, 8, 8, 8, 9

The mode = **4**

4 is the most frequently occurring value

Illustration - 2

Determine the median for the following numbers.

213 345 609 073 167 243 444 524 199 682

Arrange terms in ascending order:

073, 167, 199, 213, 243, 345, 444, 524, 609, 682

There are 10 terms.

Since there are an even number of terms, the median is the average of the two middle terms:

$$\text{Median} = \frac{243 + 345}{2} = \frac{588}{2} = 294$$

$$\frac{243 + 345}{2} = \frac{588}{2} = 294$$

Using the formula, the median is located at the $\frac{n+1}{2}$ th term.

$$\frac{10+1}{2} = 5.5$$

$$n = 10 \text{ therefore } \frac{10 + 1}{2} = \frac{11}{2} = 5.5^{\text{th}} \text{ term.}$$

The median is located halfway between the 5th and 6th terms.

$$5^{\text{th}} \text{ term} = 243 \quad 6^{\text{th}} \text{ term} = 345$$

Illustration - 3

Compute the 35th percentile, the 55th percentile, Q1, Q2, and Q3 for the following data.

16 28 29 13 17 20 11 34 32 27 25 30 19 18 33

Rearranging the data into ascending order:

11, 13, 16, 17, 18, 19, 20, 25, 27, 28, 29, 30, 32, 33, 34

$$i = \frac{35}{100}(15) = 5.25$$

P₃₅ is located at the 5 + 1 = 6th term

$$P_{35} = \mathbf{19}$$

$$i = \frac{55}{100}(15) = 8.25$$

P₅₅ is located at the 8 + 1 = 9th term

$$P_{55} = \mathbf{27}$$

$$Q_1 = P_{25}$$

$$i = \frac{25}{100}(15) = 3.75$$

$Q_1 = P_{25}$ is located at the $3 + 1 = 4^{\text{th}}$ term

$$Q_1 = 17$$

$Q_2 = \text{Median}$

The median is located at the $\left(\frac{15+1}{2}\right)^{\text{th}} = 8^{\text{th}}$ term

$$Q_2 = 25$$

$Q_3 = P_{75}$

$$i = \frac{75}{100}(15) = 11.25$$

$Q_3 = P_{75}$ is located at the $11 + 1 = 12^{\text{th}}$ term

$$Q_3 = 30$$

Illustration - 4

Compute P_{20} , P_{47} , P_{83} , Q_1 , Q_2 , and Q_3 for the following data.

120 138 97 118 172 144

138 107 94 119 139 145

162 127 112 150 143 80

105 116 142 128 116 171

Rearranging the data in ascending order:

80, 94, 97, 105, 107, 112, 116, 116, 118, 119, 120, 127, 128, 138, 138, 139, 142, 143, 144, 145,
150, 162, 171, 172

$$n = 24$$

$$i = \frac{20}{100}(24) = 4.8$$

P_{20} is located at the $4 + 1 = 5^{\text{th}}$ term

$$P_{20} = \mathbf{107}$$

$$i = \frac{47}{100}(24) = 11.28$$

P_{47} is located at the $11 + 1 = 12^{\text{th}}$ term

$$P_{47} = \mathbf{127}$$

$$i = \frac{83}{100}(24) = 19.92$$

P_{83} is located at the $19 + 1 = 20^{\text{th}}$ term

$$P_{83} = \mathbf{145}$$

$$Q_1 = P_{25}$$

$$i = \frac{25}{100}(24) = 6$$

Q_1 is located at the 6.5^{th} term

$$Q_1 = (112 + 116) / 2 = \mathbf{114}$$

$$Q_2 = \text{Median}$$

The median is located at the:

$$\left(\frac{24+1}{2}\right)^{th} = 12.5^{th} \text{ term}$$

$$Q_2 = (127 + 128) / 2 = \mathbf{127.5}$$

$$Q_3 = P_{75}$$

$$i = \frac{75}{100}(24) = 18$$

Illustration - 5

The following lists the number of fatal accidents by scheduled commercial airlines over a 17-year period according to the Air Transport Association of America. Using these data, compute the mean, median, and mode. What is the value of the third quartile? Determine P11, P35, P58, and P67.

4 4 4 1 4 2 4 3 8 6 4 4 1 4 2 3 3

$$n = 17$$

$$\text{Mean} = \frac{\sum x}{N} = \frac{61}{17} = 3.59$$

The median is located at the $\left(\frac{17+1}{2}\right)^{th} = 9^{th}$ term

Median = **4**

Mode = **4**

$$Q_3 = P_{75} \quad i = \frac{75}{100}(17) = 12.75$$

Q_3 is located at the 13th term

$$Q_3 = 4$$

$$P_{11}: i = \frac{11}{100}(17) = 1.87$$

❖ Measures of Variability

Measures of central tendency yield information about the center or middle part of a data set. However, business researchers can use another group of analytic tools, **measures of variability**, to describe the spread or the dispersion of a set of data. Using measures of variability in conjunction with measures of central tendency makes possible a more complete numerical description of the data.

❖ Range

The **range** is the difference between the largest value of a data set and the smallest value of a set.

The data in Table 3.1 represent the offer prices for the 20 largest U.S. initial public offerings in a recent year. The lowest offer price was \$7.00 and the highest price was \$43.25. The range of the offer prices can be computed as the difference of the highest and lowest values:

$$\text{Range} = \text{Highest} - \text{Lowest} = \$43.25 - \$7.00 = \$36.25$$

❖ Interquartile Range

Another measure of variability is the **interquartile range**. The interquartile range is the range of values between the first and third quartile. Essentially, it is the range of the middle 50% of the data and is determined by computing the value of $Q_3 - Q_1$. The interquartile range is especially useful in situations where data users are more interested in values toward the middle and less interested in extremes.

INTERQUARTILE RANGE

$$Q_3 - Q_1$$

❖ Mean Absolute Deviation

The **mean absolute deviation (MAD)** is the average of the absolute values of the deviations around the mean for a set of numbers.

MEAN ABSOLUTE DEVIATION

$$\text{MAD} = \frac{\sum |x - \mu|}{N}$$

❖ Variance

POPULATION VARIANCE

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

SAMPLE VARIANCE

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

SAMPLE STANDARD
DEVIATION

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

❖ z Scores

A **z score** represents the number of standard deviations a value (x) is above or below the mean of a set of numbers when the data are normally distributed. Using z scores allows translation of a value's raw distance from the mean into units of standard deviations.

z SCORE

$$z = \frac{x - \mu}{\sigma}$$

❖ Coefficient of Variation

The **coefficient of variation** is a statistic that is the ratio of the standard deviation to the mean expressed in percentage and is denoted CV.

COEFFICIENT OF VARIATION

$$CV = \frac{\sigma}{\mu}(100)$$

Illustration - 6

A data set contains the following seven values.

6 2 4 9 1 3 5

- a. Find the range.
- b. Find the mean absolute deviation.
- c. Find the population variance.
- d. Find the population standard deviation.
- e. Find the interquartile range.
- f. Find the z score for each value.

<u>x</u>		<u>x - μ</u>	<u>(x - μ)²</u>
6	6 - 4.2857 =	1.7143	2.9388
2		2.2857	5.2244
4		0.2857	.0816
9		4.7143	22.2246
1		3.2857	10.7958
3		1.2857	1.6530
<u>5</u>		<u>0.7143</u>	<u>.5102</u>
Σx = 30		Σ x - μ = 14.2857	Σ(x - μ) ² = 43.4284

$$\mu = \frac{\Sigma x}{N} = \frac{30}{7} = 4.2857$$

a.) Range = 9 - 1 = **8**

b.) M.A.D. = $\frac{\sum|x - \mu|}{N} = \frac{14.2857}{7} = \mathbf{2.041}$

c.) $\sigma^2 = \frac{\sum(x - \mu)^2}{N} = \frac{43.4284}{7} = \mathbf{6.204}$

d.) $\sigma = \sqrt{\frac{\sum(x - \mu)^2}{N}} = \sqrt{6.204} = \mathbf{2.491}$

e.) 1, 2, 3, 4, 5, 6, 9

$Q_1 = P_{25}$

$$i = \frac{25}{100}(7) = 1.75$$

Q_1 is located at the $1 + 1 = 2^{\text{th}}$ term, $Q_1 = 2$

$Q_3 = P_{75}$:

$$i = \frac{75}{100}(7) = 5.25$$

Q_3 is located at the $5 + 1 = 6^{\text{th}}$ term, $Q_3 = 6$

$IQR = Q_3 - Q_1 = 6 - 2 = \mathbf{4}$

$$f.) \quad z = \frac{6 - 4.2857}{2.491} = \mathbf{0.69}$$

$$z = \frac{2 - 4.2857}{2.491} = \mathbf{-0.92}$$

$$z = \frac{4 - 4.2857}{2.491} = \mathbf{-0.11}$$

$$z = \frac{9 - 4.2857}{2.491} = \mathbf{1.89}$$

$$z = \frac{1 - 4.2857}{2.491} = \mathbf{-1.32}$$

$$z = \frac{3 - 4.2857}{2.491} = \mathbf{-0.52}$$

$$z = \frac{5 - 4.2857}{2.491} = \mathbf{0.29}$$

Illustration - 7

A data set contains the following six values.

12 23 19 26 24 23

- a. Find the population standard deviation using the formula containing the mean (the original formula).

- b. Find the population standard deviation using the computational formula.
- c. Compare the results. Which formula was faster to use? Which formula do you prefer? Why do you think the computational formula is sometimes referred to as the “shortcut” formula?

a.)

<u>x</u>	<u>(x-μ)</u>	<u>(x-μ)²</u>
12	12-21.167= -9.167	84.034
23	1.833	3.360
19	-2.167	4.696
26	4.833	23.358
24	2.833	8.026
<u>23</u>	<u>1.833</u>	<u>3.360</u>
Σx = 127	Σ(x-μ) = -0.002	Σ(x-μ) ² = 126.834

$$\mu = \frac{\Sigma x}{N} = \frac{127}{6} = 21.167$$

$$\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}} = \sqrt{\frac{126.834}{6}} = \sqrt{21.139} = \mathbf{4.598} \quad \underline{\text{ORIGINAL FORMULA}}$$

b.)

<u>x</u>	<u>x²</u>
12	144
23	529
19	361
26	676
24	576
<u>23</u>	<u>529</u>

$$\Sigma x = 127$$

$$\Sigma x^2 = 2815$$

$$\sigma =$$

$$\sqrt{\frac{\Sigma x^2 - \frac{(\Sigma x)^2}{N}}{N}} = \sqrt{\frac{2815 - \frac{(127)^2}{6}}{6}} = \sqrt{\frac{2815 - 2688.17}{6}} = \sqrt{\frac{126.83}{6}} = \sqrt{21.138}$$

$$= 4.598 \quad \text{SHORT-CUT FORMULA}$$

The short-cut formula is faster.

Illustration -8

Determine the interquartile range on the following data.

44 18 39 40 59

46 59 37 15 73

23 19 90 58 35

82 14 38 27 24

71 25 39 84 70

14, 15, 18, 19, 23, 24, 25, 27, 35, 37, 38, 39, 39, 40, 44, 46, 58, 59, 59, 70, 71, 73, 82, 84, 90

$$Q_1 = P_{25}$$

$$i = \frac{25}{100}(25) = 6.25$$

$$Q_1 = 25$$

$$Q_3 = P_{75}$$

$$i = \frac{75}{100}(25) = 18.75$$

P_{75} is located at the $18 + 1 = 19^{\text{th}}$ term

$$Q_3 = 59$$

$$\text{IQR} = Q_3 - Q_1 = 59 - 25 = \mathbf{34}$$

P_{25} is located at the $6 + 1 = 7^{\text{th}}$ term

3.17 According to Chebyshev's theorem, at least what proportion of the data will be within k for each value of k ?

- a. $k = 2$
- b. $k = 2.5$
- c. $k = 1.6$
- d. $k = 3.2$

$$\text{a) } 1 - \frac{1}{2^2} = 1 - \frac{1}{4} = \frac{3}{4} = .75$$

$$\text{b) } 1 - \frac{1}{2.5^2} = 1 - \frac{1}{6.25} = .84$$

$$\text{c) } 1 - \frac{1}{1.6^2} = 1 - \frac{1}{2.56} = .609$$

$$\text{d) } 1 - \frac{1}{3.2^2} = 1 - \frac{1}{10.24} = .902$$

Measures of Central Tendency And Variability : Grouped Data

❖ Mean

With grouped data, the specific values are unknown. What can be used to represent the data values? The midpoint of each class interval is used to represent all the values in a class interval. This midpoint is weighted by the frequency of values in that class interval. The mean for grouped data is then computed by summing the products of the class midpoint and the class frequency for each class and dividing that sum by the total number of frequencies. The formula for the mean of grouped data follows.

MEAN OF GROUPED DATA

$$\mu_{\text{grouped}} = \frac{\sum fM}{N} = \frac{\sum fM}{\sum f} = \frac{f_1M_1 + f_2M_2 + \dots + f_iM_i}{f_1 + f_2 + \dots + f_i}$$

where

i = the number of classes

f = class frequency

N = total frequencies

❖ Median

The median for ungrouped or raw data is the middle value of an ordered array of numbers. For grouped data, solving for the median is considerably more complicated. The calculation of the median for grouped data is done by using the following formula.

MEDIAN OF GROUPED DATA

$$\text{Median} = L + \frac{\frac{N}{2} - cf_p}{f_{med}}(W)$$

where:

- L = the lower limit of the median class interval
- cf_p = a cumulative total of the frequencies up to but not including the frequency of the median class
- f_{med} = the frequency of the median class
- W = the width of the median class interval
- N = total number of frequencies

M

measures of Variability

Two measures of variability for grouped data are presented here: the variance and the standard deviation. Again, the standard deviation is the square root of the variance. Both measures have original and computational formulas.

FORMULAS FOR POPULATION VARIANCE AND STANDARD DEVIATION OF GROUPED DATA

<u>Original Formula</u>	<u>Computational Version</u>
-------------------------	------------------------------

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

where:

- f = frequency
- M = class midpoint
- $N = \sum f$, or total frequencies of the population
- μ = grouped mean for the population

FORMULAS FOR SAMPLE VARIANCE AND STANDARD DEVIATION OF GROUPED DATA

<u>Original Formula</u>	<u>Computational Version</u>
-------------------------	------------------------------

$s^2 = \frac{\sum f(M - \bar{x})^2}{n - 1}$	$s^2 = \frac{\sum fM^2 - \frac{(\sum fM)^2}{n}}{n - 1}$
$s = \sqrt{s^2}$	

where:

- f = frequency
- M = class midpoint
- $n = \sum f$, or total of the frequencies of the sample
- \bar{x} = grouped mean for the sample

Illustration - 9

Compute the mean, the median, and the mode for the following data.

Class	f
0-under 2	39
2-under 4	27
4-under 6	16
6-under 8	15
8-under 10	10
10-under 12	8
12-under 14	6

<u>Class</u>	<u>f</u>	<u>M</u>	<u>fM</u>
0 - 2	39	1	39
2 - 4	27	3	81
4 - 6	16	5	80
6 - 8	15	7	105
8 - 10	10	9	90
10 - 12	8	11	88
12 - 14	<u>6</u>	13	<u>78</u>
	$\Sigma f=121$		$\Sigma fM = 561$

$$\mu = \frac{\Sigma fM}{\Sigma f} = \frac{561}{121} = 4.64$$

Mode: The modal class is 0 – 2.

The midpoint of the modal class = the mode = 1

3.29 Determine the population variance and standard deviation for the following data by using the original formula.

Class	f
20–under 30	7
30–under 40	11
40–under 50	18
50–under 60	13
60–under 70	6
70–under 80	4

<u>Class</u>	<u>f</u>	<u>M</u>	<u>fM</u>
20-30	7	25	175
30-40	11	35	385
40-50	18	45	810
50-60	13	55	715
60-70	6	65	390
70-80	<u>4</u>	75	<u>300</u>
Total	59		2775

$$\mu = \frac{\Sigma fM}{\Sigma f} = \frac{2775}{59} = 47.034$$

$M - \mu$	$(M - \mu)^2$	$f(M - \mu)^2$
-22.0339	485.4927	3398.449
-12.0339	144.8147	1592.962
- 2.0339	4.1367	74.462
7.9661	63.4588	824.964
17.9661	322.7808	1936.685
27.9661	782.1028	3128.411
Total		10,955.933

$$\sigma^2 = \frac{\sum f(M - \mu)^2}{\sum f} = \frac{10,955.93}{59} = \mathbf{185.694}$$

$$\sigma = \sqrt{185.694} = \mathbf{13.627}$$

3.31 A random sample of voters in Nashville, Tennessee, is classified by age group, as shown by the following data.

Age Group	Frequency
18–under 24	17
24–under 30	22
30–under 36	26
36–under 42	35
42–under 48	33
48–under 54	30
54–under 60	32
60–under 66	21
66–under 72	15

- Calculate the mean of the data.
- Calculate the mode.
- Calculate the median.
- Calculate the variance.
- Calculate the standard deviation.

<u>Class</u>	<u>f</u>	<u>M</u>	<u>fM</u>	<u>fM²</u>
18 - 24	17	21	357	7,497
24 - 30	22	27	594	16,038
30 - 36	26	33	858	28,314
36 - 42	35	39	1,365	53,235
42 - 48	33	45	1,485	66,825
48 - 54	30	51	1,530	78,030
54 - 60	32	57	1,824	103,968
60 - 66	21	63	1,323	83,349
66 - 72	<u>15</u>	69	<u>1,035</u>	<u>71,415</u>
	$\Sigma f = 231$		$\Sigma fM = 10,371$	$\Sigma fM^2 = 508,671$

a.) Mean: $\bar{x} = \frac{\Sigma fM}{n} = \frac{\Sigma fM}{\Sigma f} = \frac{10,371}{231} = \mathbf{44.9}$

b.) Mode. The Modal Class = 36-42. The mode is the class midpoint = **39**

c.) $s^2 = \frac{\Sigma fM^2 - \frac{(\Sigma fM)^2}{n}}{n-1} = \frac{508,671 - \frac{(10,371)^2}{231}}{230} = \frac{43,053.5}{230} = \mathbf{187.2}$

$$d.) s = \sqrt{187.2} = 13.7$$

Illustration- 11

The following data represent the number of appointments made per 15-minute interval by telephone solicitation for a lawn-care company. Assume these are population data.

Number of Frequency	Appointments of Occurrence
0–under 1	31
1–under 2	57
2–under 3	26
3–under 4	14
4–under 5	6
5–under 6	3

- a. Calculate the mean of the data.
- b. Calculate the mode.
- c. Calculate the median.
- d. Calculate the variance.
- e. Calculate the standard deviation.

a.) Mean

<u>Class</u>	<u>f</u>	<u>M</u>	<u>fM</u>	<u>fM²</u>
0 - 1	31	0.5	15.5	7.75
1 - 2	57	1.5	85.5	128.25
2 - 3	26	2.5	65.0	162.50
3 - 4	14	3.5	49.0	171.50
4 - 5	6	4.5	27.0	121.50
5 - 6	<u>3</u>	5.5	<u>16.5</u>	<u>90.75</u>
	$\Sigma f=137$		$\Sigma fM=258.5$	$\Sigma fM^2=682.25$

$$\mu = \frac{\Sigma fM}{\Sigma f} = \frac{258.5}{137} = \mathbf{1.89}$$

b.) Mode: Modal Class = 1-2. Mode = **1.5**

c.) Variance:

$$\sigma^2 = \frac{\Sigma fM^2 - \frac{(\Sigma fM)^2}{N}}{N} = \frac{682.25 - \frac{(258.5)^2}{137}}{137} = \mathbf{1.4197}$$

d.) standard Deviation:

$$\sigma = \sqrt{\sigma^2} = \sqrt{1.4197} = \mathbf{1.1915}$$

Measures of shape

Measures of shape are tools that can be used to describe the shape of a distribution of data. In this section, we examine two measures of shape, skewness and kurtosis. We also look at box-and-whisker plots.

❖ Skewness

A distribution of data in which the right half is a mirror image of the left half is said to be symmetrical. One example of a symmetrical distribution is the normal distribution, or bell curve, shown in Figure 3.8 and presented in more detail in Chapter 6.

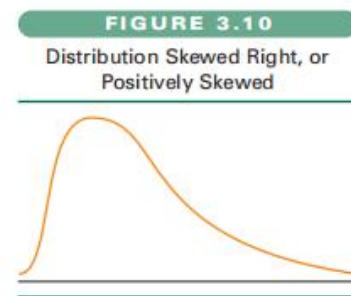
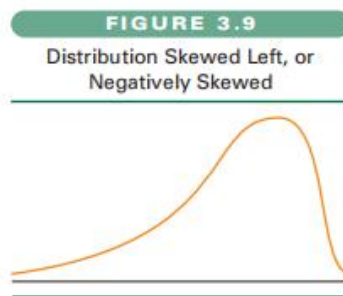
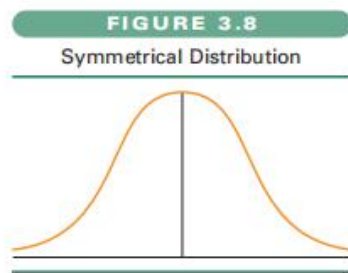
Skewness is when a distribution is asymmetrical or lacks symmetry. The distribution in

Figure 3.8 has no skewness because it is symmetric. Figure 3.9 shows a distribution that is skewed left, or negatively skewed, and Figure 3.10 shows a distribution that is skewed right, or positively skewed.

The skewed portion is the long, thin part of the curve. Many researchers use skewed distribution to denote that the data are sparse at one end of the distribution and piled up at the other end. Instructors sometimes refer to a grade distribution as skewed, meaning that few students scored at one end of the grading scale, and many students scored at the other end.

Skewness and the Relationship of the Mean, Median, and Mode

The concept of skewness helps to understand the relationship of the mean, median, and mode. In a unimodal distribution (distribution with a single peak or mode) that is skewed, the mode is the apex (high point) of the curve and the median is the middle value. The mean tends to be located toward the tail of the distribution, because the mean is particularly affected by the extreme values. A bell-shaped or normal distribution with the mean, median, and mode all at the center of the distribution has no skewness. Figure 3.11 displays the relationship of the mean, median, and mode for different types of skewness.



Coefficient of Skewness

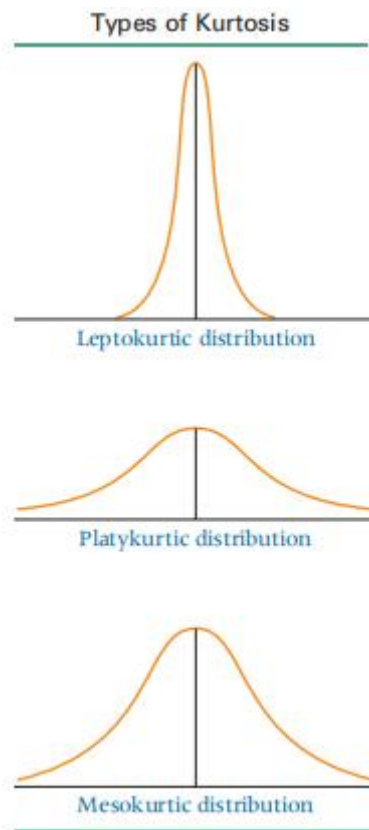
Statistician Karl Pearson is credited with developing at least two coefficients of skewness that can be used to determine the degree of skewness in a distribution. We present one of these coefficients here, referred to as a Pearsonian **coefficient of skewness**. This coefficient compares the mean and median in light of the magnitude of the standard deviation.

Note that if the distribution is symmetrical, the mean and median are the same value and hence the coefficient of skewness is equal to zero.

<p>COEFFICIENT OF SKEWNESS</p> <p style="text-align: center;">where</p> <p>S_k = coefficient of skewness M_d = median</p>	$S_k = \frac{3(\mu - M_d)}{\sigma}$
--	-------------------------------------

❖ Kurtosis

Kurtosis describes the amount of peakedness of a distribution. Distributions that are high and thin are referred to as **leptokurtic** distributions. Distributions that are flat and spread out are referred to as **platykurtic** distributions. Between these two types are distributions that are more “normal” in shape, referred to as **mesokurtic** distributions. These three types of kurtosis are.



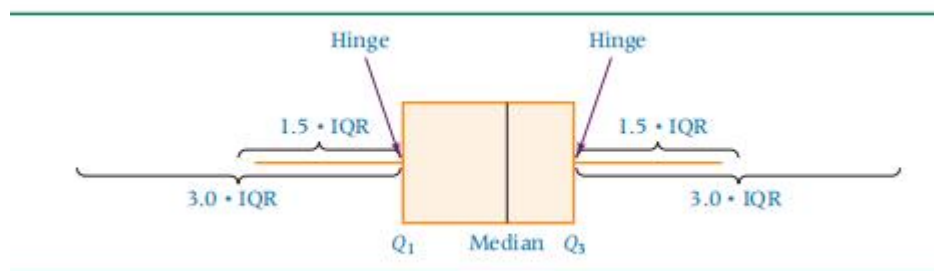
❖ Box-and-Whisker Plots

Another way to describe a distribution of data is by using a box and whisker plot. A **box-and-whisker plot**, sometimes called a box plot, is a diagram that utilizes the upper and lower quartiles along with the median and the two most extreme values to depict a distribution graphically. The plot is constructed by using a box to enclose the median. This box is extended outward from the median along a continuum to the lower and upper quartiles, enclosing not only the median but also the middle 50% of the data. From the lower and upper quartiles, lines

referred to as whiskers are extended out from the box toward the outermost data values. The box-and-whisker plot is determined from five specific numbers.

1. The median (Q₂)
2. The lower quartile (Q₁)
3. The upper quartile (Q₃)
4. The smallest value in the distribution
5. The largest value in the distribution

The box of the plot is determined by locating the median and the lower and upper quartiles on a continuum. A box is drawn around the median with the lower and upper quartiles (Q₁ and Q₃) as the box endpoints. These box endpoints (Q₁ and Q₃) are referred to as the hinges of the box.



Next the value of the interquartile range (IQR) is computed by $Q_3 - Q_1$. The interquartile range includes the middle 50% of the data and should equal the length of the box.

However, here the interquartile range is used outside of the box also. At a distance of 1.5 * IQR outward from the lower and upper quartiles are what are referred to as inner fences. A whisker, a line segment, is drawn from the lower hinge of the box outward to the smallest data value. A second whisker is drawn from the upper hinge of the box outward to the largest data value. The inner fences are established as follows.

$$Q_1 - 1.5 \cdot IQR$$

$$Q_3 + 1.5 \cdot IQR$$

If data fall beyond the inner fences, then outer fences can be constructed:

$$Q1 - 3.0 \cdot IQR$$

$$Q3 + 3.0 \cdot IQR$$

Probability

Classical Method of Assigning Probabilities

When probabilities are assigned based on laws and rules, the method is referred to as the **classical method of assigning probabilities**. This method involves an experiment, which is a process that produces outcomes, and an event, which is an outcome of an experiment.

When we assign probabilities using the classical method, the probability of an individual event occurring is determined as the ratio of the number of items in a population containing the event (n_e) to the total number of items in the population (N). That is, $P(E) = \frac{n_e}{N}$. For example, if a company has 200 workers and 70 are female, the probability of randomly selecting a female from this company is $\frac{70}{200} = .35$.

CLASSICAL METHOD OF ASSIGNING PROBABILITIES

where

N = total possible number of outcomes of an experiment

n_e = the number of outcomes in which the event occurs out of N outcomes

$$P(E) = \frac{n_e}{N}$$

❖ Relative Frequency of Occurrence

The **relative frequency of occurrence method** of assigning probabilities is based on cumulated historical data. With this method, the probability of an event occurring is equal to the number of times the event has occurred in the past divided by the total number of opportunities for the event to have occurred.

PROBABILITY BY RELATIVE FREQUENCY OF OCCURRENCE

$$\frac{\text{Number of Times an Event Occurred}}{\text{Total Number of Opportunities for the Event to Occur}}$$

❖ Subjective Probability

The **subjective method** of assigning probability is based on the feelings or insights of the person determining the probability. Subjective probability comes from the person's intuition or reasoning. Although not a scientific approach to probability, the subjective method often is

based on the accumulation of knowledge, understanding, and experience stored and processed in the human mind. At times it is merely a guess. At other times, subjective probability can potentially yield accurate probabilities. Subjective probability can be used to capitalize on the background of experienced workers and managers in decision making.

Structure of Probability

❖ Experiment

As previously stated, an **experiment** is a process that produces outcomes. Examples of business oriented experiments with outcomes that can be statistically analyzed might include the following.

- Interviewing 20 randomly selected consumers and asking them which brand of appliance they prefer
- Sampling every 200th bottle of ketchup from an assembly line and weighing the contents
- Testing new pharmaceutical drugs on samples of cancer patients and measuring the patients' improvement
- Auditing every 10th account to detect any errors
- Recording the Dow Jones Industrial Average on the first Monday of every month for 10 years

❖ Event

Because an **event** is an outcome of an experiment, the experiment defines the possibilities of the event. If the experiment is to sample five bottles coming off a production line, an event could be to get one defective and four good bottles. In an experiment to roll a die, one event could be to roll an even number and another event could be to roll a number greater than two. Events are denoted by uppercase letters; italic capital letters (e.g., *A* and *E*₁, *E*₂, . . .) represent the general or abstract case, and roman capital letters (e.g., *H* and *T* for heads and tails) denote specific things and people.

❖ Elementary Events

Events that cannot be decomposed or broken down into other events are called **elementary events**. Elementary events are denoted by lowercase letters (e.g., e_1, e_2, e_3, \dots). Suppose the experiment is to roll a die. The elementary events for this experiment are to roll a 1 or roll a 2 or roll a 3, and so on. Rolling an even number is an event, but it is not an elementary event because the even number can be broken down further into events 2, 4, and 6.

❖ Sample Space

A **sample space** is a complete roster or listing of all elementary events for an experiment. Table 4.1 is the sample space for the roll of a pair of dice. The sample space for the roll of a single die is $\{1, 2, 3, 4, 5, 6\}$.

TABLE 4.1

All Possible Elementary
Events in the Roll of a Pair
of Dice (Sample Space)

(1,1)	(2,1)	(3,1)	(4,1)	(5,1)	(6,1)
(1,2)	(2,2)	(3,2)	(4,2)	(5,2)	(6,2)
(1,3)	(2,3)	(3,3)	(4,3)	(5,3)	(6,3)
(1,4)	(2,4)	(3,4)	(4,4)	(5,4)	(6,4)
(1,5)	(2,5)	(3,5)	(4,5)	(5,5)	(6,5)
(1,6)	(2,6)	(3,6)	(4,6)	(5,6)	(6,6)

❖ Mutually Exclusive Events

Two or more events are **mutually exclusive events** if the occurrence of one event precludes the occurrence of the other event(s). This characteristic means that mutually exclusive events cannot occur simultaneously and therefore can have no intersection.

A manufactured part is either defective or okay: The part cannot be both okay and defective at the same time because “okay” and “defective” are mutually exclusive categories. In a sample of the manufactured products, the event of selecting a defective part is mutually exclusive with the event of selecting a nondefective part. Suppose an office building is for sale and two different potential buyers have placed bids on the building. It is not possible for both buyers to purchase the building; therefore, the event of buyer A purchasing the building is mutually exclusive with the event of buyer B purchasing the building. In the toss of a single coin, heads and tails are mutually exclusive events. The person tossing the coin gets either a head or a tail but never both.

MUTUALLY EXCLUSIVE
EVENTS X AND Y

$$P(X \cap Y) = 0$$

❖ Independent Events

Two or more events are **independent events** if the occurrence or nonoccurrence of one of the events does not affect the occurrence or nonoccurrence of the other event(s). Certain experiments, such as rolling dice, yield independent events; each die is independent of the other. Whether a 6 is rolled on the first die has no influence on whether a 6 is rolled on the second die. Coin tosses always are independent of each other. The event of getting a head on the first toss of a coin is independent of getting a head on the second toss. It is generally believed that certain human characteristics are independent of other events. For example, left-handedness is probably independent of the possession of a credit card. Whether a person wears glasses or not is probably independent of the brand of milk preferred.

INDEPENDENT EVENTS
X AND Y

$$P(X|Y) = P(X) \quad \text{and} \quad P(Y|X) = P(Y)$$

❖ Collectively Exhaustive Events

A list of **collectively exhaustive events** contains all possible elementary events for an experiment. Thus, all sample spaces are collectively exhaustive lists. The list of possible outcomes for the tossing of a pair of dice contained in Table 4.1 is a collectively exhaustive list. The sample space for an experiment can be described as a list of events that are mutually exclusive and collectively exhaustive. Sample space events do not overlap or intersect, and the list is complete.

❖ Complementary Events

The **complement** of event A is denoted A^c , pronounced “not A.” All the elementary events of an experiment not in A comprise its complement. For example, if in rolling one die, event A is getting an even number, the complement of A is getting an odd number. If event A is getting a 5 on the roll of a die, the complement of A is getting a 1, 2, 3, 4, or 6. The complement of event A contains whatever portion of the sample space that event A does not contain, as the Venn diagram in Figure 4.5 shows.

PROBABILITY OF THE
COMPLEMENT OF A

$$P(A') = 1 - P(A)$$

Illustration- 12

A supplier shipped a lot of six parts to a company. The lot contained three defective parts. Suppose the customer decided to randomly select two parts and test them for defects. How large a sample space is the customer potentially working with? List the sample space. Using the sample space list, determine the probability that the customer will select a sample with exactly one defect.

Enumeration of the six parts: $D_1, D_2, D_3, A_4, A_5, A_6$

D = Defective part

A = Acceptable part

Sample Space

$D_1 D_2, D_2 D_3, D_3 A_4$

$D_1 D_3, D_2 A_4, D_3 A_6$

$D_1 A_4, D_2 A_5, A_4 A_5$

$D_1 A_5, D_2 A_6, A_4 A_6$

$D_1 A_6, D_3 A_4, A_5 A_6$

There are 15 members of the sample space

The probability of selecting exactly one defect out of two is:

$$9/15 = .60$$

Illustration- 13

If a population consists of the positive even numbers through 30 and if $A = \{2, 6, 12, 24\}$, what is A' ?

If $A = \{2, 6, 12, 24\}$ and the population is the positive even numbers through 30,

$$A' = \{4, 8, 10, 14, 16, 18, 20, 22, 26, 28, 30\}$$

Illustration- 14

A company's customer service 800 telephone system is set up so that the caller has six options. Each of these six options leads to a menu with four options. For each of these four options, three more options are available. For each of these three options, another three options are presented. If a person calls the 800 number for assistance, how many total options are possible?

$$6(4)(3)(3) = 216$$

Illustration- 15

A bin contains six parts. Two of the parts are defective and four are acceptable. If three of the six parts are selected from the bin, how large is the sample space? Which counting rule did you use, and why? For this sample space, what is the probability that exactly one of the three sampled parts is defective?

Enumeration of the six parts: $D_1, D_2, A_1, A_2, A_3, A_4$

D = Defective part

A = Acceptable part

Sample Space:

$D_1 D_2 A_1, D_1 D_2 A_2, D_1 D_2 A_3,$

$D_1 D_2 A_4, D_1 A_1 A_2, D_1 A_1 A_3,$

$D_1 A_1 A_4, D_1 A_2 A_3, D_1 A_2 A_4,$



$D_1 A_3 A_4, D_2 A_1 A_2, D_2 A_1 A_3,$

$D_2 A_1 A_4, D_2 A_2 A_3, D_2 A_2 A_4,$

$D_2 A_3 A_4, A_1 A_2 A_3, A_1 A_2 A_4,$

$A_1 A_3 A_4, A_2 A_3 A_4$

Combinations are used to counting the sample space because sampling is done without replacement.

$${}^6C_3 = \frac{6!}{3!3!} = 20$$

Probability that one of three is defective is:

$$12/20 = 3/5 \quad .60$$

There are 20 members of the sample space and 12 of them have 1 defective part.

Illustration- 16

A company places a seven-digit serial number on each part that is made. Each digit of the serial number can be any number from 0 through 9. Digits can be repeated in the serial number. How many different serial numbers are possible?

$$10^7 = 10,000,000 \text{ different numbers}$$

Illustration- 17

A small company has 20 employees. Six of these employees will be selected randomly to be interviewed as part of an employee satisfaction program. How many different groups of six can be selected?

$${}_{20}C_6 = \frac{20!}{6!14!} = 38,760$$

It is assumed here that 6 different (without replacement) employees are to be selected.

❖ Marginal , Union, Joint and Conditional Probabilities

Four particular types of probability are presented in this chapter. The first type is **marginal probability**. Marginal probability is denoted $P(E)$, where E is some event. A marginal probability is usually computed by dividing some subtotal by the whole. An example of marginal probability is the probability that a person owns a Ford car. This probability is computed by dividing the number of Ford owners by the total number of car owners. The probability of a person wearing glasses is also a marginal probability. This probability is computed by dividing the number of people wearing glasses by the total number of people. A second type of probability is the union of two events. Union probability is denoted $P(E_1 \cup E_2)$, where E_1 and E_2 are two events. $P(E_1 \cup E_2)$ is the probability that E_1 will occur or that E_2 will occur or that both E_1 and E_2 will occur. An example of union probability is the probability that a person owns a Ford or a Chevrolet. To qualify for the union, the person only has to have at least one of these cars. Another example is the probability of a person wearing glasses or having red hair. All people wearing glasses are included in the union, along with all redheads and all redheads who wear glasses. In a company, the probability that a person is male or a clerical worker is a union probability. A person qualifies for the union by being male or by being a clerical worker or by being both (a male clerical worker). A third type of probability is the intersection of two events, or joint probability. The joint probability of events E_1 and E_2 occurring is denoted $P(E_1 \cap E_2)$. Sometimes $P(E_1 \cap E_2)$ is read as the probability of E_1 and E_2 . To qualify for the intersection, both events must occur. An example of joint probability is the probability of a person owning both a Ford and a Chevrolet. Owning one type of car is not sufficient. A second example of joint probability is the probability that a person is a redhead and wears glasses.

The fourth type is conditional probability. Conditional probability is denoted $P(E_1|E_2)$. This expression is read: the probability that E_1 will occur given that E_2 is known to have occurred. Conditional probabilities involve knowledge of some prior information. The

information that is known or given is written to the right of the vertical line in the probability statement. An example of conditional probability is the probability that a person owns a Chevrolet given that she owns a Ford. This conditional probability is only a measure of the proportion of Ford owners who have a Chevrolet—not the proportion of total car owners who

own a Chevrolet. Conditional probabilities are computed by determining the number of items that have an outcome out of some subtotal of the population. In the car owner example, the possibilities are reduced to Ford owners, and then the number of Chevrolet owners out of those Ford owners is determined. Another example of a conditional probability is the probability that a worker in a company is a professional given that he is male. Of the four probability types, only conditional probability does not have the population total as its denominator. Conditional probabilities have a population subtotal in the denominator. Figure 4.6 summarizes these four types of probability.

Illustration- 18

Suppose that 47% of all Americans have flown in an airplane at least once and that 28% of all Americans have ridden on a train at least once. What is the probability

that a randomly selected American has either ridden on a train or flown in an airplane? Can this problem be solved? Under what conditions can it be solved? If the problem cannot be solved, what information is needed to make it solvable?

A = event - flown in an airplane at least once

T = event - ridden in a train at least once

$$P(A) = .47 \qquad P(T) = .28$$

P (ridden either a train or an airplane) =

$$P(A \cup T) = P(A) + P(T) - P(A \cap T) = .47 + .28 - P(A \cap T)$$

Cannot solve this problem without knowing the probability of the intersection.

We need to know the probability of the intersection of A and T, the proportion who have ridden both.

Illustration- 19

According to the U.S. Bureau of Labor Statistics, 75% of the women 25 through 49 years of age participate in the labor force. Suppose 78% of the women in that age group are married. Suppose also that 61% of all women 25 through 49 years of age are married and are participating in the labor force.

- What is the probability that a randomly selected woman in that age group is married or is participating in the labor force?
- What is the probability that a randomly selected woman in that age group is married or is participating in the labor force but not both?
- What is the probability that a randomly selected woman in that age group is neither married nor participating in the labor force?

$$P(L) = .75 \quad P(M) = .78 \quad P(M \cap L) = .61$$

- $P(M \cup L) = P(M) + P(L) - P(M \cap L) = .78 + .75 - .61 = .92$
- $P(M \cup L) \text{ but not both} = P(M \cup L) - P(M \cap L) = .92 - .61 = .31$
- $P(NM \cap NL) = 1 - P(M \cup L) = 1 - .92 = .08$

Illustration

Illustration- 20

According to Nielsen Media Research, approximately 67% of all U.S. households with television have cable TV. Seventy-four percent of all U.S. households with television have two or more TV sets. Suppose 55% of all U.S. households with television have cable TV and two or more TV sets. A U.S. household with television is randomly selected.

- What is the probability that the household has cable TV or two or more TV sets?
- What is the probability that the household has cable TV or two or more TV sets but not both?

c. What is the probability that the household has neither cable TV nor two or more TV sets?

d. Why does the special law of addition not apply to this problem?

Let C = have cable TV

Let T = have 2 or more TV sets

$$P(C) = .67, P(T) = .74, P(C \cap T) = .55$$

a) $P(C \cup T) = P(C) + P(T) - P(C \cap T) = .67 + .74 - .55 = .86$

b) $P(C \cap T \text{ but not both}) = P(C \cap T) - P(C \cap T) = .86 - .55 = .31$

c) $P(\overline{C} \cap \overline{T}) = 1 - P(C \cup T) = 1 - .86 = .14$

d) The special law of addition does not apply because $P(C \cap T)$ is not .0000. Possession of cable TV and 2 or more TV sets are not mutually exclusive.

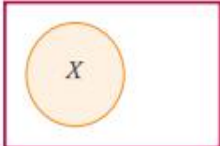
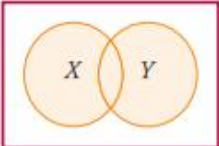
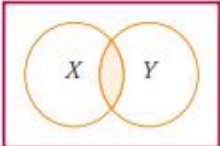
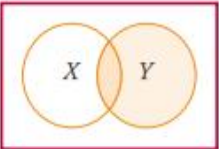
Marginal	Union	Joint	Conditional
$P(X)$	$P(X \cup Y)$	$P(X \cap Y)$	$P(X Y)$
The probability of X occurring	The probability of X or Y occurring	The probability of X and Y occurring	The probability of X occurring given that Y has occurred
Uses total possible outcomes in denominator	Uses total possible outcomes in denominator	Uses total possible outcomes in denominator	Uses subtotal of the possible outcomes in denominator
			

Illustration- 21

A study by Peter D. Hart Research Associates for the Nasdaq Stock Market revealed that 43% of all American adults are stockholders. In addition, the study determined that 75% of all American adult stockholders have some college education. Suppose 37% of all American adults have some college education. An American adult is randomly selected.

- What is the probability that the adult does not own stock?
- What is the probability that the adult owns stock and has some college education?
- What is the probability that the adult owns stock or has some college education?
- What is the probability that the adult has neither some college education nor owns stock?
- What is the probability that the adult does not own stock or has no college education?
- What is the probability that the adult has some college education and owns no stock?

Let S = stockholder

Let C = college

$$P(S) = .43 \quad P(C) = .37 \quad P(C|S) = .75$$

$$a) P(NS) = 1 - .43 = \mathbf{.57}$$

$$b) P(S \Rightarrow C) = P(S) \cdot P(C|S) = (.43)(.75) = \mathbf{.3225}$$

$$c) P(S \Leftarrow C) = P(S) + P(C) - P(S \Rightarrow C) = .43 + .37 - .3225 = \mathbf{.4775}$$

$$d) P(NS \Rightarrow NC) = 1 - P(S \Leftarrow C) = 1 - .4775 = \mathbf{.5225}$$

$$e) P(NS \Leftarrow NC) = P(NS) + P(NC) - P(NS \Rightarrow NC) = .57 + .63 - .5225 = \mathbf{.6775}$$

$$f) P(C \Rightarrow NS) = P(C) - P(C \Rightarrow S) = .37 - .3225 = \mathbf{.0475}$$

Illustration- 22

The U.S. Energy Department states that 60% of all U.S. households have ceiling fans. In addition, 29% of all U.S. households have an outdoor grill. Suppose 13% of all U.S. households have both a ceiling fan and an outdoor grill. A U.S. household is randomly selected.

- What is the probability that the household has a ceiling fan or an outdoor grill?
- What is the probability that the household has neither a ceiling fan nor an outdoor grill?
- What is the probability that the household does not have a ceiling fan and does have an outdoor grill?
- What is the probability that the household does have a ceiling fan and does not have an outdoor grill?

Let C = ceiling fans

Let O = outdoor grill

$$P(C) = .60 \quad P(O) = .29 \quad P(C \cap O) = .13$$

$$a) P(C \cup O) = P(C) + P(O) - P(C \cap O) = .60 + .29 - .13 = .76$$

$$b) P(\overline{C} \cap \overline{O}) = 1 - P(C \cup O) = 1 - .76 = .24$$

$$c) P(\overline{C} \cap O) = P(O) - P(C \cap O) = .29 - .13 = .16$$

$$d) P(C \cap \overline{O}) = P(C) - P(C \cap O) = .60 - .13 = .47$$

No.	Questions	Ans.
1	Virtually every area of business uses statistics in which?	decision making
2	Who give comprehensive definition of statistics as a science dealing with the collection, analysis, interpretation, and presentation of numerical data?	Webster's
3	One of the main ways is to subdivide statistics into how many branches?	Two
4	Name of this subdivide statistics branches:	descriptive statistics and inferential statistics
5	When researchers gather data from the whole population for a given measurement of interest, they called?	Census
6	Every how many years, the government attempts to measure all persons living in this country?	10 years
7	Which is a portion of the whole and, if properly taken, is representative of the whole?	Sample
8	If a business analyst is using data gathered on a group to describe or reach conclusions about that same group, the statistics are called?	descriptive statistics
9	Inferential statistics are sometimes referred to as	inductive statistics
10	Market researchers use inferential statistics to study the ...	impact of advertising on market segment
11	A descriptive measure of the population is called a?	Parameter
12	Parameters are usually denoted by....?	Greek letters
13	The appropriateness of the data analysis depends on the level of ...	measurement of the data gathered



14	How many type of Data measurement techniques are ?	Four
15	Four common levels of data measurement are?	1. Nominal 2. Ordinal 3. Interval 4. Ratio
16	The lowest level of data measurement is the ?	nominal level
17	Employee identification numbers are an example of which data?	nominal data
18	which data measurement is higher than the nominal level?	Ordinal-level data
19	Some questionnaire which type scales are considered by many researchers to be ordinal in level.	Likert-type scales
20	Which data measurement is the highest level of data in which the distances between consecutive numbers have meaning and the data are always numerical?	Interval-level data
21	interval data have equal?	Intervals
22is just another point on the scale and does not mean the absence of the phenomenon.	Zero
23	Ratio data have the same properties as interval data, but ratio data have an absolute zero, and the ratio of	two numbers is meaningful
24	height, weight, time, volume, and Kelvin temperature is the example of ?	ratio data



25	metric data and are sometimes referred to as ?	quantitative data
26	Statistical techniques can be separated into how many categories:	Two
27	Name of Statistical techniques are...	parametric statistics and nonparametric statistics
28	Which statistics require that data be interval or ratio?	Parametric
29	If the data are nominal or ordinal which statistics must be used?	Non parametric
30	raw data, or data that have not been summarized in any way, are sometimes referred to as ?	ungrouped data
31	Data that have been organized into a frequency distribution are called ?	grouped data
32	one particularly useful tool for grouping data is the?	frequency distribution
33	Which often is defined as the difference between the largest and smallest numbers	range
34	Range is difference between the?	Largest – smallest value
35	The midpoint of each class interval is called ?	the class midpoint
36	the class midpoint sometimes referred to as?	class mark
37	It is the value halfway across the class interval and can be calculated as theof the two class endpoints.	Average
38	Which frequency is the proportion of the total frequency that is in any given class interval in a	Relative frequency



	frequency distribution?	
39	Relative frequency is the individual class frequency divided by the ...?	total frequency
40	Which frequency is a running total of frequencies through the classes of a frequency distribution?	cumulative frequency
41	Range / Number of Classes equal to	Class Width
42	One of the most effective mechanisms for presenting data in a form meaningful to decision makers is?	graphical depiction
43	How many types of quantitative data graphs?	five
44is a series of contiguous bars or rectangles that represent the frequency of data in given class intervals.	Histogram
45	A histogram is a useful tool for differentiating the frequencies of...?	class intervals
46	in a frequencyeach class frequency is plotted as a dot at the class midpoint, and the dots are connected by a series of line segments.	Polygon
47is a cumulative frequency polygon	ogive (o-jive)
48	Ogives are most useful when the decision maker wants to see ?	running totals
49	A relatively simple statistical chart that is generally used to display continuous, quantitative data is the?	dot plot
50	dot plot, each data value is plotted along the	a dot



	horizontal axis and is represented on the chart by ?	
51 is constructed by separating the digits for each number of the data into two groups, a stem and a leaf	stem-and-leaf plot
52	The leftmost digits are the stem and consist of thevalued digits.	Higher
53	The rightmost digits are the leaves and contain thevalues	Lower
54	we will examine how many types of qualitative data graphs:	three
55	Name of qualitative data graphs...	(1) pie charts, (2) bar charts, and (3) Pareto charts
56	A pie chart is a circular depiction of data where the area of the whole pie representsof the data	100%
57of the pie represent a percentage breakdown of the sublevels	Slices
58	Pie charts show the relative magnitudes of the parts to ...	the whole
59	A bar graph or chart contains how many or more categories?	Two
60	In Excel, horizontal bar graphs are referred to as...,	bar charts
61	vertical bar graphs are referred to as	column chart
62	The bar chart is called as....	Pareto chart

63	Pareto charts were named after an Italian economist?	Vilfredo Pareto
64is a two-dimensional graph plot of pairs of points from two numerical variables	A scatter plot
65	Measures ofyield information about the center, or middle part, of a group of numbers.	central tendency
66is the most frequently occurring value in a set of data.	Mode
67	In the case of a tie for the most frequently occurring value, two modes are listed. Then the data are said to be?	Bimodal
68	Data sets with more than two modes are referred to as?	Multimodal
69is the middle value in an ordered array of numbers	Median
70	For an array with an odd number of terms, the median is the?	middle number
71	For an array with an even number of terms, the median is the ?	average of the two middle numbers.
72is the average of a group of numbers and is computed by summing all numbers and dividing by the number of numbers.	arithmetic mean
73	The population mean is represented by the Greek letter	Mu (μ)



74	sample mean is represented by	xbar
75	The capital Greek letteris commonly used in mathematics to represent a summation of all the numbers in a grouping	sigma Σ
76are measures of central tendency that divide a group of data into 100 parts	Percentiles
77	Percentiles are widely used in ...	reporting test results
78	Quartiles are measures of central tendency that divide a group of data into how many subgroups or parts?	Four
79	The three quartiles are denoted as	Q1, Q2, and Q3
80	The value of Q1 is found at the...	25th percentile, P25
81	The value of Q2 is equal to the....	Median
82	The value of Q3 is determined by ...	P75
83to describe the spread or the dispersion of a set of data.	measures of variability
84range is the range of values between the first and third quartile	Interquartile
85	The mean absolute deviation (MAD) is theof the absolute values of the deviations around the mean for a set of numbers	Average

86is the average of the squared deviations about the arithmetic mean for a set of numbers	Variance
87	The sum of the squared deviations about the mean of a set of values—called ?	the sum of squares of x
88	Two ways of applying the standard deviation are the	empirical rule and Chebyshev's theorem
89	Which is rule is an important rule of thumb?	empirical rule
90	The empirical rule applies only when data are known to be approximately	normally distributed
91	when the shape of the distribution is unknown ..	Chebyshev's theorem applies
92	Chebyshev's theorem states that at least	$1 - 1 / K^2$
93represents the number of standard deviations a value (x) is above or below the mean of a set of numbers when the data are normally distributed	z score
94	The coefficient of variation is a statistic that is the ratio of the ...	standard deviation to the mean
95	CV means	coefficient of variation
96	Three measures of central tendency are presented here for grouped data:	the mean, the median, and the mode.
97for ungrouped or raw data is the middle value of an ordered array of numbers	Median



98	Two measures of variability for grouped data are presented here:	variance and the standard deviation.
99	Measures ofare tools th that can be used to describe the shape of a distribution of data	Shape
100	How many types of shape	Shape
101	Name of them:	skewness and kurtosis
102is when a distribution is asymmetrical or lacks symmetry	Skewness
103	Statisticianis credited with developing at least two coefficients of skewness	Karl Pearson
104	We present one of these coefficients here, referred to as a Pearsonian ...	coefficient of skewness
105describes the amount of peakedness of a distribution.	Kurtosis
106	Distributions that are high and thin are referred to as which distributions?	Leptokurtic
107	Distributions that are flat and spread out are referred to as which distributions?	Platykurtic
108	Between these two types are distributions that are more “normal” in shape, referred to as which	Mesokurtic



	distributions?	
109	A box-and-whisker plot, sometimes called a ...	box plot
110	The relative frequency of occurrence method of assigning probabilities is based on ...	cumulated historical data.
111	The method of assigning probability is based on the feelings or insights of the person determining the probability.	Subjective
112 is a process that produces outcomes	Experiment
113	event is an outcome of an ?	Experiment
114	Events that cannot be decomposed or broken down into other events are called ?	elementary events
115is a complete roster or listing of all elementary events for an experiment	sample space
116	Two or more events areif the occurrence of one event precludes the occurrence of the other event(s)	mutually exclusive events
117	mutually exclusive events cannot occur simultaneously and therefore can have no ?	Intersection
118	Two or more events areif the occurrence or nonoccurrence of one of the events does not affect the occurrence or nonoccurrence of the other event	independent events
119events contains all possible elementary events	collectively exhaustive



	for an experiment.	
120	All the elementary events of an experiment not in A comprise its ?	Complement
121	The complement of event A is denoted A , pronounced ..	not A.
122	A....probability is usually computed by dividing some subtotal by the whole.	Marginal
123displays the marginal probabilities and the intersection probabilities of a given problem	probability matrix
124	The general law of multiplication is used to find ?	the joint probability
125	special law of multiplication can be used to find the ?	intersection of X and Y
126	If X, Y are two events, the X occurring given that Y is known	conditional probability
127	Bayes' rule, which was developed by ?	Thomas Bayes (1702–1761)
128	For a symmetrical distribution:	$\beta_1 = 0$
129	The scatter in a series of values about the average is called:	Dispersion
130	The measures of dispersion can never be:	Negative



131	Which of the following is an absolute measure of dispersion?	Standard deviation
132	If the observations of a variable X are, -4, -20, -30, -44 and -36, then the value of the range will be:	40
133	If the maximum value in a series is 25 and its range is 15, the maximum value of the series is:	10
134	Mean deviation computed from a set of data is always:	Less than standard deviation
135	Which measure of dispersion has a different unit other than the unit of measurement of values:	Variance
136	The positive square root of the mean of the squares of the deviations of observations from their mean is called:	Standard deviation
137	$S.D(X) = 6$ and $S.D(Y) = 8$. If X and Y are independent random variables, then $S.D(X-Y)$ is:	10
138	The ratio of the standard deviation to the arithmetic mean expressed as a percentage is called:	Coefficient of variation
139	To compare the variation of two or more than two series, we use	Coefficient of variation
140	If standard deviation of the values 2, 4, 6, 8 is 2.236, then standard deviation of the values 4, 8, 12, 16 is:	4.472



141	The moments about mean are called:	Central moments
142	Moment ratios β_1 and β_2 are:	Unit less quantities
143	If the third moment about mean is zero, then the distribution is:	Symmetrical
144	If mean=25, median=30 and standard deviation=15, the distribution will be:	Negatively skewed
145	The degree of peaked ness or flatness of a unimodel distribution is called:	Kurtosis
146	The range of the scores 29, 3, 143, 27, 99 is:	140
147	The sum of absolute deviations is minimum if these deviations are taken from the:	Median
148	For a positively skewed distribution, mean is always:	Greater than the mode
149	The range of the values -5, -8, -10, 0, 6, 10 is:	20
150	Which of the following measures of dispersion is	Coefficient of variation

	independent of the units employed?	
151	The lack of uniformity or symmetry is called:	Skewness

Module - 2

Probability Distribution

Discrete Distribution

A **random variable** is a variable that contains the outcomes of a chance experiment. For example, suppose an experiment is to measure the arrivals of automobiles at a turnpike tollbooth during a 30-second period. The possible outcomes are: 0 cars, 1 car, 2 cars, ..., n cars. These numbers (0, 1, 2, ..., n) are the values of a random variable. Suppose another experiment is to measure the time between the completion of two tasks in a production line. The values will range from 0 seconds to n seconds. These time measurements are the values of another random variable.

The two categories of random variables are

(1) discrete random variables and (2) continuous random variables.

A random variable is a **discrete random variable** if the set of all possible values is at most a finite or a countably infinite number of possible values. In most statistical situations, discrete random variables produce values that are nonnegative whole numbers. For example, if six people are randomly selected from a population and how many of the six are left-handed is to be determined, the random variable produced is discrete. The only possible numbers of left-handed people in the sample of six are 0, 1, 2, 3, 4, 5, and 6. There cannot be 2.75 left handed people in a group of six people; obtaining nonwhole number values is impossible.

Other examples of experiments that yield discrete random variables include the following:

1. Randomly selecting 25 people who consume soft drinks and determining how many people prefer diet soft drinks
2. Determining the number of defects in a batch of 50 items

3. Counting the number of people who arrive at a store during a five-minute period
4. Sampling 100 registered voters and determining how many voted for the president in the last election

Continuous random variables take on values at every point over a given interval. Thus continuous random variables have no gaps or unassumed values. It could be said that continuous random variables are generated from experiments in which things are “measured” not “counted.” For example, if a person is assembling a product component, the time it takes to accomplish that feat could be any value within a reasonable range such as 3 minutes 36.4218 seconds or 5 minutes 17.5169 seconds. A list of measures for which continuous random variables might be generated would include time, height, weight, and volume. Other examples of experiments that yield continuous random variables include the following:

1. Sampling the volume of liquid nitrogen in a storage tank
2. Measuring the time between customer arrivals at a retail outlet
3. Measuring the lengths of newly designed automobiles
4. Measuring the weight of grain in a grain elevator at different points of time

Once continuous data are measured and recorded, they become discrete data because the data are rounded off to a discrete number. Thus in actual practice, virtually all business data are discrete. However, for practical reasons, data analysis is facilitated greatly by using continuous distributions on data that were continuous originally.

The outcomes for random variables and their associated probabilities can be organized into distributions. The two types of distributions are **discrete distributions**, constructed from discrete random variables, and **continuous distributions**, based on continuous random variables.

❖ Discrete Distributions

❖ Mean or Expected Value

The **mean** or **expected value** of a discrete distribution is the long-run average of occurrences. We must realize that any one trial using a discrete random variable yields only one outcome. However, if the process is repeated long enough, the average of the outcomes are most likely

to approach a long-run average, expected value, or mean value. This mean, or expected, value is computed as follows.

**MEAN OR EXPECTED VALUE
OF A DISCRETE
DISTRIBUTION**

$$\mu = E(x) = \sum[x \cdot P(x)]$$

where

$E(x)$ = long-run average

x = an outcome

$P(x)$ = probability of that outcome

Variance and Standard Deviation of a Discrete Distribution

The variance and standard deviation of a discrete distribution are solved for by using the outcomes (x) and probabilities of outcomes [$P(x)$] in a manner similar to that of computing a mean. In addition, the computations for variance and standard deviations use the mean of the discrete distribution. The formula for computing the variance follows.

**VARIANCE OF A DISCRETE
DISTRIBUTION**

$$\sigma^2 = \sum[(x - \mu)^2 \cdot P(x)]$$

where

x = an outcome

$P(x)$ = probability of a given outcome

μ = mean

The standard deviation is then computed by taking the square root of the variance.

**STANDARD DEVIATION OF A
DISCRETE DISTRIBUTION**

$$\sigma = \sqrt{\sum[(x - \mu)^2 \cdot P(x)]}$$

❖ Binomial Distribution

Perhaps the most widely known of all discrete distributions is the **binomial distribution**.

The binomial distribution has been used for hundreds of years. Several assumptions underlie the use of the binomial distribution:

**ASSUMPTIONS OF THE
BINOMIAL DISTRIBUTION**

- The experiment involves n identical trials.
- Each trial has only two possible outcomes denoted as success or as failure.
- Each trial is independent of the previous trials.
- The terms p and q remain constant throughout the experiment, where the term p is the probability of getting a success on any one trial and the term $q = (1 - p)$ is the probability of getting a failure on any one trial.

❖ Mean and Standard Deviation of a Binomial Distribution

A binomial distribution has an expected value or a long-run average, which is denoted by m . The value of m is determined by $n p$. For example, if $n = 10$ and $p = .4$, then $m = n p = (10)(.4) = 4$. The long-run average or expected value means that, if n items are sampled over and over for a long time and if p is the probability of getting a success on one trial, the average number of successes per sample is expected to be $n p$. If 40% of all graduate business students at a large university are women and if random samples of 10 graduate business students are selected many times, the expectation is that, on average, four of the 10 students would be women.

MEAN AND STANDARD
DEVIATION OF A BINOMIAL
DISTRIBUTION

$$\mu = n \cdot p$$
$$\sigma = \sqrt{n \cdot p \cdot q}$$

Illustration -1

Solve the following problems by using the binomial formula.

- If $n = 4$ and $p = .10$, find $P(x = 3)$.
- If $n = 7$ and $p = .80$, find $P(x = 4)$.
- If $n = 10$ and $p = .60$, find $P(x \geq 7)$.
- If $n = 12$ and $p = .45$, find $P(5 \leq x \leq 7)$

a) $n = 4$ $p = .10$ $q = .90$

$$P(x=3) = {}_4C_3(.10)^3(.90)^1 = 4(.001)(.90) = \mathbf{.0036}$$

b) $n = 7$ $p = .80$ $q = .20$

$$P(x=4) = {}_7C_4(.80)^4(.20)^3 = 35(.4096)(.008) = \mathbf{.1147}$$

c) $n = 10$ $p = .60$ $q = .40$

$$P(x \geq 7) = P(x=7) + P(x=8) + P(x=9) + P(x=10) =$$

$${}_{10}C_7(.60)^7(.40)^3 + {}_{10}C_8(.60)^8(.40)^2 + {}_{10}C_9(.60)^9(.40)^1 + {}_{10}C_{10}(.60)^{10}(.40)^0 =$$

$$120(.0280)(.064) + 45(.0168)(.16) + 10(.0101)(.40) + 1(.0060)(1) =$$

$$.2150 + .1209 + .0403 + .0060 = \mathbf{.3822}$$

d) $n = 12$ $p = .45$ $q = .55$

$$P(5 \leq x \leq 7) = P(x=5) + P(x=6) + P(x=7) =$$

$${}_{12}C_5(.45)^5(.55)^7 + {}_{12}C_6(.45)^6(.55)^6 + {}_{12}C_7(.45)^7(.55)^5 =$$

$$792(.0185)(.0152) + 924(.0083)(.0277) + 792(.0037)(.0503) =$$

$$.2225 + .2124 + .1489 = \mathbf{.5838}$$

Illustration -2

Solve the following problems by using the binomial tables (Table A.2).

- If $n = 20$ and $p = .50$, find $P(x = 12)$.
- If $n = 20$ and $p = .30$, find $P(x > 8)$.
- If $n = 20$ and $p = .70$, find $P(x < 12)$.
- If $n = 20$ and $p = .90$, find $P(x \leq 16)$.
- If $n = 15$ and $p = .40$, find $P(4 \leq x \leq 9)$.
- If $n = 10$ and $p = .60$, find $P(x \geq 7)$.

By Table A.2:

a) $n = 20$ $p = .50$

$$P(x=12) = \mathbf{.120}$$



b) $n = 20$ $p = .30$

$$P(x > 8) = P(x=9) + P(x=10) + P(x=11) + \dots + P(x=20) =$$
$$.065 + .031 + .012 + .004 + .001 + .000 = \mathbf{.113}$$

c) $n = 20$ $p = .70$

$$P(x < 12) = P(x=11) + P(x=10) + P(x=9) + \dots + P(x=0) =$$
$$.065 + .031 + .012 + .004 + .001 + .000 = \mathbf{.113}$$

d) $n = 20$ $p = .90$

$$P(x \leq 16) = P(x=16) + P(x=15) + P(x=14) + \dots + P(x=0) =$$
$$.090 + .032 + .009 + .002 + .000 = \mathbf{.133}$$

e) $n = 15$ $p = .40$

$$P(4 \leq x \leq 9) = P(x=4) + P(x=5) + P(x=6) + P(x=7) + P(x=8) + P(x=9)$$
$$= .127 + .186 + .207 + .177 + .118 + .061 = \mathbf{.876}$$

f) $n = 10$ $p = .60$

$$P(x \geq 7) = P(x=7) + P(x=8) + P(x=9) + P(x=10) =$$



$$.215 + .122 + .040 + .006 = .382$$

Illustration -3

The Wall Street Journal reported some interesting statistics on the job market. One statistic is that 40% of all workers say they would change jobs for “slightly higher pay.” In addition, 88% of companies say that there is a shortage of qualified job candidates. Suppose 16 workers are randomly selected and asked if they would change jobs for “slightly higher pay.”

- a. What is the probability that nine or more say yes?
- b. What is the probability that three, four, five, or six say yes?
- c. If 13 companies are contacted, what is the probability that exactly 10 say there is a shortage of qualified job candidates?
- d. If 13 companies are contacted, what is the probability that all of the companies say there is a shortage of qualified job candidates?
- e. If 13 companies are contacted, what is the expected number of companies that would say there is a shortage of qualified job candidates?

$n = 16$ $p = .40$

$P(x \geq 9)$: from Table A.2:

<u>x</u>	<u>Prob</u>
9	.084
10	.039
11	.014
12	.004
13	<u>.001</u>
	.142

$P(3 \leq x \leq 6)$:

<u>x</u>	<u>Prob</u>
3	.047
4	.101
5	.162
6	<u>.198</u>
	.508

$$n = 13 \quad p = .88$$

$$P(x = 10) = {}_{13}C_{10}(.88)^{10}(.12)^3 = 286(.278500976)(.001728) = \mathbf{.1376}$$

$$P(x = 13) = {}_{13}C_{13}(.88)^{13}(.12)^0 = (1)(.1897906171)(1) = \mathbf{.1898}$$

$$\text{Expected Value} = \mu = n p = 13(.88) = \mathbf{11.44}$$

Illustration -4

In the past few years, outsourcing overseas has become more frequently used than ever before by U.S. companies. However, outsourcing is not without problems. A recent survey by Purchasing indicates that 20% of the companies that outsource overseas use a consultant. Suppose 15 companies that outsource overseas are randomly selected.

- What is the probability that exactly five companies that outsource overseas use a consultant?
- What is the probability that more than nine companies that outsource overseas use a consultant?
- What is the probability that none of the companies that outsource overseas use a consultant?
- What is the probability that between four and seven (inclusive) companies that outsource overseas use a consultant?



e. Construct a graph for this binomial distribution. In light of the graph and the expected value, explain why the probability results from parts (a) through (d) were obtained.

$$n = 15 \quad p = .20$$

$$\text{a) } P(x = 5) = {}_{15}C_5(.20)^5(.80)^{10} =$$

$$3003(.00032)(.1073742) = \mathbf{.1032}$$

b) $P(x > 9)$: Using Table A.2

$$\begin{aligned} P(x = 10) + P(x = 11) + \dots + P(x = 15) = \\ .000 + .000 + \dots + .000 = \mathbf{.000} \end{aligned}$$

$$\text{c) } P(x = 0) = {}_{15}C_0(.20)^0(.80)^{15} =$$

$$(1)(1)(.035184) = \mathbf{.0352}$$

d) $P(4 \leq x \leq 7)$: Using Table A.2

$$\begin{aligned} P(x = 4) + P(x = 5) + P(x = 6) + P(x = 7) = \\ .188 + .103 + .043 + .014 = \mathbf{.348} \end{aligned}$$

e)

Binomial Distribution for $n=15$ and $p=.20$

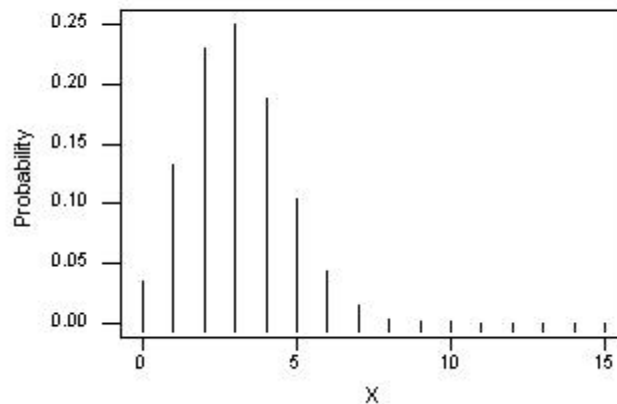


Illustration -5 According to Cerulli Associates of Boston, 30% of all CPA financial advisors have an average client size between \$500,000 and \$1 million. Thirty-four percent have an average client size between \$1 million and \$5 million. Suppose a complete list of all CPA financial advisors is available and 18 are randomly selected from that list.

- What is the expected number of CPA financial advisors that have an average client size between \$500,000 and \$1 million? What is the expected number with an average client size between \$1 million and \$5 million?
- What is the probability that at least eight CPA financial advisors have an average client size between \$500,000 and \$1 million?
- What is the probability that two, three, or four CPA financial advisors have an average client size between \$1 million and \$5 million?
- What is the probability that none of the CPA financial advisors have an average client size between \$500,000 and \$1 million? What is the probability that none have an average client size between \$1 million and \$5 million? Which probability is higher and why?

$n = 18$

a) $p = .30$ $\mu = 18(.30) = 5.4$



$$p = .34 \quad \mu = 18(.34) = \mathbf{6.12}$$

b) $P(x \geq 8) \quad n = 18 \quad p = .30$

from Table A.2

<u>x</u>	<u>Prob</u>
8	.081
9	.039
10	.015
11	.005
12	<u>.001</u>
	.141

c) $n = 18 \quad p = .34$

$$P(2 \leq x \leq 4) = P(x = 2) + P(x = 3) + P(x = 4) =$$

$${}_{18}C_2(.34)^2(.66)^{16} + {}_{18}C_3(.34)^3(.66)^{15} + {}_{18}C_4(.34)^4(.66)^{14} =$$

$$.0229 + .0630 + .1217 = \mathbf{.2076}$$

d) $n = 18 \quad p = .30 \quad x = 0$

$${}_{18}C_0(.30)^0(.70)^{18} = \mathbf{.00163}$$

$n = 18 \quad p = .34 \quad x = 0$

$${}_{18}C_0(.34)^0(.66)^{18} = \mathbf{.0005}$$



The probability that none are in the \$500,000 to \$1,000,000 is higher because there is a smaller percentage in that category which is closer to zero.

❖ POISSON DISTRIBUTION

The Poisson distribution is another discrete distribution. It is named after Simeon-Denis Poisson (1781–1840), a French mathematician, who published its essentials in a paper in 1837. The Poisson distribution and the binomial distribution have some similarities but also several differences. The binomial distribution describes a distribution of two possible outcomes designated as successes and failures from a given number of trials. The **Poisson distribution** focuses only on the number of discrete occurrences over some interval or continuum. A Poisson experiment does not have a given number of trials (n) as a binomial experiment does. For example, whereas a binomial experiment might be used to determine how many U.S.-made cars are in a random sample of 20 cars, a Poisson experiment might focus on the number of cars randomly arriving at an automobile repair facility during a 10-minute interval.

The Poisson distribution describes the occurrence of rare events. In fact, the Poisson formula has been referred to as the law of improbable events. For example, serious accidents at a chemical plant are rare, and the number per month might be described by the Poisson distribution. The Poisson distribution often is used to describe the number of random arrivals per some time interval. If the number of arrivals per interval is too frequent, the time interval can be reduced enough so that a rare number of occurrences is expected.

Another example of a Poisson distribution is the number of random customer arrivals per five-minute interval at a small boutique on weekday mornings. The Poisson distribution also has an application in the field of management science. The models used in queuing theory (theory of waiting lines) usually are based on the assumption that the Poisson distribution is the proper distribution to describe random arrival rates over a period of time. The Poisson distribution has the following characteristics:

- It is a discrete distribution.
- It describes rare events.
- Each occurrence is independent of the other occurrences.
- It describes discrete occurrences over a continuum or interval.
- The occurrences in each interval can range from zero to infinity.
- The expected number of occurrences must hold constant throughout the experiment.

Examples of Poisson-type situations include the following:

1. Number of telephone calls per minute at a small business
2. Number of hazardous waste sites per county in the United States
3. Number of arrivals at a turnpike tollbooth per minute between 3 A.M. and 4 A.M. in January on the Kansas Turnpike
4. Number of sewing flflaws per pair of jeans during production
5. Number of times a tire blows on a commercial airplane per week

POISSON FORMULA

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

where

$$x = 0, 1, 2, 3, \dots$$

λ = long-run average

$$e = 2.718282$$

Illustration -5

Find the following values by using the Poisson tables in Appendix A.

- a. $\text{Prob}(x=5 \mid \lambda = 2.3)$
- b. $\text{Prob}(x=2 \mid \lambda = 3.9)$
- c. $\text{Prob}(x \leq 3 \mid \lambda = 4.1)$
- d. $\text{Prob}(x=0 \mid \lambda = 2.7)$
- e. $\text{Prob}(x=1 \mid \lambda = 5.4)$
- f. $\text{Prob}(4 < x < 8 \mid \lambda = 4.4)$

a) $\text{Prob}(x=5 \mid \lambda = 2.3) =$

$$\frac{(2.3^5)(e^{-2.3})}{5!} = \frac{(64.36343)(.1002588)}{(120)} = .0538$$

b) $\text{Prob}(x=2 \mid \lambda = 3.9) =$



$$\frac{(3.9^2)(e^{-3.9})}{2!} = \frac{(15.21)(.02024)}{(2)} = .1539$$

c) $\text{Prob}(x \leq 3 \mid \lambda = 4.1) =$

$$\text{Prob}(x=3) + \text{Prob}(x=2) + \text{Prob}(x=1) + \text{Prob}(x=0) =$$

$$\frac{(4.1^3)(e^{-4.1})}{3!} = \frac{(68.921)(.016574)}{6} = .1904$$

$$+ \frac{(4.1^2)(e^{-4.1})}{2!} = \frac{(16.81)(.016573)}{2} = .1393$$

$$+ \frac{(4.1^1)(e^{-4.1})}{1!} = \frac{(4.1)(.016573)}{1} = .0679$$

$$+ \frac{(4.1^0)(e^{-4.1})}{0!} = \frac{(1)(.016573)}{1} = .0166$$

$$.1904 + .1393 + .0679 + .0166 = .4142$$

d) $\text{Prob}(x=0 \mid \lambda = 2.7) =$

$$\frac{(2.7^0)(e^{-2.7})}{0!} = \frac{(1)(.0672)}{1} = .0672$$

$$e) \text{ Prob}(x=1 \mid \lambda = 5.4) =$$

$$\frac{(5.4^1)(e^{-5.4})}{1!} = \frac{(5.4)(.0045)}{1} = .0244$$

$$f) \text{ Prob}(4 < x < 8 \mid \lambda = 4.4) =$$

$$\text{Prob}(x=5 \mid \lambda = 4.4) + \text{Prob}(x=6 \mid \lambda = 4.4) + \text{Prob}(x=7 \mid \lambda = 4.4) =$$

$$\frac{(4.4^5)(e^{-4.4})}{5!} + \frac{(4.4^6)(e^{-4.4})}{6!} + \frac{(4.4^7)(e^{-4.4})}{7!} =$$

$$\frac{(1649.162)(.012277)}{120} + \frac{(7256.314)(.012277)}{720} + \frac{(31927.781)(.012277)}{5040}$$

$$= .1687 + .1237 + .0778 = .3702$$

Illustration -6

On Monday mornings, the First National Bank only has one teller window open for deposits and withdrawals. Experience has shown that the average number of arriving customers in a four-minute interval on Monday mornings is 2.8, and each teller can serve more than that number efficiently. These random arrivals at this bank on Monday mornings are Poisson distributed.

- What is the probability that on a Monday morning exactly six customers will arrive in a four-minute interval?
- What is the probability that no one will arrive at the bank to make a deposit or withdrawal during a four-minute interval?
- Suppose the teller can serve no more than four customers in any four-minute interval at this window on a Monday morning. What is the probability that, during any given four-minute interval, the teller will be unable to meet the demand? What is the probability that the teller will be able to meet the demand? When demand cannot be met during any given interval, a

second window is opened. What percentage of the time will a second window have to be opened?

d. What is the probability that exactly three people will arrive at the bank during a two-minute period on Monday mornings to make a deposit or a withdrawal?

What is the probability that five or more customers will arrive during an eight minute period?

$$\lambda = 2.8 \mid 4 \text{ minutes}$$

a) $\text{Prob}(x=6 \mid \lambda = 2.8)$

from Table A.3 **.0407**

b) $\text{Prob}(x=0 \mid \lambda = 2.8) =$

from Table A.3 **.0608**

c) Unable to meet demand if $x > 4 \mid 4 \text{ minutes}$

<u>x</u>	<u>Prob.</u>
5	.0872
6	.0407
7	.0163
8	.0057
9	.0018
10	.0005
<u>11</u>	<u>.0001</u>
$x > 4$.1523

.1523 probability of being unable to meet the demand

Probability of meeting the demand = $1 - (.1523) = \mathbf{.8477}$

15.23% of the time a second window will need to be opened.

d) $\lambda = 2.8 \text{ arrivals} \mid 4 \text{ minutes}$

$\text{Prob}(x=3 \text{ arrivals} \mid 2 \text{ minutes}) = ??$

Lambda must be changed to the same interval ($\frac{1}{2}$ the size)

New lambda = $1.4 \text{ arrivals} \mid 2 \text{ minutes}$

$\text{Prob}(x=3 \mid \lambda=1.4) =$ from Table A.3 = **.1128**

$$\text{Prob}(x \geq 5 \mid 8 \text{ minutes}) = ??$$

Lambda must be changed to the same interval(twice the size):

$$\text{New lambda} = 5.6 \text{ arrivals} \mid 8 \text{ minutes}$$

$$\text{Prob}(x \geq 5 \mid \lambda = 5.6):$$

From Table A.3:

<u>X</u>	<u>Prob.</u>
5	.1697
6	.1584
7	.1267
8	.0887
9	.0552
10	.0309
11	.0157
12	.0073
13	.0032
14	.0013
15	.0005
16	.0002
<u>17</u>	<u>.0001</u>
$x \geq 5$.6579

Illustration -7

According to the United National Environmental Program and World Health Organization, in Mumbai, India, air pollution standards for particulate matter are exceeded an average of 5.6 days in every three-week period. Assume that the distribution of number of days exceeding the standards per three-week period is Poisson distributed.

a. What is the probability that the standard is not exceeded on any day during a three-week period?

- b. What is the probability that the standard is exceeded exactly six days of a three-week period?
- c. What is the probability that the standard is exceeded 15 or more days during a three-week period? If this outcome actually occurred, what might you conclude?

$$\lambda = 5.6 \text{ days} \mid 3 \text{ weeks}$$

a) $\text{Prob}(x=0 \mid \lambda = 5.6)$:

from Table A.3 = **.0037**

b) $\text{Prob}(x=6 \mid \lambda = 5.6)$:

from Table A.3 = **.1584**

c) $\text{Prob}(x \geq 15 \mid \lambda = 5.6)$:

<u>x</u>	<u>Prob.</u>
15	.0005
16	.0002
<u>17</u>	<u>.0001</u>
$x \geq 15$.0008

Because this probability is so low, if it actually occurred, the researcher would actually have to question the Lambda value as too low for this period.

Illustration -8

The average number of annual trips per family to amusement parks in the United States is Poisson distributed, with a mean of 0.6 trips per year. What is the probability of randomly selecting an American family and finding the following?

- a. The family did not make a trip to an amusement park last year.
- b. The family took exactly one trip to an amusement park last year.



- c. The family took two or more trips to amusement parks last year.
- d. The family took three or fewer trips to amusement parks over a three-year period.
- e. The family took exactly four trips to amusement parks during a six-year period.

$$\lambda = 0.6 \text{ trips} \mid 1 \text{ year}$$

a) $\text{Prob}(x=0 \mid \lambda = 0.6)$:

from Table A.3 = **.5488**

b) $\text{Prob}(x=1 \mid \lambda = 0.6)$:

from Table A.3 = **.3293**

c) $\text{Prob}(x \geq 2 \mid \lambda = 0.6)$:

from Table A.3

<u>x</u>	<u>Prob.</u>
2	.0988
3	.0198
4	.0030
5	.0004
<u>6</u>	<u>.0000</u>
x ≥ 2	.1220

d) $\text{Prob}(x \leq 3 \mid 3 \text{ year period})$:

The interval length has been increased (3 times)

New Lambda = $\lambda = 1.8 \text{ trips} \mid 3 \text{ years}$

$\text{Prob}(x \leq 3 \mid \lambda = 1.8)$:

from Table A.3	x	<u>Prob.</u>
	0	.1653
	1	.2975
	2	.2678
	<u>3</u>	<u>.1607</u>
	$x \leq 3$.8913

e) Prob($x=4$ | 6 years):

The interval has been increased (6 times)

New Lambda = $\lambda = 3.6$ trips | 6 years

Prob($x=4$ | $\lambda = 3.6$):

from Table A.3 = **.1912**

❖ Continuous Distribution

THE UNIFORM DISTRIBUTION

The **uniform distribution**, sometimes referred to as the **rectangular distribution**, is a relatively simple continuous distribution in which the same height, or $f(x)$, is obtained over a range of values. The following probability density function defines a uniform distribution.

**PROBABILITY DENSITY
FUNCTION OF A UNIFORM
DISTRIBUTION**

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{for all other values} \end{cases}$$

**MEAN AND STANDARD
DEVIATION OF A UNIFORM
DISTRIBUTION**

$$\mu = \frac{a+b}{2}$$

$$\sigma = \frac{b-a}{\sqrt{12}}$$

As an example,

suppose a production line is set up to manufacture machine braces in lots of five per minute during a shift. When the lots are weighed, variation among the weights is detected, with lot weights ranging from 41 to 47 grams in a uniform distribution. The height of this distribution is

$$f(x) = \text{Height} = \frac{1}{(b - a)} = \frac{1}{(47 - 41)} = \frac{1}{6}$$

The mean and standard deviation of this distribution are

$$\text{Mean} = a+b / 2 = 41+47 / 2 =44$$

$$\text{Standard Deviation} = \frac{b - a}{\sqrt{12}} = \frac{47 - 41}{\sqrt{12}} = \frac{6}{3.464} = 1.732$$

● Determining Probabilities in a Uniform Distribution

With discrete distributions, the probability function yields the value of the probability. For continuous distributions, probabilities are calculated by determining the area over an interval of the function. With continuous distributions, there is no area under the curve for a single point. The following equation is used to determine the probabilities of x for a uniform distribution between a and b .

PROBABILITIES IN A
UNIFORM DISTRIBUTION

$$P(x) = \frac{x_2 - x_1}{b - a}$$

where:

$$a \leq x_1 \leq x_2 \leq b$$

6.2 x is uniformly distributed over a range of values from 8 to 21.

- What is the value of $f(x)$ for this distribution?
- Determine the mean and standard deviation of this distribution.
- Probability of $P(10 \leq x < 17) = ?$
- Probability of $P(x > 22) = ?$
- Probability of $P(x \geq 7) = ?$

$$a = 8 \quad b = 21$$

$$\text{a) } f(x) = \frac{1}{b - a} = \frac{1}{21 - 8} = \frac{1}{13}$$

$$b) \mu = \frac{a+b}{2} = \frac{8+21}{2} = \frac{29}{2} = \mathbf{14.5}$$

$$\sigma = \frac{b-a}{\sqrt{12}} = \frac{21-8}{\sqrt{12}} = \frac{13}{\sqrt{12}} = \mathbf{3.7528}$$

$$c) P(10 \leq x < 17) = \frac{17-10}{21-8} = \frac{7}{13} = \mathbf{.5385}$$

$$d) P(x > 22) = \mathbf{.0000}$$

$$e) P(x \geq 7) = \mathbf{1.0000}$$

Illustration -9

The average fill volume of a regular can of soft drink is 12 ounces. Suppose the fill volume of these cans ranges from 11.97 to 12.03 ounces and is uniformly distributed. What is the height of this distribution? What is the probability that a randomly selected can contains more than 12.01 ounces of fluid? What is the probability that the fill volume is between 11.98 and 12.01 ounces?

$$a = 11.97 \quad b = 12.03$$

$$\text{Height} = \frac{1}{b-a} = \frac{1}{12.03-11.97} = \mathbf{16.667}$$

$$P(x > 12.01) = \frac{12.03-12.01}{12.03-11.97} = \mathbf{.3333}$$

$$P(11.98 < x < 12.01) = \frac{12.01 - 11.98}{12.03 - 11.97} = .5000$$

Suppose the average U.S. household spends \$2,100 a year on all types of insurance.

Suppose the figures are uniformly distributed between the values of \$400 and \$3,800.

What are the standard deviation and the height of this distribution? What proportion

of households spends more than \$3,000 a year on insurance? More than \$4,000?

Between \$700 and \$1,500?

$$\mu = 2100 \quad a = 400 \quad b = 3800$$

$$\sigma = \frac{b - a}{\sqrt{12}} = \frac{3800 - 400}{\sqrt{12}} = \mathbf{981.5}$$

$$\text{Height} = \frac{1}{b - a} = \frac{3800 - 400}{\sqrt{12}} = \mathbf{.000294}$$

$$P(x > 3000) = \frac{3800 - 3000}{3800 - 400} = \frac{800}{3400} = \mathbf{.2353}$$

$$P(x > 4000) = \mathbf{.0000}$$

$$P(700 < x < 1500) = \frac{1500 - 700}{3800 - 400} = \frac{800}{3400} = \mathbf{.2353}$$

Normal Distribution

Probably the most widely known and used of all distributions is the **normal distribution**. It fits many human characteristics, such as height, weight, length, speed, IQ, scholastic achievement, and years of life expectancy, among others. Like their human counterparts, living things in nature, such as trees, animals, insects, and others, have many characteristics that are normally distributed. Many variables in business and industry also are normally distributed. Some

examples of variables that could produce normally distributed measurements include the annual cost of household insurance, the cost per square foot of renting warehouse space, and managers' satisfaction with support from ownership on a five-point scale. In addition, most items produced or filled by machines are normally distributed.

History of the Normal Distribution

Discovery of the normal curve of errors is generally credited to mathematician and astronomer Karl Gauss (1777–1855), who recognized that the errors of repeated measurement of objects are often normally distributed.* Thus the normal distribution is sometimes referred to as the Gaussian distribution or the normal curve of error. A modern-day analogy of Gauss's work might be the distribution of measurements of machine-produced parts, which often yield a normal curve of error around a mean specification.

To a lesser extent, some credit has been given to Pierre-Simon de Laplace (1749–1827) for discovering the normal distribution. However, many people now believe that Abraham de Moivre (1667–1754), a French mathematician, first understood the normal distribution. De Moivre determined that the binomial distribution approached the normal distribution as a limit. De Moivre worked with remarkable accuracy. His published table values for the normal curve are only a few ten-thousandths off the values of currently published tables.

The normal distribution exhibits the following characteristics.

- It is a continuous distribution.
- It is a symmetrical distribution about its mean.
- It is asymptotic to the horizontal axis.
- It is unimodal.
- It is a family of curves.
- Area under the curve is 1

The normal distribution is symmetrical. Each half of the distribution is a mirror image of the other half. Many normal distribution tables contain probability values for only one side of the distribution because probability values for the other side of the distribution are identical because of symmetry.

Probability Density Function of the Normal Distribution

The normal distribution is described or characterized by two parameters: the mean, μ , and the standard deviation, σ . The values of μ and σ produce a normal distribution. The density function of the normal distribution is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2[(x-\mu)/\sigma]^2}$$

where

μ = mean of x
 σ = standard deviation of x
 π = 3.14159 . . . , and
 e = 2.71828 . . .

z FORMULA

$$z = \frac{x - \mu}{\sigma}, \quad \sigma \neq 0$$

Illustration -10

Determine the probabilities for the following normal distribution problems.

- $\text{Prob}(x \leq 635 \mid \mu = 604, \sigma = 56.8)$
 - $\text{Prob}(x < 20 \mid \mu = 48, \sigma = 12)$
 - $\text{Prob}(100 \leq x < 150 \mid \mu = 111, \sigma = 33.8)$
 - $\text{Prob}(250 < x < 255 \mid \mu = 264, \sigma = 10.9)$
 - $\text{Prob}(x > 35 \mid \mu = 37, \sigma = 4.35)$
 - $\text{Prob}(x \geq 170 \mid \mu = 156, \sigma = 11.4)$
- $\text{Prob}(x \leq 635 \mid \mu = 604, \sigma = 56.8):$



$$z = \frac{x - \mu}{\sigma} = \frac{635 - 604}{56.8} = 0.55$$

Table A.5 value for $z = 0.55$: .2088

$$\text{Prob}(x \leq 635) = .2088 + .5000 = \mathbf{.7088}$$

b) $\text{Prob}(x < 20 \mid \mu = 48, \sigma = 12)$:

$$z = \frac{x - \mu}{\sigma} = \frac{20 - 48}{12} = -2.33$$

Table A.5 value for $z = -2.33$: .4901

$$\text{Prob}(x < 20) = .5000 - .4901 = \mathbf{.0099}$$

c) $\text{Prob}(100 \leq x < 150 \mid \mu = 111, \sigma = 33.8)$:

$$z = \frac{x - \mu}{\sigma} = \frac{150 - 111}{33.8} = 1.15$$

Table A.5 value for $z = 1.15$: .3749

$$z = \frac{x - \mu}{\sigma} = \frac{100 - 111}{33.8} = -0.33$$

Table A.5 value for $z = -0.33$: .1293

$$\text{Prob}(100 \leq x < 150) = .3749 + .1293 = \mathbf{.5042}$$

d) $\text{Prob}(250 < x < 255 \mid \mu = 264, \sigma = 10.9)$:

$$z = \frac{x - \mu}{\sigma} = \frac{250 - 264}{10.9} = -1.28$$

Table A.5 value for $z = -1.28$: .3997

$$z = \frac{x - \mu}{\sigma} = \frac{255 - 264}{10.9} = -0.83$$

Table A.5 value for $z = -0.83$: .2967

$$\text{Prob}(250 < x < 255) = .3997 - .2967 = \mathbf{.1030}$$

e) $\text{Prob}(x > 35 \mid \mu = 37, \sigma = 4.35)$:

$$z = \frac{x - \mu}{\sigma} = \frac{35 - 37}{4.35} = -0.46$$

Table A.5 value for $z = -0.46$: .1772

$$\text{Prob}(x > 35) = .1772 + .5000 = \mathbf{.6772}$$

f) $\text{Prob}(x \geq 170 \mid \mu = 156, \sigma = 11.4)$:

$$z = \frac{x - \mu}{\sigma} = \frac{170 - 156}{11.4} = 1.23$$

Table A.5 value for $z = 1.23$: .3907

$$\text{Prob}(x \geq 170) = .5000 - .3907 = \mathbf{.1093}$$

Illustration -11

Tompkins Associates reports that the mean clear height for a Class A warehouse in the United States is 22 feet. Suppose clear heights are normally distributed and that the standard deviation is 4 feet. A Class A warehouse in the United States is randomly selected.

- What is the probability that the clear height is greater than 17 feet?
- What is the probability that the clear height is less than 13 feet?
- What is the probability that the clear height is between 25 and 31 feet?

$$\mu = 22 \quad \sigma = 4$$

a) $\text{Prob}(x > 17)$:

$$z = \frac{x - \mu}{\sigma} = \frac{17 - 22}{4} = -1.25$$

area between $x = 17$ and $\mu = 22$ from table A.5 is .3944

$$\text{Prob}(x > 17) = .3944 + .5000 = \mathbf{.8944}$$

b) Prob($x < 13$):

$$z = \frac{x - \mu}{\sigma} = \frac{13 - 22}{4} = -2.25$$

from table A.5, area = .4878

$$\text{Prob}(x < 13) = .5000 - .4878 = \mathbf{.0122}$$

c) P($25 \leq x \leq 31$):

$$z = \frac{x - \mu}{\sigma} = \frac{31 - 22}{4} = 2.25$$

from table A.5, area = .4878

$$z = \frac{x - \mu}{\sigma} = \frac{25 - 22}{4} = 0.75$$

from table A.5, area = .2734

$$\text{Prob}(25 \leq x \leq 31) = \mathbf{.4878 - .2734 = .2144}$$

Illustration -12

Tool workers are subject to work-related injuries. One disorder, caused by strains to the hands and wrists, is called carpal tunnel syndrome. It strikes as many as 23,000 workers per year. The U.S. Labor Department estimates that the average cost of this disorder to employers and insurers is approximately \$30,000 per injured worker. Suppose these costs are normally distributed, with a standard deviation of \$9,000.

a. What proportion of the costs are between \$15,000 and \$45,000?

- b. What proportion of the costs are greater than \$50,000?
- c. What proportion of the costs are between \$5,000 and \$20,000?
- d. Suppose the standard deviation is unknown, but 90.82% of the costs are more than \$7,000. What would be the value of the standard deviation?
- e. Suppose the mean value is unknown, but the standard deviation is still \$9,000. How much would the average cost be if 79.95% of the costs were less than \$33,000?

$$\mu = \$30,000 \quad \sigma = \$9,000$$

- a) $\text{Prob}(\$15,000 \leq x \leq \$45,000)$:

$$z = \frac{x - \mu}{\sigma} = \frac{45,000 - 30,000}{9,000} = 1.67$$

From Table A.5, $z = 1.67$ yields: .4525

$$z = \frac{x - \mu}{\sigma} = \frac{15,000 - 30,000}{9,000} = -1.67$$

From Table A.5, $z = -1.67$ yields: .4525

$$\text{Prob}(\$15,000 \leq x \leq \$45,000) = .4525 + .4525 = \mathbf{.9050}$$

- b) $\text{Prob}(x > \$50,000)$:

$$z = \frac{x - \mu}{\sigma} = \frac{50,000 - 30,000}{9,000} = 2.22$$



From Table A.5, $z = 2.22$ yields: 4868

$$\text{Prob}(x > \$50,000) = .5000 - .4868 = \mathbf{.0132}$$

c) $\text{Prob}(\$5,000 \leq x \leq \$20,000)$:

$$z = \frac{x - \mu}{\sigma} = \frac{5,000 - 30,000}{9,000} = -2.78$$

From Table A.5, $z = -2.78$ yields: .4973

$$z = \frac{x - \mu}{\sigma} = \frac{20,000 - 30,000}{9,000} = -1.11$$

From Table A.5, $z = -1.11$ yields .3665

$$\text{Prob}(\$5,000 \leq x \leq \$20,000) = .4973 - .3665 = \mathbf{.1308}$$

d) 90.82% of the values are greater than $x = \$7,000$.

Then $x = \$7,000$ is in the lower half of the distribution and $.9082 - .5000 = 0.4082$ lie between x and μ .

From Table A.5, $z = -1.33$ is associated with an area of .4082.

Solving for σ :

$$z = \frac{x - \mu}{\sigma}$$

$$-1.33 = \frac{7,000 - 30,000}{\sigma}$$

$$\sigma = \mathbf{17,293.23}$$

e) $\sigma = \$9,000$. If 79.95% of the costs are less than \$33,000, $x = \$33,000$ is in the upper half of the distribution and $.7995 - .5000 = .2995$ of the values lie between \$33,000 and the mean.

From Table A.5, an area of .2995 is associated with $z = 0.84$

Solving for μ :

$$z = \frac{x - \mu}{\sigma}$$

$$0.84 = \frac{33,000 - \mu}{9,000}$$

$$\mu = \mathbf{\$25,440}$$

Illustration -13

Suppose you are working with a data set that is normally distributed, with a mean of 200 and a standard deviation of 47. Determine the value of x from the following information.

- a. 60% of the values are greater than x .
- b. x is less than 17% of the values.
- c. 22% of the values are less than x .
- d. x is greater than 55% of the values.

$$\mu = 200, \quad \sigma = 47 \quad \text{Determine } x$$

- a) 60% of the values are greater than x :

Since 50% of the values are greater than the mean, $\mu = 200$, 10% or .1000 lie between x and the mean. From Table A.5, the z value associated with an area of .1000 is $z = -0.25$. The z value is negative since x is below the mean.

Substituting $z = -0.25$, $\mu = 200$, and $\sigma = 47$ into the formula and solving for x :

$$z = \frac{x - \mu}{\sigma}$$

$$-0.25 = \frac{x - 200}{47}$$

$$x = \mathbf{188.25}$$

- b) x is less than 17% of the values.

Since x is only less than 17% of the values, 33% (.5000- .1700) or .3300 lie

between x and the mean. Table A.5 yields a z value of 0.95 for an area of .3300. Using this $z = 0.95$, $\mu = 200$, and $\sigma = 47$, x can be solved for:

$$z = \frac{x - \mu}{\sigma}$$

$$0.95 = \frac{X - 200}{47}$$

$$x = \mathbf{244.65}$$

c) 22% of the values are less than x .

Since 22% of the values lie below x , 28% lie between x and the mean (.5000 - .2200). Table A.5 yields a z of -0.77 for an area of .2800. Using the z value of -0.77, $\mu = 200$, and $\sigma = 47$, x can be solved for:

$$z = \frac{x - \mu}{\sigma}$$

$$-0.77 = \frac{x - 200}{47}$$

$$x = \mathbf{163.81}$$

d) x is greater than 55% of the values.

Since x is greater than 55% of the values, 5% (.0500) lie between x and the

mean. From Table A.5, a z value of 0.13 is associated with an area of .05.

Using $z = 0.13$, $\mu = 200$, and $\sigma = 47$, x can be solved for:

$$z = \frac{x - \mu}{\sigma}$$

$$0.13 = \frac{x - 200}{47}$$

$$x = \mathbf{206.11}$$

Illustration -14

Data accumulated by the National Climatic Data Center shows that the average wind speed in miles per hour for St. Louis, Missouri, is 9.7. Suppose wind speed measurements are normally distributed for a given geographic location. If 22.45% of the time the wind speed measurements are more than 11.6 miles per hour, what is the standard deviation of wind speed in St. Louis?

$\mu = 9.7$ Since 22.45% are greater than 11.6, $x = 11.6$ is in the upper half of the distribution and .2755 (.5000 - .2245) lie between x and the mean. Table A.5 yields a $z = 0.76$ for an area of .2755.

Solving for σ :

$$z = \frac{x - \mu}{\sigma}$$

$$0.76 = \frac{11.6 - 9.7}{\sigma}$$

$$\sigma = \mathbf{2.5}$$

Suppose the mean clear height of all U.S. Class A warehouses is unknown but the



standard deviation is known to be 4 feet. What is the value of the mean clear height if 29% of U.S. Class A warehouses have a clear height less than 20 feet?

$$\text{Prob}(x < 20) = .2900$$

x is less than μ because of the percentage. Between x and μ is $.5000 - .2900 = .2100$ of the area. The z score associated with this area is -0.55. Solving for μ :

$$z = \frac{x - \mu}{\sigma}$$

$$-0.55 = \frac{20 - \mu}{4}$$

$$\mu = \mathbf{22.20}$$

Hypothesis Testing

Hypothesis testing was introduced by Ronald Fisher, Jerzy Neyman, Karl Pearson and Pearson's son, Egon Pearson. Hypothesis testing is a statistical method that is used in making statistical decisions using experimental data. Hypothesis Testing is basically an assumption that we make about the population parameter.

Key terms and concepts:

- **Null hypothesis:** Null hypothesis is a statistical hypothesis that assumes that the observation is due to a chance factor. Null hypothesis is denoted by; $H_0: \mu_1 = \mu_2$, which shows that there is no difference between the two population means.
- **Alternative hypothesis:** Contrary to the null hypothesis, the alternative hypothesis shows that observations are the result of a real effect.
- **Level of significance:** Refers to the degree of significance in which we accept or reject the null hypothesis. 100% accuracy is not possible for accepting or rejecting a hypothesis, so we therefore select a level of significance that is usually 5%.
- **Type I error:** When we reject the null hypothesis, although that hypothesis was true. Type I error is denoted by alpha. In hypothesis testing, the normal curve that shows the critical region is called the alpha region.
- **Type II errors:** When we accept the null hypothesis, but it is false. Type II errors are denoted by beta. In Hypothesis testing, the normal curve that shows the acceptance region is called the beta region.
- **Power:** Usually known as the probability of correctly accepting the null hypothesis. $1 - \beta$ is called power of the analysis.
- **One-tailed test:** When the given statistical hypothesis is one value like $H_0: \mu_1 = \mu_2$, it is called the one-tailed test.
- **Two-tailed test:** When the given statistics hypothesis assumes a less than or greater than value, it is called the two-tailed test.

What is Hypothesis Testing?

A **statistical hypothesis** is an assumption about a population parameter. This assumption may or may not be true. **Hypothesis testing** refers to the formal procedures used by statisticians to accept or reject statistical hypotheses.

Statistical Hypotheses

The best way to determine whether a statistical hypothesis is true would be to examine the entire population. Since that is often impractical, researchers typically examine a random sample from the population. If sample data are not consistent with the statistical hypothesis, the hypothesis is rejected.

There are two types of statistical hypotheses.

- **Null hypothesis.** The null hypothesis, denoted by H_0 , is usually the hypothesis that sample observations result purely from chance.
- **Alternative hypothesis.** The alternative hypothesis, denoted by H_1 or H_a , is the hypothesis that sample observations are influenced by some non-random cause.

For example, suppose we wanted to determine whether a coin was fair and balanced. A null hypothesis might be that half the flips would result in Heads and half, in Tails. The alternative hypothesis might be that the number of Heads and Tails would be very different. Symbolically, these hypotheses would be expressed as

$$H_0: P = 0.5$$

$$H_a: P \neq 0.5$$

Suppose we flipped the coin 50 times, resulting in 40 Heads and 10 Tails. Given this result, we would be inclined to reject the null hypothesis. We would conclude, based on the evidence, that the coin was probably not fair and balanced.

Can We Accept the Null Hypothesis?

Some researchers say that a hypothesis test can have one of two outcomes: you accept the null hypothesis, or you reject the null hypothesis. Many statisticians, however, take issue with the notion of "accepting the null hypothesis." Instead, they say: you reject the null hypothesis, or you fail to reject the null hypothesis.

Why the distinction between "acceptance" and "failure to reject?" Acceptance implies that the null hypothesis is true. Failure to reject implies that the data are not sufficiently persuasive for us to prefer the alternative hypothesis over the null hypothesis

Hypothesis Tests

Statisticians follow a formal process to determine whether to reject a null hypothesis, based on sample data. This process, called **hypothesis testing**, consists of four steps.

- State the hypotheses. This involves stating the null and alternative hypotheses. The hypotheses are stated in such a way that they are mutually exclusive. That is, if one is true, the other must be false.
- Formulate an analysis plan. The analysis plan describes how to use sample data to evaluate the null hypothesis. The evaluation often focuses around a single test statistic.
- Analyse sample data. Find the value of the test statistic (mean score, proportion, t statistic, z-score, etc.) described in the analysis plan.
- Interpret results. Apply the decision rule described in the analysis plan. If the value of the test statistic is unlikely, based on the null hypothesis, reject the null hypothesis.

Decision Errors

Two types of errors can result from a hypothesis test.

- **Type I error.** A Type I error occurs when the researcher rejects a null hypothesis when it is true. The probability of committing a Type I error is called the **significance level**. This probability is also called **alpha**, and is often denoted by α .
- **Type II error.** A Type II error occurs when the researcher fails to reject a null hypothesis that is false. The probability of committing a Type II error is called **Beta**, and is often denoted by β . The probability of not committing a Type II error is called the **Power** of the test.

● Types of Hypotheses

Three types of hypotheses that will be explored here:

1. Research hypotheses
2. Statistical hypotheses
3. Substantive hypotheses

Although much of the focus will be on testing statistical hypotheses, it is also important for business decision makers to have an understanding of both research and substantive hypotheses.

● Research Hypotheses

Research hypotheses are most nearly like hypotheses defined earlier. A **research hypothesis** is a statement of what the researcher believes will be the outcome of an experiment or a study. Before studies are undertaken, business researchers often have some idea or theory based on experience or previous work as to how the study will turn out. These ideas, theories, or notions established before an experiment or study is conducted are research hypotheses.

Some examples of research hypotheses in business might include:

- Older workers are more loyal to a company.
- Companies with more than \$1 billion in assets spend a higher percentage of their annual budget on advertising than do companies with less than \$1 billion in assets.
- The implementation of a Six Sigma quality approach in manufacturing will result in greater productivity.
- The price of scrap metal is a good indicator of the industrial production index six

months later.

- Airline company stock prices are positively correlated with the volume of OPEC oil production.

Virtually all inquisitive, thinking business people have similar research hypotheses concerning relationships, approaches, and techniques in business. Such hypotheses can lead decision makers to new and better ways to accomplish business goals. However, to formally test research hypotheses, it is generally best to state them as statistical hypotheses.

- **Statistical Hypotheses**

In order to scientifically test research hypotheses, a more formal hypothesis structure needs to be set up using **statistical hypotheses**. Suppose business researchers want to “prove” the research hypothesis that older workers are more loyal to a company. A “loyalty” survey instrument is either developed or obtained. If this instrument is administered to both older and younger workers, how much higher do older workers have to score on the “loyalty” instrument (assuming higher scores indicate more loyal) than younger workers to prove the research hypothesis? What is the “proof threshold”? Instead of attempting to prove or disprove research hypotheses directly in this manner, business researchers convert their research hypotheses to statistical hypotheses and then test the statistical hypotheses using standard procedures. All statistical hypotheses consist of two parts, a null hypothesis and an alternative hypothesis. These two parts are constructed to contain all possible outcomes of the experiment or study. Generally, the **null hypothesis** states that the “null” condition exists; that is, there is nothing new happening, the old theory is still true, the old standard is correct, and the system is in control. The **alternative hypothesis**, on the other hand, states that the new theory is true, there are new standards, the system is out of control, and/or something is happening. As an example, suppose flour packaged by a manufacturer is sold by weight; and a particular size of package is supposed to average 40 ounces. Suppose the manufacturer wants to test to determine whether their packaging process is out of control as determined by the weight of the flour packages. The null hypothesis for this experiment is that the average weight of the flour packages is 40 ounces (no problem).

The alternative hypothesis is that the average is not 40 ounces (process is out of control). It is common symbolism to represent the null hypothesis as H_0 and the alternative hypothesis as H_a .

- **Substantive Hypotheses**



In testing a statistical hypothesis, a business researcher reaches a conclusion based on the data obtained in the study. If the null hypothesis is rejected and therefore the alternative hypothesis is accepted, it is common to say that a statistically significant result has been obtained. For example, in the market share problem, if the null hypothesis is rejected, the result is that the market share is “significantly greater” than 18%. The word significant to statisticians and business researchers merely means that the result of the experiment is unlikely due to chance and a decision has been made to reject the null hypothesis. However, in everyday business life, the word significant is more likely to connote “important” or “a large amount.” One problem that can arise in testing statistical hypotheses is that particular characteristics of the data can result in a statistically significant outcome that is not a

significant business outcome. As an example, consider the market share study. Suppose a large sample of potential customers is taken, and a sample market share of 18.2% is obtained. Suppose further that a statistical analysis of these data results in statistical significance. We would conclude statistically that the market share is significantly higher than 18%. This finding actually means that it is unlikely that the difference between the sample proportion and the population proportion of .18 is due just to chance. However, to the business decision maker, a market share of 18.2% might not be significantly higher than 18%. Because of the way the word significant is used to denote rejection of the null hypothesis rather than an important business difference, business decision makers need to exercise caution in interpreting the outcomes of statistical tests.

Two-tailed tests always use = and in the statistical hypotheses and are directionless in that the alternative hypothesis allows for either the greater than () or less than () possibility. In this particular example, if the process is “out of control,” plant officials might not know whether machines are overfilling or under filling packages and are interested in testing for either possibility.

$H_0: p = .18$

$H_a: p > .18$

One-tailed tests are always directional, and the alternative hypothesis uses either the greater than (>) or the less than (<) sign. A one-tailed test should only be used when the researcher knows for certain that the outcome of an experiment is going to occur only in one direction or the researcher is only interested in one direction of the experiment as in the case of the market share problem. In one-tailed problems, the researcher is trying to “prove” that something is older, younger, higher, lower, more, less, greater, and so on. These words are considered “directional” words in that they indicate the direction of the focus of the research. Without these words, the alternative hypothesis of a one-tailed test cannot be established.

Process of Testing Hypothesis

In conducting business research, the process of testing hypotheses involves four major tasks:

- Task 1. Establishing the hypotheses
- Task 2. Conducting the test
- Task 3. Taking statistical action
- Task 4. Determining the business implications

Typically, statisticians and researchers present the hypothesis testing process in terms of an eight-step approach:

- Step 1. Establish a null and alternative hypothesis.
- Step 2. Determine the appropriate statistical test.
- Step 3. Set the value of alpha, the Type I error rate.
- Step 4. Establish the decision rule.
- Step 5. Gather sample data.
- Step 6. Analyze the data.
- Step 7. Reach a statistical conclusion.
- Step 8. Make a business decision.

Rejection and Nonrejection Regions

Using the critical values established at step 4 of the hypothesis testing process, the possible statistical outcomes of a study can be divided into two groups:

1. Those that cause the rejection of the null hypothesis
2. Those that do not cause the rejection of the null hypothesis.

Conceptually and graphically, statistical outcomes that result in the rejection of the null hypothesis lie in what is termed the **rejection region**. Statistical outcomes that fail to result in the rejection of the null hypothesis lie in what is termed the **nonrejection region**.

Type I and Type II Errors

Because the hypothesis testing process uses sample statistics calculated from random data to reach conclusions about population parameters, it is possible to make an incorrect decision about the null hypothesis. In particular, two types of errors can be made in testing hypotheses: Type I error and Type II error.

A **Type I error** is committed by rejecting a true null hypothesis. With a Type I error, the null hypothesis is true, but the business researcher decides that it is not. As an example, suppose the flour-packaging process actually is “in control” and is averaging 40 ounces of flour per package. Suppose also that a business researcher randomly selects 100 packages, weighs the contents of each, and computes a sample mean. It is possible, by chance, to randomly select 100 of the more extreme packages (mostly heavy weighted or mostly light weighted) resulting in a mean that falls in the rejection region. The decision is to reject the null hypothesis even though the population mean is actually 40 ounces. In this case, the business researcher has committed a Type I error.

the rejection regions represent the possibility of committing a Type I error. Means that fall beyond the critical values will be considered so extreme that the business researcher chooses to reject the null hypothesis. However, if the null hypothesis is true, any mean that falls in a rejection region will result in a decision that produces a Type I error. The probability of committing a Type I error is called **alpha (α)** or **level of significance**. Alpha equals the area under the curve that is in the rejection region beyond the critical value(s). The value of alpha is always set before the experiment or study is undertaken. As mentioned previously, common values of alpha are .05, .01, .10, and .001.

A **Type II error** is committed when a business researcher fails to reject a false null

hypothesis. In this case, the null hypothesis is false, but a decision is made to not reject it. Suppose in the case of the flour problem that the packaging process is actually producing a population mean of 41 ounces even though the null hypothesis is 40 ounces. A sample of 100 packages yields a sample mean of 40.2 ounces, which falls in the nonrejection region. The business decision maker decides not to reject the null hypothesis. A Type II error has been committed. The packaging procedure is out of control and the hypothesis testing process does not identify it.

The probability of committing a Type II error is **beta (β)**. Unlike alpha, beta is not usually stated at the beginning of the hypothesis testing procedure. Actually, because beta occurs only when the null hypothesis is not true, the computation of beta varies with the many possible alternative parameters that might occur. For example, in the flour-packaging problem, if the

population mean is not 40 ounces, then what is it? It could be 41, 38, or 42 ounces. A value of beta is associated with each of these alternative means.

		<i>State of nature</i>	
		Null true	Null false
<i>Action</i>	Fail to reject null	Correct decision	Type II error (β)
	Reject null	Type I error (α)	Correct decision (power)

Sr. No.	Question	Answer
1	The formula to calculate standardized normal random variable is	$x - \mu / \sigma$
2	In random experiment, the observations of random variable are classified as	trials
3	In binomial distribution, the formula of calculating standard deviation is	square root of npq
4	The mean of binomial probability distribution is 857.6 and the probability is 64% then the number of values of binomial distribution	1340
5	The tail or head, the one or zero and the girl and boy are examples of	complementary events
6	In binomial probability distribution, the success and failure generated by the trial is respectively denoted by	p and q
7	By taking a level of significance of 5% it is the same as saying	We are 95% confident that the results have not occurred by chance
8	One or two tail test will determine	If the region of rejection is located in one or two tails of the distribution

9	Two types of errors associated with hypothesis testing are Type I and Type II. Type II error is committed when	We accept a null hypothesis when it is not true
10	Type 1 error occurs when?	We reject H ₀ if it is True
11	The probability of Type 1 error is referred as?	α
12	Alternative Hypothesis is also called as?	Research Hypothesis
13	Which of the following is defined as the rule or formula to test a Null Hypothesis?	Test statistic
14	If the Critical region is evenly distributed then the test is referred as?	Two tailed
15	The point where the Null Hypothesis gets rejected is called as?	Critical Value
16	The rejection probability of Null Hypothesis when it is true is called as?	Level of Significance
17	A hypothesis which defines the population distribution is called?	Simple Hypothesis
18	A statement whose validity is tested on the basis of a sample is called?	Statistical Hypothesis
19	If the assumed hypothesis is tested for rejection considering it to be true is called?	Null Hypothesis
20	A statement made about a population for testing purpose is called?	Hypothesis
21is a variable that contains the outcomes of a chance experiment.	random variable
22	A random variable is aif the set of all possible values is at most a finite or a countably infinite number of possible values.	discrete random variable
23random variables take on values at every point over a given interval.	Continuous
24	How many types of distributions?	Two
25	How many types continuous distributions have?	Six
26	How many types of discrete distribution have?	3

27	Theorof a discrete distribution is the long-run average of occurrences.	Mean & expected value
28	The experiment involves n identical trials it is assumption of ?	Binomial distribution
29	Binomial distribution trial has only how many possible outcomes ?	Two
30	P means ?	probability of getting a success
31	q means?	probability of getting a failure
32	q means?	(1- P)
33	In a binomial experiment, the trials must be ?	Independent
34	n = ?	sample size
35	N = ?	population size
36	Ahas an expected value or a long-run average?	binomial distribution
37	A binomial distribution has an expected value or a long-run average, which is denoted by ?	μ
38	Poisson distribution is named after	Simeon-Denis Poisson (1781–1840)
39	Simeon-Denis Poisson (1781–1840) is ?	a French mathematician
40	who published its essentials in a paper in 1837.	Simeon-Denis Poisson
41focuses only on the number of discrete occurrences over some interval or continuum	Poisson distribution
42does not have a given number of trials (n)	Poisson

		experiment
43	The Poisson distribution describes the occurrence of which events?	Rare
44	It describes discrete occurrences over a continuum or interval.	Poisson distribution
45	The occurrences in each interval can range from zero to infinity	Poisson distribution
46	Number of sewing flaws per pair of jeans during production is example of ?	Poisson distribution
47	This average is denoted ?	Lambda (λ)
48	Value of e =?	e = 2.718282
49	Statisticians often use theto complement the types of analyses that can be made by using the binomial distribution.	hypergeometric distribution
50	The hypergeometric distribution applies only to experiments in which the trials are done ?	without replacement
51	The population, N, is finite and known is characteristics of ?	hypergeometric distribution
52	the uniform distribution, sometimes referred to as the?	rectangular distribution
53	height, weight, length, speed, IQ, scholastic achievement, and years of life expectancy are example of ?	Normal distribution
54	Discovery of the normal curve of errors is generally credited to ?	astronomer Karl Gauss (1777–1855)
55	astronomer Karl Gauss (1777–1855) was a?	Mathematician

56 recognized that the errors of repeated measurement of objects are often normally distributed.	astronomer Karl Gauss
57	Who determined that the binomial distribution approached the normal distribution as a limit?	De Moivre
58	It is a symmetrical distribution about its mean is characteristics of what ?	Normal distribution
59	2.71828 is value of ?	π
60is the number of standard deviations that a value, x, is above or below the mean	z score
61	The z distribution is a normal distribution with a mean ofand a standard deviation of?	0 & 1
62	theis continuous and describes a probability distribution of the times between random occurrences.	exponential distribution
63	■ It is skewed to the right is characteristics of what ?	exponential distribution
64are tentative explanations of a principle operating in nature?	Hypotheses
65is is a statement of what the researcher believes will be the outcome of an experiment or a study	research hypotheses
66	If the null hypothesis is rejected and therefore the alternative hypothesis is	Accepted
67	Hypothesis testing have how many steps ?	8
68	Conceptually and graphically, statistical outcomes that result in the rejection of the null hypothesis lie in what is termed the?	rejection region
69	Statistical outcomes that fail to result in the rejection of the null hypothesis lie in what is termed the?	nonrejection region
70	How many types of errors can be made in testing	Two

	hypotheses?	
71	A which type of error is committed by rejecting a true null hypothesis.	Type I
72	Type I error is called ?	Alpha
73	Alpha is known as ?	level of significance
74	A.....error is committed when a business researcher fails to reject a false null hypothesis	Type II
75	a Type II error is known ?	Beta
76	The p-value defines the smallest value of alpha for which the null hypothesis can be?	Rejected
77	A statement made about a population for testing purpose is called?	Hypothesis
78	$1-\alpha$ is the probability of	<ul style="list-style-type: none"> • Acceptance Region
79	Analysis of Variance (ANOVA) is a test for equality of	Means
80	The critical value of a test statistic is determined from	The sampling distribution of the statistics assuming Null Hypothesis
81	If we reject the null hypothesis, we might be making	Type-I Error
82	The t distributions are	Symmetrical
83	A Type I error occurs when we:	reject a true null hypothesis



84	In a criminal trial, a Type I error is made when:	an innocent person is convicted (sent to jail)
85	A Type II error occurs when we:	do not reject a false null hypothesis
86	In a criminal trial, a Type II error is made when:	a guilty defendant is acquitted (set free)
87	If we reject the null hypothesis, we conclude that:	There is enough statistical evidence to infer that the alternative hypothesis is true
88	The area under the curve represents the sum of probabilities for all possible outcomes .. This statement is	True
89	_____ Is a symmetrical distribution	Normal
90	_____ is a bell-shaped distribution	Normal
91	A multiple-choice test has 30 questions. There are 4 choices for each question. A student who has not studied for the test decides to answer all the questions randomly by guessing the answer to each question. Which of the following probability distributions can be used to calculate the student's chance of getting at least 20 questions right?	Binomial distribution
92	The rejection probability of Null Hypothesis when it is true is called as?	Level of Significance
93	The point where the Null Hypothesis gets rejected is called as?	Critical Value
94	If the Critical region is evenly distributed then the test is referred as?	Two tailed
95	Consider a hypothesis H_0 where $\phi_0 = 5$ against H_1 where	Right tailed

	$\phi_1 > 5$. The test is?	
96	Type 1 error occurs when?	We reject H_0 if it is True
97	The probability of Type 1 error is referred as?	α
98	Alternative Hypothesis is also called as?	Research Hypothesis
99	Which of the following mentioned standard Probability density functions is applicable to discrete Random Variables?	Poisson Distribution
100	A variable that can assume any value between two given points is called _____	Continuous random variable
101	Standard deviation of a binomial distribution	$\sigma = \sqrt{n \cdot p \cdot q}$
102	Mean of a binomial distribution	$n \cdot p$
103	The _____ <i>focuses only on the number of discrete occurrences over some interval or continuum.</i>	Poisson distribution
104	_____ describes rare events.	Poisson distribution

Z-Test

A z-test is a statistical test used to determine whether two population means are different when the variances are known and the sample size is large. The test statistic is assumed to have a normal distribution, and nuisance parameters such as standard deviation should be known in order for an accurate z-test to be performed.

Z-test is a statistical test to determine whether two population means are different when the variances are known and the sample size is large.

Z-test is a hypothesis test in which the z-statistic follows a normal distribution.

A z-statistic, or z-score, is a number representing the result from the z-test.

Z-tests are closely related to t-tests, but t-tests are best performed when an experiment has a small sample size.

Z-tests assume the standard deviation is known, while t-tests assume it is unknown.

Applications of Z test

Z-test is performed in studies where the sample size is larger, and the variance is known.

It is also used to determine if there is a significant difference between the mean of two independent samples.

The z-test can also be used to compare the population proportion to an assumed proportion or to determine the difference between the population proportion of two samples.

Formula

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

\bar{x} = sample mean

μ = population mean

σ = standard deviation of population

n = number of observation

Differences

Z-Test	T-Test
Data size is greater than 30	Data size is smaller than 30
Data points isn't related or doesn't affect another data point. For example, in our example, both the control and the experimental group were independent from each other	Data points might be related to each other, where the behaviour or value of one point may affect another
Data is normally distributed (if data size > 30 we can assume it is)	Data is not normally distributed
Data was randomly selected from a broader population	Data wasn't randomly selected
Sample sizes should be equal if possible	Sample sizes are not equally

z-test for the difference in mean:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

where \bar{x}_1 and \bar{x}_2 are the means of two samples, σ is the standard deviation of the samples, and n_1 and n_2 are the numbers of observations of two samples.

Question- 1

A random sample of size 35 is taken with sample mean of 29.5 and population mean of 32 and a population standard deviation of 9. The data is normally distributed and level of significance is 5%.

Step-1: $H_0: \mu=32$

$H_a: \mu \neq 32$

Step-2: $\alpha=0.05$

Step-3

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$
$$= \frac{29.5 - 32}{9 / \sqrt{35}}$$

= -1.64

Step-4 Table value $p=1.645$

Step-5 Accept the null hypothesis

Question- 2

In population the average IQ is 3.25 with standard deviation of 2.61. A sample of 900 students who has a mean of 3.4 is taken to test the medication. Did the medication affect the IQ?

Step-1: $H_0: \mu=3.25$

$H_a: \mu \neq 3.25$

Step-2: $\alpha=0.05$

Step-3

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$
$$= \frac{3.4 - 3.25}{2.61 / \sqrt{900}}$$

=1.73

Step-4 Table value $p=1.645$

Step-5 Reject the null hypothesis

Question- 3

In population the average IQ is 74914 with standard deviation of 14530. A sample of 112 students who has a mean of 78695 is taken to test the medication. Did the medication affect the IQ?

(Answer 2.75)

Question- 4

In population the average is 80 with standard deviation of 20. A sample of 50 students who has a mean of 100 is taken to test the effect.

(Answer 7.07)

Z-test for difference in mean of 2 groups

Question-1

A sample of 87 professional working women showed that the average amount to them is 3352. The population standard deviation is 1100. A sample of 76 professional working men showed that the average amount to them is 5727. The population standard deviation is 1700. $\alpha=1\%$

Step-1: $H_0: \mu_1 - \mu_2 = 0$

$H_a: \mu_1 - \mu_2 \neq 0$

Step-2: $\alpha=0.01$

Step-3

$$z = \frac{(3352 - 5727) - (0)}{\sqrt{\frac{1100^2}{87} + \frac{1700^2}{76}}} = \frac{-2375}{227.9} = -10.42$$

Step-4 Table value $p=2.326$

Step-5 Accept the null hypotheses

Question-2

A sample of 32 boys showed that the average amount to them is 70.700. The population variance is 264.160. A sample of 34 girls showed that the average amount to them is 62.187. The population variance is 166.410.

Step-1: $H_0: \mu_1 - \mu_2 = 0$

$H_a: \mu_1 - \mu_2 \neq 0$

Step-2: $\alpha=0.05$

Step-3

$$z = \frac{(70.700 - 62.187) - (0)}{\sqrt{\frac{264.160}{32} + \frac{166.410}{34}}} = 2.35$$

Step-4 Table value $p=1.645$

Step-5 Reject the null hypotheses

Question-3

Use the following information to construct the difference between 2 population means.

Group – 1	Group – 2
$N_1 = 32$	$N_2 = 31$
$\bar{x}_1 = 70.4$	$\bar{x}_2 = 68.7$
$\sigma_1 = 5.76$	$\sigma_2 = 6.1$

T-Test

Introduction

A t-test is used to determine if there is a significant difference between the means of two groups, which may be related in certain features. It is mostly used when the data sets, like the data set recorded as the outcome from flipping a coin 100 times, would follow a normal distribution and may have unknown variances

A t-test looks at the t-statistic, the t-distribution values, and the degrees of freedom to determine the statistical significance. To conduct a test with three or more means, one must use an analysis of variance.

Essentially, a t-test allows us to compare the average values of the two datasets and determine if they came from the same population. In the above examples, if we were to take a sample of students from class A and another sample of students from class B, we would not expect them to have exactly the same mean and standard deviation.

Formula

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

$$s / \sqrt{n}$$

\bar{x} = sample mean

μ = population mean

s = standard deviation

n = number of observations

Question- 1

A random sample of size 20 is taken with sample mean of 16.45 and population mean of 16 and a sample standard deviation of 3.59. The data is normally distributed and level of significance is 5%.

Step-1: Ho: $\mu=16$ Ha: $\mu \neq 16$

Step-2: $\alpha=0.05$

Step-3 df= $n - 1 = 20 - 1 = 19$

Step-4:

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

$$s / \sqrt{n}$$

$$= \frac{16.45 - 16}{3.56 / \sqrt{20}}$$

$$3.56 / \sqrt{20}$$

$$= 0.56$$

Step-5 Table value $p=2.093$

Step-6 Reject the null hypothesis

Note: If in question more than or less than is given then it is one tail test so when looking at the table you have to divide by 2.

Question- 2

A random sample of size 20 is taken with sample mean of 25.51 population mean of 25 and a sample standard deviation of 2.193. Level of significance is 5%. Compare whether the data is less than the required?

Step-1: $H_0: \mu=20$ $H_a: \mu<20$

Step-2: $\alpha=0.05$

Step-3 $df= n - 1= 20 - 1 =19$

Step-4:

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

$$s / \sqrt{n}$$

$$= \frac{25.51 - 20}{2.193 / \sqrt{20}}$$

$$= 1.04$$

Step-5 Table value $p=2.093$

Step-6 Accept the null hypothesis

Question- 3

A new process for producing diamond can be operated at profitable level only. If the average weight of diamond is greater than 0.5 carat. To evaluate the process, six diamonds are generated with weights of 0.46, 0.61, 0.52, 0.48, 0.57 and 0.54. Does measurement present sufficient evidence to indicate that the average weight of the diamond excess 0.5 carat?

Step-1: $H_0: \mu=0.5$ $H_a: \mu>0.5$

Step-2: $\alpha=0.05$

Step-3 $df= n - 1= 6 - 1 =5$

Step-4

Calculate $\bar{x} = \frac{\sum x}{N} = \frac{0.46+0.61+ 0.52+0.48+ 0.57 + 0.54}{6} = 0.53$

N

6

Calculate standard deviation

	$(x - \bar{x})$	$(x - \bar{x})^2$
0.46	-0.07	0.0049
0.61	0.08	0.0064
0.52	-0.01	0.0001
0.48	-0.05	0.0025
0.57	0.04	0.0016
0.54	0.01	0.0001
		0.0156

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

$$= \sqrt{0.0156/5}$$

$$= 0.05585$$

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

$$s / \sqrt{n}$$

$$= \frac{0.53 - 0.5}{0.05585 / \sqrt{6}}$$

$$0.05585 / \sqrt{6}$$

$$= 1.32$$

Step-5 Table value $p=2.015$

Step-6 Accept the null hypothesis.

Question- 5

The following data were gathered from a random sample of 11 items.

1200 1175 1080 1275 1201 1387 1280 1400 1287 1225 1090

Use these data at 5% level of significance to test the hypotheses, assuming the data to be normally distributed using t test.

Paired t -test

What is the paired t -test?

The paired t -test is a method used to test whether the mean difference between pairs of measurements is zero or not.

When to use this test?

You can use the test when your data values are paired measurements. For example, you might have before-and-after measurements for a group of people. Also, the distribution of differences between the paired measurements should be normally distributed.

What are some other names for the paired t -test?

The paired t -test is also known as the dependent samples t -test, the paired-difference t -test, the matched pairs t -test and the repeated-samples t -test.

What if my data isn't nearly normally distributed?

If your sample sizes are very small, you might not be able to test for normality. You might need to rely on your understanding of the data. Or, you can perform a *nonparametric* test that doesn't assume normality.

Assumption of paired t -test

- Subjects must be independent. Measurements for one subject do not affect measurements for any other subject.
- Each of the paired measurements must be obtained from the same subject. For example, the before-and-after weight for a smoker in the example above must be from the same person.

- The measured differences are normally distributed.

Formula

$$t = \frac{\bar{d} - D}{s / \sqrt{n}}$$

\bar{d} = sample difference between the

pair

D = mean population difference

s = standard deviation

n = number of observations

To find standard deviation

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

Question-1

Suppose a stock market investor is interested in determining whether there is significant difference in P/E ratio for companies from one year to next. Data are given below. Assume $\alpha=0.01$

Company	1	2	3	4	5	6	7	8	9
P/E ratio (Year 1)	8.9	38.1	43.0	34.0	34.5	15.2	20.3	19.9	61.9
P/E ratio (Year 2)	12.7	45.4	10.0	27.2	22.8	24.1	32.3	40.1	106.5

Step-1: Ho: D=0

Ha: D≠0

Step-2: α=0.01

Step-3 df= n – 1= 9 – 1 =8

Step-4

Company	P/E ratio (Year 1)	P/E ratio (Year 2)	d	d ²
1	8.9	12.7	-3.8	14.44
2	38.1	45.4	-7.3	53.29
3	43.0	10.0	33.0	1089
4	34.0	27.2	6.8	46.24
5	34.5	22.8	11.7	136.89
6	15.2	24.1	-8.9	79.21
7	20.3	32.3	-12.0	144
8	19.9	40.1	-20.2	408.04
9	61.9	106.5	-44.6	1989.16
			-45.3	3960.27

$$\bar{d} = \sum d/n = -45.3/9 = -5.033$$

$$\sqrt{\frac{g d^2 - \frac{(g d)^2}{n}}{n - 1}}$$

$$s = \sqrt{3960.27 - (-45.3)^2 / 9}$$

$$9 - 1$$

$$= 21.599$$

$$t = \frac{\bar{d} - D}{s / \sqrt{n}}$$

$$s / \sqrt{n}$$

$$= \frac{-5.033 - 0}{21.599 / \sqrt{9}}$$

$$21.599 / \sqrt{9}$$

$$= -0.70$$

Step-5 Table value $p=3.355$

Step-6 Accept the null hypothesis.

Question-2

A company ask individual to rate video before and after the break. Data are given below. Assume $\alpha=0.05$ to determine whether there is significant increase in the rating or not?

Individual	1	2	3	4	5	6	7
Before	32	11	21	17	30	38	14

After	39	15	35	13	41	39	22
-------	----	----	----	----	----	----	----

Step-1: Ho: D=0

Ha: D>0

Step-2: $\alpha=0.05$

Step-3 $df= n - 1= 7 - 1 =6$

Step-4

Individual	Before	After	d	d ²
1	32	39	-7	49
2	11	15	-4	16
3	21	35	-14	196
4	17	13	4	16
5	30	41	-11	121
6	38	39	-1	1
7	14	22	-8	64
			-41	463

$$\bar{d} = \sum d/n = -41/7 = -5.857$$

$$\sqrt{\frac{g d^2 - \frac{(g d)^2}{n}}{n - 1}}$$

$$s = \sqrt{463 - \frac{(-41)^2}{7}}$$

$$7 - 1$$

$$= 6.0945$$

$$t = \frac{\bar{d} - D}{s / \sqrt{n}}$$

$$= \frac{-41-0}{6.0945 / \sqrt{7}}$$

$$= -2.54$$

Step-5 Table value $p=1.943$

Step-6 Reject the null hypothesis.

Question-3

Use the given data and a 1% level of significance to test the following hypotheses. Assume the differences are normally distributed in the population.

$H_0: D=0$ $H_a: D>0$

Pair	1	2	3	4	5	6	7	8	9
Sample 1	38	27	30	41	36	38	33	35	44
Sample 2	22	28	21	38	38	26	19	31	35

Independent t -test

The Independent Samples T Test compares the means of two independent groups in order to determine whether there is statistical evidence that the associated population means are significantly different. The Independent Samples t Test is a parametric test.

The Independent Samples t Test is commonly used to test the following:

- Statistical differences between the means of two groups
- Statistical differences between the means of two interventions
- Statistical differences between the means of two change scores

Note: The Independent Samples t Test can only compare the means for two (and only two) groups. It cannot make comparisons among more than two groups. If you wish to compare the means across more than two groups, you will likely want to run an ANOVA.

Hypotheses

The null hypothesis (H_0) and alternative hypothesis (H_1) of the Independent Samples t Test can be expressed in two different but equivalent ways:

$H_0: \mu_1 = \mu_2$ ("the two-population means are equal")

$H_a: \mu_1 \neq \mu_2$ ("the two-population means are not equal")

OR

$H_0: \mu_1 - \mu_2 = 0$ ("the difference between the two-population means is equal to 0")

$H_a: \mu_1 - \mu_2 \neq 0$ ("the difference between the two-population means is not 0")

where μ_1 and μ_2 are the population means for group 1 and group 2, respectively. Notice that the second set of hypotheses can be derived from the first set by simply subtracting μ_2 from both sides of the equation.

When the two independent samples are assumed to be drawn from populations with identical population variances (i.e., $\sigma^2 = \sigma_1^2$), the test statistic t is computed as:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

When variances are not identical (it can be used in both identical as well non identical.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$
$$df = n_1 + n_2 - 2$$

To find standard deviation

$$\hat{\sigma} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

\bar{x}_1 = Mean of first sample

\bar{x}_2 = Mean of second sample

n_1 = Sample size (i.e., number of observations) of first sample

n_2 = Sample size (i.e., number of observations) of second sample

s_1 = Standard deviation of first sample

s_2 = Standard deviation of second sample

$$df = n_1 + n_2 - 2$$

Example

The concentration of cholesterol (a type of fat) in the blood is associated with the risk of developing heart disease, such that higher concentrations of cholesterol indicate a higher level of risk, and lower concentrations indicate a lower level of risk. If you lower the concentration of cholesterol in the blood, your risk of developing heart disease can be reduced. Being overweight and/or physically inactive increases the concentration of cholesterol in your blood. Both exercise and weight loss can reduce cholesterol concentration. However, it is not known whether exercise or weight loss is best for lowering cholesterol concentration. Therefore, a researcher decided to investigate whether an exercise or weight loss intervention is more effective in lowering cholesterol levels. To this end, the researcher recruited a random sample of inactive males that were classified as overweight. This sample was then randomly split into two groups: Group 1 underwent a calorie-controlled diet and Group 2 undertook the exercise-training programme. In order to determine which treatment programme was more effective, the mean cholesterol concentrations were compared between the two groups at the end of the treatment programmes.

Question-1

To test the difference between 2 methods, the manager selected randomly the employees and the data are given below.

Method A	Method B
$N_1 = 15$	$N_2 = 12$
$\bar{x}_1 = 47.73$	$\bar{x}_2 = 56.5$
$S_1^2 = 19.495$	$S_2^2 = 18.273$

Step-1: $H_0: \mu_1 - \mu_2 = 0$

$H_a: \mu_1 - \mu_2 \neq 0$

Step-2: $\alpha=0.05$

Step-3 $df= n_1 + n_2 - 2$

$$= 15 + 12 - 2 = 25$$

Step-4

$$t = \frac{(47.73 - 56.50) - (0)}{\sqrt{\frac{(19.495)(14) + (18.273)(11)}{(15 + 12 - 2)} \left(\frac{1}{15} + \frac{1}{12} \right)}} = -5.20$$

Step-5 Table value $p = 2.060$

Step-6 Reject the null hypotheses (ignore the sign of -)

Question-2

To test the difference between 2 company, the manager selected randomly the buyers and the data are given below. Level of significance 1%.

Taiwan buyer	Chinese buyer
N1 = 46	N2 = 26
$\bar{x}_1 = 5.42$	$\bar{x}_2 = 5.04$
S1= 0.58	S2 =0.49
S1² = 0.3364	S2² = 0.2401

Step-1: $H_0: \mu_1 - \mu_2 = 0$ $H_a: \mu_1 - \mu_2 \neq 0$

Step-2: $\alpha=0.01$

Step-3 $df= n_1 + n_2 - 2$

$$= 46 + 26 - 2$$

$$= 70$$

Step-4

$$t = \frac{(5.42 - 5.04) - (0)}{\sqrt{\frac{(.3364)(45) + (.2401)(25)}{46 + 26 - 2}} \sqrt{\frac{1}{46} + \frac{1}{26}}} = 2.8$$

Step-5 Table value $p = 2.648$

Step-6 Reject the null hypotheses (ignore the sign of -)

Question-3

Suppose a company is interested in comparing the prices of products. A small survey was conducted through phone. A random sample of 21 person resulting in a sample average price of 11690 with a standard deviation of 230. Another random sample of 26 person resulting in a sample average price of 11400 with a standard deviation of 175. Assume data to be normally distributed. Level of significance 10%.

Question-4

Suppose a company is interested in comparing the prices of products. A small survey was conducted through phone. A random sample of 10

person resulting in a sample average price of 900 with a standard deviation of 23. Another random sample of 13 person resulting in a sample average price of 490 with a standard deviation of 15. Assume data to be normally distributed.

Regression Analysis

Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables. It can be utilized to assess the strength of the relationship between variables and for modelling the future relationship between them.

In linear regression the variable to be predicted is called the dependent variable and is denoted by y . The predictor is called independent variable/ explanatory variable and is denoted by x . In linear regression straight linerelationship between two variables are examined.

Regression Analysis – Linear Model Assumptions

Linear regression analysis is based on six fundamental assumptions:

1. The dependent and independent variables show a linear relationship between the slope and the intercept (expected value of y).
2. The independent variable is not random.
3. The value of the residual (error) is zero.
4. The value of the residual (error) is constant across all observations.
5. The value of the residual (error) is not correlated across all observations.
6. The residual (error) values follow the normal distribution.

Regression Analysis – Simple Linear Regression

Simple linear regression is a model that assesses the relationship between a dependent variable and an independent variable. The simplelinear model is expressed using the following equation:

$$Y = a + bX + \epsilon$$

Where:

- **Y** – Dependent variable
- **X** – Independent (explanatory) variable
- **a** – Intercept
- **b** – Slope

- **ϵ** – Residual (error)

Least square analysis is a process whereby a regression model is developed by producing the minimum sum of the square error values.

Following are the formulas for regression analysis

$$SS_{xx} = \sum(x - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$SS_{xy} = \sum(x - \bar{x})(y - \bar{y}) = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

$$b_0 = \bar{y} - b_1\bar{x} = \frac{\sum y}{n} - b_1 \frac{(\sum x)}{n}$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}}$$

Regression line(y) = b₀ + b₁x

Question-1

The data contains the information regarding number of passengers and cost. Compute the regression analysis.

X	61	63	67	69	70	74	76	81	86	91	95	97
Y	4.28	4.08	4.42	4.17	4.48	4.3	4.82	4.7	5.11	5.13	5.64	5.56

Calculation

X	Y(in 000)	X ²	Xy
61	4280	3721	261.080
63	4080	3969	257.040
67	4420	4489	296.140
69	4170	4761	287.730
70	4480	4900	313.600
74	4300	5476	318.200
76	4820	5776	366.320
81	4700	6561	380.700
86	5110	7396	439.460
91	5130	8281	466.830

95	5640	9025	535.800
97	5560	9409	539.320
$\Sigma x=930$	$\Sigma y=56690$	$\Sigma x^2=73764$	$\Sigma xy=4462220$

$$SS_{xy} = \frac{\Sigma xy}{n} - \frac{\Sigma x \Sigma y}{n^2}$$

$$= \frac{4462220}{12} - \frac{(930)(56690)}{12^2}$$

$$= 68745$$

$$SS_{xx} = \frac{\Sigma x^2}{n} - \frac{(\Sigma x)^2}{n^2}$$

$$= \frac{73764}{12} - \frac{(930)^2}{12^2}$$

$$= 1689$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}}$$

$$= \frac{68745}{1689}$$

$$= 0.0407$$

$$b_0 = \frac{\Sigma y}{n} - b_1 \frac{\Sigma x}{n}$$

$$= \frac{56690}{12} - 0.0407 \frac{930}{12}$$

$$= \frac{56.69}{12} - (0.0407) \frac{930}{12}$$

$$= 1.57$$

$$\text{Regression line}(y) = b_0 + b_1x$$

$$= 1.57 + 0.0407x$$

Question-2

In hospital number of beds and number of employees survey was undertaken. The data are given below:

x(Beds)	y(Hospital)	x ²	xy
23	69	529	1587
29	95	841	2755
29	102	841	2958
35	118	1225	4130
42	126	1764	5292
46	125	2116	5750
50	138	2500	6900
54	178	2916	9612
64	156	4096	9984
66	184	4356	12144
76	176	5776	13376
78	225	6084	17550
Σx= 592	Σy= 1692	Σx²= 33044	Σxy= 92038

$$S_{xy} = \frac{\sum xy - \sum x \sum y}{n}$$

n

$$= 92038 - \frac{(592)(1692)}{12}$$

12

$$= 8566$$

$$SS_{xx} = \frac{\sum x^2 - (\sum x)^2}{n}$$

n

$$= 33044 - \frac{(592)^2}{12}$$

12

$$= 3838.667$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}}$$

SS_{xx}

$$= \frac{8566}{3838.667}$$

$$= 2.232$$

$$b_0 = \frac{\sum y}{n} - b_1 \left(\frac{\sum x}{n} \right)$$

$$= \frac{1692}{12} - (2.232) \left(\frac{592}{12} \right)$$

$$= 30.888$$

$$\text{Regression line}(y) = b_0 + b_1x$$

$$= 30.888 + 2.232x$$

Question-3

A corporation owns several companies. In long term planning they gather data of sales and advertising from several companies and data is given below. Develop regression line.

Advertising (x)	Sales (y)
12.5	148
3.7	55
21.6	338
60.0	994
37.6	541
6.1	89
16.8	126
41.2	379

1.	The test to compare between 2 means?	T-test
2.	A t-test looks at the?	t-statistics
3.	A t-test allows us to compare the average values of the two data sets and determine if they came from the	same population.
4.	There is _____ in variance of t-test?	Homogeneous
5.	There is _____ shaped distribution in t-test?	Bell
6.	The <i>t</i> -test is a method used to test whether the mean difference between pairs of measurements is zero or not.	paired
7.	Before-after is in which test?	Paired
8.	The _____ Samples <i>T</i> Test compares the means of two independent groups in order to determine whether there is statistical evidence that the associated population means are significantly different.	Independent
9.	To compare the means across more than two groups, you will likely want to run an _____	ANOVA
10.	A _____ is a statistical test used to determine whether two population means are different when the variances are known and the sample size is large.	z-test
11.	Sample size is greater than 30 in which test?	z-test
12.	Sample size is less than 30 in which test?	t-test
13.	_____ is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables	Regression analysis
14.	In linear regression the variable to be predicted is called the dependent variable and is denoted by ?	y



15.	Independent Variable is denoted by?	X
16.	The value of residual is 0 in regression.	True
17.	The residual (error) values follow the normal distribution in regression.	True
18.	___ analysis is a process whereby a regression model is developed by producing the minimum sum of the square error values.	Least square
19.	In regression b stands for?	slope
20.	In regression ϵ stands for?	error

Analysis of Variance (ANOVA)

Introduction

Analysis of Variance (ANOVA) is a statistical method used to test differences between two or more means. It may seem odd that the technique is called "Analysis of Variance" rather than "Analysis of Means." As you will see, the name is appropriate because inferences about means are made by analysing variance.

ANOVA is used to test general rather than specific differences among means. This can be seen best by example.

Hypotheses of ANOVA

Ho: $\mu_1 = \mu_2 = \mu_3$

Ha: At least one of the means is different from others.

Assumption of ANOVA

- Each group sample is drawn from a normally distributed population.
 - All populations have a common variance.
 - All samples are drawn independently of each other.
- Within each sample, the observations are sampled randomly and independently of each other.

Steps to calculate ANOVA

Calculate mean of all groups.

Calculate mean of mean.

Calculate SSC

Calculate SSE

Calculate SST

Complete the table

Formula

$$SSC = \sum_{j=1}^C n_j (\bar{x}_j - \bar{x})^2$$

$$SSE = \sum_{i=1}^{n_j} \sum_{j=1}^C (x_{ij} - \bar{x}_j)^2$$

$$SST = \sum_{i=1}^{n_j} \sum_{j=1}^C (x_{ij} - \bar{x})^2$$

$$df_C = C - 1$$

$$df_E = N - C$$

$$df_T = N - 1$$

$$MSC = \frac{SSC}{df_C}$$

$$MSE = \frac{SSE}{df_E}$$

$$F = \frac{MSC}{MSE}$$

where

i = a particular member of a treatment level

j = a treatment level

C = number of treatment levels

n_j = number of observations in a given treatment level

\bar{x} = grand mean

\bar{x}_j = column mean

x_{ij} = individual value

The Calculations

variation within treatments, and we reject the null hypothesis of equal means. If F is small, the variability between treatments is small relative to the variation within treatments, and we do not reject the null hypothesis of equal means. (In this case, the sample data is consistent with the hypothesis that population means are equal between groups.) To compute this ratio (the F statistic) is difficult and time consuming. Therefore, we are always going to let the computer do this for us. The computer generates what is called an ANOVA table:

Source	SS	Df	MS	F
Treatment/Group	SSG	$k - 1$	$MSG = \frac{SSG}{k - 1}$	$\frac{MSG}{MSE}$
Error	SSE	$n - k$	$MSE = \frac{SSE}{n - k}$	
Total	SST	$n - 1$		

The *source* (of variability) column tells us SS=Sum of Squares (sum of squared deviations):

- SST measures variation of the data around the overall mean \bar{x}
- SSG measures variation of the group means around the overall mean
- SSE measures the variation of each observation around its group mean \bar{x}_i



Degrees of freedom

- $n - k$ for SSE, since it measures the variation of the n observations about k group means.
- $n - 1$ for SST, since it measures the variation of all n observations about the overall mean.
- $k-1$ for SSG, since it measures the variation of the k group in overall mean.

Note- For ANOVA we will look at table F.

A company has three manufacturing plants, and company officials want to determine whether there is a difference in the average age of workers at the three locations. The following data are the ages of five randomly selected workers at each plant. Perform a one-way ANOVA to determine whether there is a significant difference in the mean ages of the workers at the three plants. Use $\alpha = .01$ and note that the sample sizes are equal.

Solution

HYPOTHESIZE:

STEP 1. The hypotheses follow.

$$H_0: \mu_1 = \mu_2 = \mu_3$$

H_a : At least one of the means is different from the others.

TEST:

STEP 2. The appropriate test statistic is the F test calculated from ANOVA.

STEP 3. The value of α is .01.

STEP 4. The degrees of freedom for this problem are $3 - 1 = 2$ for the numerator and $15 - 3 = 12$ for the denominator. The critical F value is $F_{.01,2,12} = 6.93$.

Because ANOVAs are always one tailed with the rejection region in the upper tail, the decision rule is to reject the null hypothesis if the observed value of F is greater than 6.93.

STEP 5.

Plant (Employee Ages)

1	2	3
29	32	25
27	33	24
30	31	24
27	34	25
28	30	26

STEP 6.

$$T_j: T_1 = 141 \quad T_2 = 160 \quad T_3 = 124 \quad T = 425$$

$$n_j: n_1 = 5 \quad n_2 = 5 \quad n_3 = 5 \quad N = 15$$

$$\bar{x}_j: \bar{x}_1 = 28.2 \quad \bar{x}_2 = 32.0 \quad \bar{x}_3 = 24.8 \quad \bar{x} = 28.33$$

$$SSC = 5(28.2 - 28.33)^2 + 5(32.0 - 28.33)^2 + 5(24.8 - 28.33)^2 = 129.73$$

$$SSE = (29 - 28.2)^2 + (27 - 28.2)^2 + \dots + (25 - 24.8)^2 + (26 - 24.8)^2 = 19.60$$

$$SST = (29 - 28.33)^2 + (27 - 28.33)^2 + \dots + (25 - 28.33)^2 \\ + (26 - 28.33)^2 = 149.33$$

$$df_C = 3 - 1 = 2$$

$$df_E = 15 - 3 = 12$$

$$df_T = 15 - 1 = 14$$

Source of Variance	SS	df	MS	F
Between	129.73	2	64.87	39.80
Error	19.60	12	1.63	
Total	149.33	14		

ACTION:

STEP 7. The decision is to reject the null hypothesis because the observed F value of 39.80 is greater than the critical table F value of 6.93.

BUSINESS IMPLICATIONS:

STEP 8. There is a significant difference in the mean ages of workers at the three plants. This difference can have hiring implications. Company leaders should understand that because motivation, discipline, and experience may differ with age, the differences in ages may call for different managerial approaches in each plant.

The chart on the next page displays the dispersion of the ages of workers from the three samples, along with the mean age for each plant sample. Note the difference in group means. The significant F value says that the difference between the mean ages is relatively greater than the differences of ages within each group.

1.	_____ is a statistical method used to test differences between two or more means	ANOVA
2.	All populations have a common variance.	True
3.	All samples are drawn independently of each other.	True
4.	A _____ test is a statistical technique that analyses the effect of the independent variables on the expected outcome along with their relationship to the outcome itself.	Two-way ANOVA
5.	In two-way ANOVA there are 2 independent variables.	True
6.	A _____ is primarily designed to enable the equality testing between three or more means.	One-way ANOVA
7.	A _____ is an $n \times n$ array filled with n different symbols (letter A, B, C) each occurring exactly once in each row and exactly once in each column.	Latin Square
8.	Latin Square is denoted by A, B, C, D etc.	True
9.	Latin Square randomly permute the columns.	True
10.	Latin Square randomly permute the rows	True
11.	Latin Square assign the treatments to the Latin letters in a random fashion.	True
12.	An assumption that we make when using a Latin square design is that the three factors (treatments, and two nuisance factors) do not interact.	True
13.	Two-way ANOVA tests the interdependence between two independent variables.	True
14.	SST means Sum of Square of Total	True
15.	SSR means?	Sum of Square of Rows

Chi-Square Test

Meaning of Chi-Square Test:

The Chi-square (χ^2) test represents a useful method of comparing experimentally obtained results with those to be expected theoretically on some hypothesis.

Thus, Chi-square is a measure of actual divergence of the observed and expected frequencies. It is very obvious that the importance of such a measure would be very great in sampling studies where we have invariably to study the divergence between theory and fact.

Chi-square as we have seen is a measure of divergence between the expected and observed frequencies and as such if there is no difference between expected and observed frequencies the value of Chi-square is 0.

If there is a difference between the observed and the expected frequencies then the value of Chi-square would be more than 0. That is, the larger the Chi-square the greater the probability of a real divergence of experimentally observed from expected results.

If the calculated value of chi-square is very small as compared to its table value it indicates that the divergence between actual and expected frequencies is very little and consequently the fit is good. If, on the other hand, the calculated value of chi-square is very big as compared to its table value it indicates that the divergence between expected and observed frequencies is very great and consequently the fit is poor.

To evaluate Chi-square, we enter Table E with the computed value of chi-square and the appropriate number of degrees of freedom. The number of $df = (r - 1)(c - 1)$ in which r is the number of rows and c the number of columns in which the data are tabulated. :

Thus in 2×2 table degrees of freedom are $(2 - 1)(2 - 1)$ or 1. Similarly in 3×3 table, degrees of freedom is $(3 - 1)(3 - 1)$ or 4 and in 3×4 table the degrees of freedom are $(3 - 1)(4 - 1)$ or 6.

Levels of Significance of Chi-Square Test:

The calculated values of χ^2 (Chi-square) are compared with the table values, to conclude whether the difference between expected and observed frequencies is due to the sampling fluctuations and as such significant or whether the difference is due to some other reason and as such significant. The divergence of theory and fact is always tested in terms of certain probabilities

The probabilities indicate the extent of reliance that we can place on the conclusion drawn. The table values of χ^2 are available at various probability levels. These levels are called levels of

significance. Usually, the value of χ^2 at .05 and .01 level of significance for the given degrees of freedom is seen from the tables.

If the calculated value of χ^2 is greater than the tabulated value, it is said to be significant. In other words, the discrepancy between the observed and expected frequencies cannot be attributed to chance and we reject the null hypothesis.

Thus, we conclude that the experiment does not support the theory. On the other hand, if calculated value of χ^2 is less than the corresponding

tabulated value then it is said to be non- significant at the required level of significance.

This implies that the discrepancy between observed values (experiment) and the expected values(theory) may be attributed to chance, i.e., fluctuations of sampling.

Chi-Square Test under Null Hypothesis:

Suppose we are given a set of observed frequencies obtained under some experiment and we want to test if the experimental results support a particular hypothesis or theory. Karl Pearson in 1900, developed a test for testing the significance of the discrepancy between experimental values and the theoretical values obtained under some theory or hypothesis.

This test is known as χ^2 -test and is used to test if the deviation between observation (experiment) and theory may be attributed to chance (fluctuations of sampling) or if it is really due to the inadequacy of the theory to fit the observed data.

Under the Null Hypothesis we state that there is no significant difference between the observed (experimental) and the theoretical or hypothetical values, i.e., there is a good compatibility between theory and experiment.

The equation for chi-square (χ^2) is stated as follows:

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right]$$

in which f_o = frequency of occurrence of observed or experimentally determined facts : f_e = expected frequency of occurrence on some hypothesis.

Thus chi-square is the sum of the values obtained by dividing the square of the difference between observed and expected frequencies by the expected frequencies in each case. In other words the differences between observed and expected frequencies are squared and divided by the expected number in each case, and the sum of these quotients is χ^2 .

Several illustrations of the chi-square test will clarify the discussion given above. The differences of f_o and f_e are written always + ve.

Testing the divergence of observed results from those expected on the hypothesis of equal probability (null hypothesis):

Example 1:

Ninety-six subjects are asked to express their attitude towards the proposition "Should AIDS education be integrated in the curriculum of Higher secondary stage" by marking F (favourable), I (indifferent) or U (unfavourable).

It was observed that 48 marked 'F', 24 'I' and 24 'U':

Test whether the observed results diverge significantly from the results to be expected if there are no preferences in the group.

Test the hypothesis that "there is no difference between preferences in the group".
Interpret the findings.

Solution:

Following steps may be followed for the computation of χ^2 and drawing the conclusions:

Step 1:

Compute the expected frequencies (f_e) corresponding to the observed frequencies in each case under some theory or hypothesis.

In our example the theory is of equal probability (null hypothesis). In the second row the distribution of answers to be expected on the null hypothesis is selected equally.

	Favourable	Indifferent	Unfavourable	
Observed (f_o)	48	24	24	96
Expected (f_e)	32	32	32	96
$(f_o - f_e)$	16	8	8	
$(f_o - f_e)^2$	256	64	64	
$\frac{(f_o - f_e)^2}{f_e}$	8	2	2	

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right] = 12 \quad df = 2 \quad P \text{ is less than } .01$$

Step 2::

Compute the deviations ($f_o - f_e$) for each frequency. Each of these differences is squared and divided by its f_e (256/32, 64/32 and 64/32).

Step 3:

Add these values to compute:

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right]$$

$$\left(\frac{256}{32} + \frac{64}{32} + \frac{64}{32} \right) \text{ to give } \chi^2 = 8 + 2 + 2 = 12$$

From Table E
Tabulated χ^2 with 2 df at
.01 level = 9.21

Step 4:

The degrees of freedom in the table is calculated from the formula $df = (r - 1) (c - 1)$ to be $(3 - 1) (2 - 1)$ or 2.

Step 5:

Look up the calculated (critical) values of χ^2 for 2 df at certain level of significance, usually 5% or 1%.

With $df = 2$, the χ^2 value to be significant at .01 level is 9.21 (Table E). The obtained χ^2 value of $12 > 9.21$.

- i. Hence the marked divergence is significant.
- ii. The null hypothesis is rejected.
- iii. We conclude that our group really favours the proposition.

We reject the “equal answer” hypothesis and conclude that our group favours the proposition.

Example 2:

The number of automobile accidents per week in a certain community were as follows:

12, 8, 20, 2, 14, 10, 15, 6, 9, 4

Are these frequencies in agreement with the belief that accident conditions were the same during this 10-week period?

Solution:

Null Hypothesis—Set up the null hypothesis that the given frequencies (of number of accidents per week in a certain community) are consistent with the belief that the accident conditions were same during the 10-week period.

Since the total number of accidents over the 10 weeks are:

$$12 + 8 + 20 + 2 + 14 + 10 + 15 + 6 + 9 + 4 = 100.$$

Under the null hypothesis, these accidents should be uniformly distributed over the 10-week period and hence the expected number of accidents for each of the 10 weeks are $100/10 = 10$.

Computation of χ^2

Week	Observed No. of accidents (f_o)	Expected No. of accidents (f_e)	($f_o - f_e$)	($f_o - f_e$) ²	$\frac{(f_o - f_e)^2}{f_e}$
1.	12	10	2	4	0.4
2.	8	10	2	4	0.4
3.	20	10	10	100	10.0
4.	2	10	8	64	6.4
5.	14	10	4	16	1.6
6.	10	10	0	0	0.0
7.	15	10	5	25	2.5
8.	6	10	4	16	1.6
9.	9	10	1	1	0.1
10.	4	10	6	36	3.6
	100	100			26.6

$$\therefore \chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right] = 26.6$$

$df = 9$. P is less than .01

From Table E

Tabulated χ^2 with 9 df

at .05 level = 16.916

at .01 level = 21.666

Since calculated value of $\chi^2 = 26.6$ is greater than the tabulated value, 21.666. It is significant and the null hypothesis rejected at .01 level of significance. Hence we conclude that the accident conditions are certainly not uniform (same) over the 10-week period.

Testing the divergence of observed results from those expected on the hypothesis of anormal distribution:

The hypothesis, instead of being equally probable, may follow the normal distribution. An example illustrates how this hypothesis may be tested by chi-square.

Example 3:

Two hundred salesmen have been classified into three groups very good, satisfactory, and poor—by consensus of sales managers.

	Good	Satisfactory	Poor	Total
Observed (f_o)	76	96	28	200
Expected (f_e)	32	136	32	200
$(f_o - f_e)$	44	40	4	
$(f_o - f_e)^2$	1936	1600	16	
$\frac{(f_o - f_e)^2}{f_e}$	60.50	11.76	0.50	

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right] = 60.50 + 11.76 + .50 = 72.76$$

Does this distribution of rating differ significantly from that to be expected if sellingability is normally distributed in our population of salesmen?

We set up the hypothesis that selling ability is normally distributed. The normal curve extends from -3σ to $+3\sigma$. If the selling ability is normally distributed the base line can be divided into three equal segments, i.e.

Rating	σ range between	% in Table A	% of 200 or (f_o)
Good	$+3.00\sigma$ and $+1.00\sigma$	16%	32
Satisfactory	$+1.00\sigma$ and -1.00σ	68%	136
Poor	-1.00σ and -3.00σ	16%	32
		100%	200

($+1\sigma$ to $+3\sigma$), (-1σ to $+1\sigma$) and (-3σ to -1σ) representing good, satisfactory and poor salesmen respectively. By referring Table A we find that 16% of cases lie between $+1\sigma$ and $+3\sigma$, 68% in between -1σ and $+1\sigma$ and 16% in between -3σ and -1σ . In case of our problem 16% of 200 = 32 and 68% of 200 = 136. $df = 2$. P is less than .01. The calculated $\chi^2 = 72.76$

The calculated χ^2 of 72.76 > 9.21. Hence P is less than .01.

From Table E
Tabulated χ^2 for 2df at
.01 level = 9.21

\therefore The discrepancy between observed frequencies and expected frequencies is quite significant. On this ground the hypothesis of a normal distribution of selling ability in this group must be rejected. Hence we conclude that the distribution of ratings differ from that to be expected.

- Chi-square test when our expectations are

based on predetermined results:

- **Example 4:**
- **In an experiment on breeding of peas a researcher obtained the following data:**

The theory predicts the proportion of beans, in four groups A, B, C and D should be 9:3:3:1. In an experiment among 1,600 beans, the numbers in four groups were 882,313, 287 and 118. Does the experiment results support the genetic theory? (Test at .05 level).

Solution:

We set up the null hypothesis that there is no significant difference between the experimental values and the theory. In other words there is good correspondence between theory and experiment, i.e., the theory supports the experiment.

Category	Expected frequency (f_e)
A	$\frac{9}{16} \times 1600 = 900$
B	$\frac{3}{16} \times 1600 = 300$
C	$\frac{3}{16} \times 1600 = 300$
D	$\frac{1}{16} \times 1600 = 100$

$$9 + 3 + 3 + 1 = 16$$

Computation of χ^2

	A	B	C	D
Observed frequency f_o	882	313	287	118
Expected frequency f_e	900	300	300	100

$(f_o - f_e)$	18	13	13	18
$(f_o - f_e)^2$	324	169	169	324
$\frac{(f_o - f_e)^2}{f_e}$.360	.563	.563	3.240

$$\Sigma \left[\frac{(f_o - f_e)^2}{f_e} \right] = .360 + .563 + .563 + 3.240$$

$$= 4.726$$

$df = 3$ P is near about .20

The calculated $\chi^2 = 4.726$

From Table E
Tabulated χ^2 for 3df
at .05 level = 7.81

Since the calculated χ^2 value of $4.726 < 7.81$, it is not significant. Hence null hypothesis may be accepted at .05 level of significance and we may conclude that the experimental results support the genetic theory.

The Chi-square test when table entries are small:

When table entries are small and when table is 2 x 2 fold, i.e., $df = 1$, χ^2 is subject to considerable error unless a correction for continuity (called Yates' Correction) is made.

Example 5:

Forty rats were offered opportunity to choose between two routes. It was found that 13 chose lighted routes (i.e., routes with more illumination) and 27 chose dark routes.

- (i) Test the hypothesis that illumination makes no difference in the rats' preference for routes (Test at .05 level).
- (ii) Test whether the rats have a preference towards dark routes.

Solution:

If illumination makes no difference in preference for routes i.e., if H_0 be true, the proportionate preference would be 1/2 for each route (i.e., 20).

In our example we are to subtract .5 from each ($f_o - f_e$) difference for the following reason:

In 2×2 fold tables, especially when entries are small, the χ^2 curve is not continuous. Hence, the deviation of 27 from 20 must be written as 6.5 ($26.5 - 20$) instead of 7 ($27 - 20$), as 26.5 is the lower limit of 27 in a continuous series. In like manner the deviation of 13 from 20 must be taken from the upper limit of 13, namely, 13.5.

The data can be tabulated as follows:

	Dark routes	Lighted routes	Total
Observed frequencies f_o	27	13	40
Expected frequencies f_e	20	20	40

$(f_o - f_e)$	7	7
$[(f_o - f_e) - .5]$	6.5	6.5
$[(f_o - f_e) - .5]^2$	42.25	42.25
$\frac{[(f_o - f_e) - .5]^2}{f_e}$	2.11	2.11

$$\therefore \chi^2 = 2.11 + 2.11 = 4.22.$$

When the expected entries in 2×2 fold table are the same as in our problem the formula for chi-square may be written in a somewhat shorter form as follows:

$$\chi^2 = \frac{2[(f_o - f_e) - .5]^2}{f_e}$$

$$= \frac{2(6.5)^2}{20} = \frac{2 \times 42.25}{20} = 4.22$$

From Table E' ... (56)
The tabulated value of χ^2 for 1 df at .05 level = 3.841.

df = 1 P is .043 (by interpolation)

Calculated $\chi^2 = 4.22$

(i) The critical value of χ^2 at .05 level is 3.841. The obtained χ^2 of 4.22 is more than 3.841. Hence the null hypothesis is rejected at .05 level. Apparently light or dark is a factor in the rats' choice for routes.

(ii) In our example we have to make a one-tailed test. Entering table E we find that χ^2 of 4.22 has a P = .043 (by interpolation).

$\therefore P/2 = .0215$ or 2%. In other words there are 2 chances in 100 that such a divergence would occur.

Hence we mark the divergence to be significant at 02 level.

Therefore, we conclude that the rats have a preference for dark routes.

The Chi-square test of independence in contingency tables:

Sometimes we may encounter situations which require us to test whether there is any relationship (or association) between two variables or attributes. In other words χ^2 can be made when we wish to investigate the relationship between traits or attributes which can be classified into two or more categories.

For example, we may be required to test whether the eye-colour of father is associated with the eye-colour of sons, whether the socio-economic status of the family is associated with the preference of different brands of a commodity, whether the education of couple and family size are related, whether a particular vaccine has a controlling effect on a particular disease etc.

To make a test we prepare a contingency table and to calculate f_e (expected frequency) for each cell of the contingency table and then compute χ^2 by using formula:

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right]$$

Null hypothesis:

χ^2 is calculated with an assumption that the two attributes are independent of each other, i.e. there is no relationship between the two attributes.

The calculation of expected frequency of a cell is as follows:

$$f_e \text{ of a cell} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand total}}$$

Example 6:

In a certain sample of 2,000 families 1,400 families are consumers of tea where 1236 are Hindu families and 164 are non-Hindu.

And 600 families are not consumers of tea where 564 are Hindu families and 36 are non-Hindu. Use χ^2 – test and state whether there is any significant difference between consumption of tea among Hindu and non-Hindu families.

Solution:

The above data can be arranged in the form of a 2 x 2 contingency table as given below:

	Hindu	Non-Hindu	Total
Families consuming tea	(I) 1236	(II) 164	1400
Families not consuming tea	(III) 564	(IV) 36	600
Grand Total	1800	200	2000

We set up the null hypothesis (H_0) that the two attributes viz., 'consumption of tea' and the 'community' are independent. In other words, there is no significant difference between the consumption of tea among Hindu and non-Hindu families.

Calculation of (f_e) :

	Hindu	Non-Hindu	Total
Families consuming tea	(I) $\frac{1800 \times 1400}{2000}$ = 1260	(II) $\frac{200 \times 1400}{2000}$ = 140	1400
Families not consuming tea	(III) $\frac{1800 \times 600}{2000}$ = 540	(IV) $\frac{200 \times 600}{2000}$ = 60	600
Total	1800	200	2000

$$f_e \text{ of each cell} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand total}}$$

Calculation of χ^2

Cells	f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
I	1236	1260	24	576	0.4571
II	164	140	24	576	4.1143
III	564	540	24	576	1.0667
IV	36	60	24	576	9.6000

$(f_o - f_e)$ is written disregarding sign.

$$\chi^2 = 15.2381$$

$$df = (2 - 1) (2 - 1) = 1$$

P is less than .01

$$\text{Calculated } \chi^2 = 15.2381$$

From Table E

Tabulated value of χ^2 for 1 df at

.05 level = 3.841

.01 level = 6.635

Since the calculated value of χ^2 , viz., 15.24 is much greater than the tabulated value of χ^2 at .01 level of significance; the value of χ^2 is highly significant and null hypothesis is rejected.

Hence we conclude that the two communities (Hindu and Non-Hindus) differ significantly as regards the consumption of tea among them.

Example 7:

The table given below shows the data obtained during an epidemic of cholera.

	Attacked	Non Attacked	Total
Inoculated	31	469	500
Not Inoculated	185	1315	1500
Total	216	1784	2000

Test the effectiveness of inoculation in preventing the attack of cholera.

Solution:

We set up the null hypothesis (H_0) that the two attributes viz., inoculation and absence of attack from cholera are not associated. These two attributes in the given table are independent.

	Attacked	Non Attacked	Total
Inoculated	(I) 31	(II) 469	500
Not Inoculated	(III) 185	(IV) 1315	1500
Total	216	1784	2000

Basing on our hypothesis we can calculate the expected frequencies as follows: Calculation of (f_e):

	Attacked	Not Attacked	Total
Inoculated	(I) $\frac{500 \times 216}{2000} = 54$	(II) $\frac{500 \times 1784}{2000} = 446$	500
Not Inoculated	(III) $\frac{1500 \times 216}{2000} = 162$	(IV) $\frac{1500 \times 1784}{2000} = 1338$	1500
Total			2000

$$f_e \text{ of each cell} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

Calculation of χ^2

Cells	f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
I	31	54	23	529	9.796
II	469	446	23	529	1.186
III	185	162	23	529	3.265
IV	1315	1338	23	529	0.395

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right] = 9.796 + 1.186 + 3.265 + 0.395 = 14.642$$

$$df = (2 - 1)(2 - 1) = 1. P \text{ is less than } .01$$

$$\text{Calculated } \chi^2 = 14.64$$

From Table E

Tabulated value of χ for 1
df at .05 level = 3.841

The five percent value of χ^2 for 1 df is 3.841, which is much less than the calculated value of χ^2 . So in the light of this, conclusion is evident that the hypothesis is incorrect and inoculation and absence of attack from cholera are associated.

Conditions for the Validity of Chi-Square Test:

The Chi-square test statistic can be used if the following conditions are satisfied:

1. N, the total frequency, should be reasonably large, say greater than 50.
2. The sample observations should be independent. This implies that no individual item should be included twice or more in the sample.
3. The constraints on the cell frequencies, if any, should be linear (i.e., they should not involve square and higher powers of the frequencies) such as $\sum f_o = \sum f_e = N$.
4. No theoretical frequency should be small. Small is a relative term. Preferably each theoretical frequency should be larger than 10 but, in any case, not less than

If any theoretical frequency is less than 5 then we cannot apply χ^2 -test as such. In that case we use the technique of "pooling" which consists in adding the frequencies which are less than 5 with the

preceding or succeeding frequency (frequencies) so that the resulting sum is greater than 5 and adjust for the degrees of freedom accordingly.

5. The given distribution should not be replaced by relative frequencies or proportions but the data should be given in original units.

6. Yates' correction should be applied in special circumstances when $df = 1$ (i.e. in 2×2 tables) and when the cell entries are small.

7. χ^2 -test is mostly used as a non-directional test (i.e. we make a two-tailed test.). However, there may be cases when χ^2 tests can be employed in making a one-tailed test.

In one-tailed test we double the P-value. For example, with $df = 1$, the critical value of χ^2 at 05 level is 2.706 (2.706 is the value written under .10 level) and the critical value of χ^2 at .01 level is 5.412 (the value is written under the .02 level).

The Additive Property of Chi-Square Test:

χ^2 has a very useful property of addition. If a number of sample studies have been conducted in the same field, then the results can be pooled together for obtaining an accurate idea about the real position.

Suppose ten experiments have been conducted to test whether a particular vaccine is effective against a particular disease. Now here we shall have ten different values of χ^2 and ten different values of df .

We can add the ten χ^2 to obtain one value and similarly ten values of df can also be added together. Thus, we shall have one value of χ^2 and one value of degrees of freedom. Now we can test the results of all these ten experiments combined together and find out the value of P.

Suppose five independent experiments have been conducted in a particular field. Suppose in each case there was one df and following values of χ^2 were obtained.

Experiment Number	Value of χ^2	df
1.	4.3	1
2.	5.7	1
3.	2.1	1
4.	3.9	1
5.	8.3	1

Now at 5% level of

significance (or for P –

.05) the value χ^2 for one df is 3.841. From the calculated values of χ^2 given above we notice that in only one case i.e., experiment No. 3 the observed value of χ^2 is less than the tabulated value of 3.841.

It means that so far as this experiment is concerned the difference is insignificant but in the remaining four cases the calculated value of χ^2 is more than 3.841 and as such at 5% level of significance the difference between the expected and the actual frequencies is significant.

If we add all the values of χ^2 we get $(4.3 + 5.7 + 2.1 + 3.9 + 8.3)$ or 24.3. The total of the degrees of freedom is 5. It means that the calculated value of χ^2 for 5 df is 24.3.

If we look in the table of χ^2 we shall find that at 5% level of significance for 5 df the value of χ^2 is 11.070. The calculated value of χ^2 which is 24.3 is much higher than the tabulated value and as such we can conclude that the difference between observed and expected frequencies is significant one.

Even if we take 1% level of significance (or $P = .01$) the table value of χ^2 is only 15.086. Thus the probability of getting a value of χ^2 equal to or more than 24.3 as a result of sampling fluctuations is much less than even .01 or in other words the difference is significant.

Applications of Chi-Test:

The applications of χ^2 -test statistic can be discussed as stated below:

1. Testing the divergence of observed results from expected results when our expectations are based on the hypothesis of equal probability.
2. Chi-square test when expectations are based on normal distribution.
3. Chi-square test when our expectations are based on predetermined results.
4. Correction for discontinuity or Yates' correction in calculating χ^2 .
5. Chi-square test of independence in

contingency tables.

Uses of Chi-Square Test

Although test is conducted in terms of frequencies it can be best viewed conceptually as a test about proportions

1. χ^2 test is used in testing hypothesis and is not useful for estimation.
2. Chi-square test can be applied to complex contingency table with several classes.
3. Chi-square test has a very useful property i.e., 'the additive property'. If a number of sample studies are conducted in the same field, the results can be pooled together. This means that χ^2 - values can be added.

1	The test represents a useful method of comparing experimentally obtained results with those to be expected theoretically on some hypothesis.	Chi-square(χ^2)
2	Chi-square is non directional.	True
3	Chi-square test has a very useful property i.e., 'the additive property'.	True
4	The _____ Chi-square the greater the probability of a real divergence of experimentally observed from expected results.	larger
5	χ^2 test is used in testing hypothesis and is not useful for estimation.	True
6	N, the total frequency, should be reasonably large, say greater than?	50
7	Under the _____ we state that there is no significant difference between the observed (experimental) and the theoretical or hypothetical values	Null Hypothesis
8	F _o stands for?	Frequency observed
9	F _e stands for?	Frequency expected
10	In chi square the sample observations should be independent.	True
11	Chi-square test when expectations are based on normal distribution.	True

12	Is the formula correct? $f_e \text{ of a cell} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand total}}$	True
13	$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right]$	True
14	If there is no difference between expected and observed frequencies the value of Chi-square is?	0
15	If any theoretical frequency is less than ___ then we cannot apply χ^2 -test as such.	5

Module 4

Mann-Whitney *U* test

The **Mann-Whitney *U* test** is a *nonparametric counterpart of the *t* test used to compare the means of two independent populations*. This test was developed by Henry B. Mann and D. R. Whitney in 1947. Recall that the *t* test for independent samples presented in Chapter 10 can be used when data are at least interval in measurement and the populations are normally distributed. However, if the assumption of a normally distributed population is invalid or if the data are only ordinal in measurement, the *t* test should not be used. In such cases, the Mann-Whitney *U* test is an acceptable option for analyzing the data. The following assumptions underlie the use of the Mann-Whitney *U* test.

1. The samples are independent.
2. The level of data is at least ordinal.

The two-tailed hypotheses being tested with the Mann-Whitney *U* test are as follows.

H₀: The two populations are identical.

H_a: The two populations are not identical.

Computation of the *U* test begins by arbitrarily designating two samples as group 1

and group 2. The data from the two groups are combined into one group, with each data value retaining a group identifier of its original group. The pooled values are then ranked from 1 to *n*, with the smallest value being assigned a rank of 1. The sum of the ranks of values from group 1 is computed and designated as *W*₁ and the sum of the ranks of values from group 2 is designated as *W*₂.

The Mann-Whitney U test is implemented differently for small samples than for large samples. If both $n_1, n_2 \leq 10$, the samples are considered small. If either n_1 or n_2 is greater than 10, the samples are considered large.

Small-Sample Case

With small samples, the next step is to calculate a U statistic for W_1 and for W_2 as

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - W_1 \quad \text{and} \quad U_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - W_2$$

The test statistic is the smallest of these two U values. Both values do not need to be calculated; instead, one value of U can be calculated and the other can be found by using the transformation

$$U' = n_1 \cdot n_2 - U$$

Is there a difference between health service workers and educational service workers in the amount of compensation employers pay them per hour? Suppose a random sample of seven health service workers is taken along with a random sample of eight educational service workers from different parts of the country. Each of their employers is interviewed and figures are obtained on the amount paid per hour for employee compensation for these workers. The data on the following page indicate total compensation per hour. Use a Mann-Whitney U test to determine whether these two populations are different in employee compensation.

Health Service Worker	Educational Service Worker
\$20.10	\$26.19
19.80	23.88
22.36	25.50
18.75	21.64
21.90	24.85
22.96	25.30
20.75	24.12
	23.45

Solution

HYPOTHESIZE:

STEP 1. The hypotheses are as follows.

H₀: The health service population is identical to the educational service population on employee compensation.

H_a: The health service population is not identical to the educational service population on employee compensation.

TEST:

STEP 2. Because we cannot be certain the populations are normally distributed, we chose a nonparametric alternative to the t test for independent populations: the small-sample Mann-Whitney U test.

STEP 3. Let alpha be .05.

STEP 4. If the final p-value from Table A.13 (after doubling for a two-tailed test here) is less than .05, the decision is to reject the null hypothesis.

STEP 5. The sample data were already provided.

STEP 6. We combine scores from the two groups and rank them from smallest to largest while retaining group identifier information.

Total Employee Compensation	Rank	Group
\$18.75	1	H
19.80	2	H
20.10	3	H
20.75	4	H
21.64	5	E
21.90	6	H
22.36	7	H
22.96	8	H
23.45	9	E
23.88	10	E
24.12	11	E
24.85	12	E
25.30	13	E
25.50	14	E
26.19	15	E

$$W_1 = 1 + 2 + 3 + 4 + 6 + 7 + 8 = 31$$

$$W_2 = 5 + 9 + 10 + 11 + 12 + 13 + 14 + 15 = 89$$

$$U_1 = (7)(8) + \frac{(7)(8)}{2} - 31 = 53$$

$$U_2 = (7)(8) + \frac{(8)(9)}{2} - 89 = 3$$

Because U_2 is the smaller value of U , we use $U = 3$ as the test statistic for Table A.13. Because it is the smallest size, let $n_1 = 7$; $n_2 = 8$.

ACTION:

STEP 7. Table A.13 yields a p-value of .0011. Because this test is two tailed, we double the table p-value, producing a final p-value of .0022. Because the p-value is less than $\alpha = .05$, the null hypothesis is rejected. The statistical conclusion is that the populations are not identical.

BUSINESS IMPLICATIONS:

STEP 8. An examination of the total compensation figures from the samples indicates that employers pay educational service workers more per hour than they pay health service workers.

Large-Sample Case

For large sample sizes, the value of U is approximately normally distributed. Using an average expected U value for groups of this size and a standard deviation of U 's allows computation of a z score for the U value. The probability of yielding a z score of this magnitude, given no difference between the groups, is computed. A decision is then made whether to reject the null hypothesis. A z score can be calculated from U by the following formulas.

LARGE-SAMPLE FORMULAS
MANN-WHITNEY U TEST
(17.1)

$$\mu_U = \frac{n_1 \cdot n_2}{2}, \quad \sigma_U = \sqrt{\frac{n_1 \cdot n_2 (n_1 + n_2 + 1)}{12}}, \quad z = \frac{U - \mu_U}{\sigma_U}$$

For example, the Mann-Whitney U test can be used to determine whether there is a difference in the average income of families who view PBS television and families who do not view PBS television. Suppose a sample of 14 families that have identified themselves as PBS television viewers and a sample of 13 families that have identified themselves as non-PBS television viewers are selected randomly.

HYPOTHESIZE:

STEP 1. The hypotheses for this example are as follows.

H_0 : The incomes of PBS and non-PBS viewers are identical.

H_a : The incomes of PBS and non-PBS viewers are not identical.

TEST:

STEP 2. Use the Mann-Whitney U test for large

samples.

STEP 3. Let $\alpha = .05$.

STEP 4. Because this test is two-tailed with $\alpha = .05$, the critical values are $z_{0.025} = 1.96$.

If the test statistic is greater than 1.96 or less than -1.96, the decision is to reject the null hypothesis.

STEP 5. The average annual reported income for each family in the two samples is given in Table 17.1.

STEP 6. The first step toward computing a Mann-Whitney U test is to combine these two columns of data into one group and rank the data from lowest to highest, while maintaining the identification of each original group. Table 17.2 shows the results of this step.

Note that in the case of a tie, the ranks associated with the tie are averaged across the values that tie. For example, two incomes of \$43,500 appear in the sample. These incomes represent ranks 19 and 20. Each value therefore is awarded a ranking of 19.5, or the average of 19 and 20.

If PBS viewers are designated as group 1, W_1 can be computed by summing the ranks of all the incomes of PBS viewers in the sample.

$$W_1 = 4 + 7 + 11 + 12 + 13 + 14 + 18 + 19.5 + 22 + 23 + 24 + 25 + 26 + 27 = 245.5$$

Then W_1 is used to compute the U value. Because $n_1 = 14$ and $n_2 = 13$, then

Then W_1 is used to compute the U value. Because $n_1 = 14$ and $n_2 = 13$, then

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - W_1 = (14)(13) + \frac{(14)(15)}{2} - 245.5 = 41.5$$

Because $n_1, n_2 > 10$, U is approximately normally distributed, with a mean of

$$\mu_U = \frac{n_1 \cdot n_2}{2} = \frac{(14)(13)}{2} = 91$$

and a standard deviation of

$$\sigma_U = \sqrt{\frac{n_1 \cdot n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{(14)(13)(28)}{12}} = 20.6$$

A z value now can be computed to determine the probability of the sample U value coming from the distribution with $\mu_U = 91$ and $\sigma_U = 20.6$ if there is no difference in the populations.

$$z = \frac{U - \mu_U}{\sigma_U} = \frac{41.5 - 91}{20.6} = \frac{-49.5}{20.6} = -2.40$$

ACTION:

STEP 7. The observed value of z is -2.40, which is less than so the results are in the rejection region. That is, there is a difference between the income of a PBS viewer and that of a non-PBS viewer. Examination of the sample data confirms that in general, the income of a PBS viewer is higher than that of a non-PBS viewer.

Income	Rank	Group	Income	Rank	Group
\$16,000	1	Non-PBS	39,500	15	Non-PBS
21,000	2	Non-PBS	40,500	16	Non-PBS
21,500	3	Non-PBS	41,000	17	Non-PBS
24,500	4	PBS	43,000	18	PBS
27,600	5	Non-PBS	43,500	19.5	PBS
27,800	6	Non-PBS	43,500	19.5	Non-PBS
32,000	7	PBS	51,900	21	Non-PBS
32,400	8	Non-PBS	53,000	22	PBS
32,500	9	Non-PBS	55,000	23	PBS
33,000	10	Non-PBS	57,960	24	PBS
34,000	11	PBS	61,000	25	PBS
36,800	12	PBS	61,400	26	PBS
39,000	13	PBS	62,500	27	PBS
39,400	14	PBS			

BUSINESS IMPLICATIONS:

STEP 8. The fact that PBS viewers have higher average income can affect the type of programming on PBS in terms of both trying to please present viewers and offering programs that might attract viewers of other income levels. In addition, fund-raising drives can be made to appeal to the viewers with higher incomes.

Illustration :-

Do construction workers who purchase lunch from street vendors spend less per meal than construction workers who go to restaurants for lunch? To test this question, a researcher selects two random samples of construction workers, one group that purchases lunch from street vendors and one group that purchases lunch from restaurants. Workers are asked to record how much they spend on lunch that day. The data follow. Use the data and a Mann-Whitney U test to analyze the data to determine whether street-vendor lunches are significantly cheaper than restaurant lunches. Let $\alpha = .01$.

Vendor	Restaurant
\$2.75	\$4.10
3.29	4.75
4.53	3.95
3.61	3.50
3.10	4.25
4.29	4.98
2.25	5.75
2.97	4.10
4.01	2.70
3.68	3.65
3.15	5.11
2.97	4.80
4.05	6.25
3.60	3.89
	4.80
	5.50
$n_1 = 14$	$n_2 = 16$

Solution

HYPOTHESIZE:

STEP 1. The hypotheses follow.

H₀: The populations of construction-worker spending for lunch at vendors and restaurants are the same.

H_a: The population of construction-worker spending at vendors is shifted to the left of the population of construction-worker spending at restaurants.

TEST:

STEP 2. The large-sample Mann-Whitney U test is appropriate. The test statistic is the z.

STEP 3. Alpha is .01.

STEP 4. If the p-value of the sample statistic is less than .01, the decision is to reject the null hypothesis.

STEP 5. The sample data are given.

STEP 6. Determine the value of W₁ by combining the groups, while retaining group identification and ranking all the values from 1 to 30 (14 + 16), with 1 representing the smallest value.

Value	Rank	Group	Value	Rank	Group
\$2.25	1	V	\$4.01	16	V
2.70	2	R	4.05	17	V
2.75	3	V	4.10	18.5	R
2.97	4.5	V	4.10	18.5	R
2.97	4.5	V	4.25	20	R
3.10	6	V	4.29	21	V
3.15	7	V	4.53	22	V
3.29	8	V	4.75	23	R
3.50	9	R	4.80	24.5	R
3.60	10	V	4.80	24.5	R
3.61	11	V	4.98	26	R
3.65	12	R	5.11	27	R
3.68	13	V	5.50	28	R
3.89	14	R	5.75	29	R
3.95	15	R	6.25	30	R

Summing the ranks for the vendor sample gives Solving for U, U, and U yields Solving for the observed z value gives

$$W1 = 1 + 3 + 4.5 + 4.5 + 6 + 7 + 8 + 10 + 11 + 13 + 16 + 17 + 21 + 22 = 144$$

Solving for U , μ_U , and σ_U yields

$$U = (14)(16) + \frac{(14)(15)}{2} - 144 = 185 \quad \mu_U = \frac{(14)(16)}{2} = 112$$

$$\sigma_U = \sqrt{\frac{(14)(16)(31)}{12}} = 24.1$$

Solving for the observed z value gives

$$z = \frac{185 - 112}{24.1} = 3.03$$

ACTION:

STEP 7. The p-value associated with $z = 3.03$ is .0012. The null hypothesis is rejected.

BUSINESS IMPLICATIONS:

STEP 8. The business researcher concludes that construction-worker spending at vendors is less than the spending at restaurants for lunches.

Wilcoxon matched-pairs signed rank test

The Mann-Whitney U test presented in Section 17.2 is a nonparametric alternative to the t test for two *independent* samples. If the two samples are *related*, the U test is not applicable. A test that does handle related data is the **Wilcoxon matched-pairs signed rank test**, which serves as a *nonparametric alternative to the t test for two related samples*. Developed by Frank Wilcoxon in 1945, the Wilcoxon test, like the t test for two related samples, is used to analyze several different types of studies when the data of one group are related to the data in the other group, including before-and-after studies, studies in which measures are taken on the same person or object under two different conditions, and studies of twins or other relatives.

The Wilcoxon test utilizes the differences of the scores of the two matched groups in a manner similar to that of the t test for two related samples. After the difference scores

have been computed, the Wilcoxon test ranks all differences regardless of whether the difference is positive or negative. The values are ranked from smallest to largest, with a rank of 1 assigned to the smallest difference. If a difference is negative, the rank is given a negative sign. The sum of the positive ranks is tallied along with the sum of the negative ranks. Zero differences representing ties between scores from the two groups are ignored, and the value of n is reduced accordingly. When ties occur between ranks, the ranks are averaged over the values. The smallest sum of ranks (either + or -) is used in the analysis and is represented by T . The Wilcoxon matched-pairs signed rank test procedure for determining statistical significance differs with sample size. When the number of matched pairs, n , is greater than 15, the value of T is approximately normally distributed and a z score is computed to test the null hypothesis. When sample size is small, $n \leq 15$, a different procedure is followed.

Two assumptions underlie the use of this technique.

1. The paired data are selected randomly.
2. The underlying distributions are symmetrical.

The following hypotheses are being tested.

For two-tailed tests:

$$H_0: M_d = 0 \quad H_a: M_d \neq 0$$

For one-tailed tests:

$$H_0: M_d = 0 \quad H_a: M_d > 0$$

or

$$H_0: M_d = 0 \quad H_a: M_d < 0$$

where M_d is the median.

Small-Sample Case (n 15)

When sample size is small, a critical value against which to compare T can be found in Table A.14 to determine whether the null hypothesis should be rejected. The critical value is located by using n and α . Critical values are given in the table for $\alpha = .05, .025, .01, \text{ and } .005$ for two-tailed tests and $\alpha = .10, .05, .02, \text{ and } .01$ for one-tailed tests. If the observed value of T is less than or equal to the critical value of T , the decision is to reject the null hypothesis.

As an example, consider the survey by American Demographics that estimated the average annual household spending on healthcare. The U.S. metropolitan average was \$1,800. Suppose six families in Pittsburgh, Pennsylvania, are matched demographically with six families in Oakland, California, and their amounts of household spending on

healthcare for last year are obtained. The data follow on the next page.

Family Pair	Pittsburgh	Oakland
1	\$1,950	\$1,760
2	1,840	1,870
3	2,015	1,810
4	1,580	1,660
5	1,790	1,340
6	1,925	1,765

A healthcare analyst uses $\alpha = .05$ to test to determine whether there is a significant difference in annual household healthcare spending between these two cities.

HYPOTHESIZE:

STEP 1. The following hypotheses are being tested.

$$H_0: M_d = 0$$

$$H_a: M_d \neq 0$$

TEST:

STEP 2. Because the sample size of pairs is six, the small-sample Wilcoxon matched pairs signed ranks test is appropriate if the underlying distributions are assumed to be symmetrical.

STEP 3. Alpha is .05.

STEP 4. From Table A.14, if the observed value of T is less than or equal to 1, the decision is to reject the null hypothesis.

STEP 5. The sample data were listed earlier.

STEP 6.

Family Pair	Pittsburgh	Oakland	d	Rank
1	\$1,950	\$1,760	+190	+4
2	1,840	1,870	-30	-1
3	2,015	1,810	+205	+5
4	1,580	1,660	-80	-2
5	1,790	1,340	+450	+6
6	1,925	1,765	+160	+3

$$T = \text{minimum of } (T_+, T_-)$$

$$T_+ = 4 + 5 + 6 + 3 = 18$$

$$T_- = 1 + 2 = 3$$

$$T = \text{minimum of } (18, 3) = 3$$

ACTION:

STEP 7. Because $T = 3$ is greater than critical $T = 1$, the decision is not to reject the null hypothesis.

BUSINESS IMPLICATIONS:

STEP 8. Not enough evidence is provided to declare that Pittsburgh and Oakland differ in annual household spending on healthcare. This information may be useful to healthcare providers and employers in the two cities and particularly to businesses that either operate in both cities or are planning to move from one to the other. Rates can be established on the notion that healthcare costs are about the same in both cities. In addition, employees considering transfers from one city to the other can expect their annual healthcare costs to remain about the same.

Large-Sample Case ($n > 15$)

For large samples, the T statistic is approximately normally distributed and a z score can be used as the test statistic. Formula 17.2 contains the necessary formulas to complete this procedure.

WILCOXON MATCHED-
PAIRS SIGNED RANK TEST
(17.2)

$$\mu_T = \frac{(n)(n+1)}{4}$$
$$\sigma_T = \sqrt{\frac{(n)(n+1)(2n+1)}{24}}$$
$$z = \frac{T - \mu_T}{\sigma_T}$$

where

n = number of pairs

T = total ranks for either + or - differences, whichever is less in magnitude

This technique can be applied to the airline industry, where an analyst might want to determine whether there is a difference in the cost per mile of airfares in the United States between 1979 and 2009 for various cities. The data in Table 17.3 represent the costs per mile of airline tickets for a sample of 17 cities for both 1979 and 2009.

HYPOTHESIZE:

STEP 1. The analyst states the hypotheses as follows.

$$H_0: M_d = 0$$

$$H_a: M_d \neq 0$$

TEST:

STEP 2. The analyst applies a Wilcoxon matched-pairs signed rank test to the data to test the difference in cents per mile for the two periods of time. She assumes the underlying distributions are symmetrical.

STEP 3. Use $\alpha = .05$.

City	1979	2009	<i>d</i>	Rank
1	20.3	22.8	-2.5	-8
2	19.5	12.7	+6.8	+17
3	18.6	14.1	+4.5	+13
4	20.9	16.1	+4.8	+15
5	19.9	25.2	-5.3	-16
6	18.6	20.2	-1.6	-4
7	19.6	14.9	+4.7	+14
8	23.2	21.3	+1.9	+6.5
9	21.8	18.7	+3.1	+10
10	20.3	20.9	-0.6	-1
11	19.2	22.6	-3.4	-11.5
12	19.5	16.9	+2.6	+9
13	18.7	20.6	-1.9	-6.5
14	17.7	18.5	-0.8	-2
15	21.6	23.4	-1.8	-5
16	22.4	21.3	+1.1	+3
17	20.8	17.4	+3.4	+11.5

STEP 4. Because this test is two-tailed, $\alpha = .025$ and the critical values are $z = 1.96$.

If the observed value of the test statistic is greater than 1.96 or less than -1.96, the null hypothesis is rejected.

STEP 5. The sample data are given in Table 17.3.

STEP 6. The analyst begins the process by computing a difference score, d . Which year's data are subtracted from the other does not matter as long as consistency in direction is maintained. For the data in Table 17.3, the analyst subtracted the 2009 figures from the 1979 figures. The sign of the difference is left on the difference score. Next, she ranks the differences without regard to sign, but the sign is left on the rank as an identifier. Note the tie for ranks 6 and 7; each is given a rank of 6.5, the average of the two ranks. The same applies to ranks 11 and 12.

After the analyst ranks all difference values regardless of sign, she sums the positive ranks (T_1) and the negative ranks (T_2). She then determines the T value from these two sums as the smallest T_1 or T_2 .

$$T = \text{minimum of } (T_+, T_-)$$

$$T_+ = 17 + 13 + 15 + 14 + 6.5 + 10 + 9 + 3 + 11.5 = 99$$

$$T_- = 8 + 16 + 4 + 1 + 11.5 + 6.5 + 2 + 5 = 54$$

$$T = \text{minimum of } (99, 54) = 54$$

The T value is normally distributed for large sample sizes, with a mean and standard deviation of

$$\mu_T = \frac{(n)(n+1)}{4} = \frac{(17)(18)}{4} = 76.5$$

$$\sigma_T = \sqrt{\frac{(n)(n+1)(2n+1)}{24}} = \sqrt{\frac{(17)(18)(35)}{24}} = 21.1$$

The observed z value is

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{54 - 76.5}{21.1} = -1.07$$

ACTION:

STEP 7. The critical z value for this two-tailed test is $z_{.025} = 1.96$. The observed $z = -1.07$, so the analyst fails to reject the null hypothesis. There is no significant difference in the cost of airline tickets between 1979 and 2009.

BUSINESS IMPLICATIONS:

STEP 8. Promoters in the airline industry can use this type of information (the fact that ticket prices have not increased significantly in 30 years) to sell their product as a good buy. In addition, industry managers could use it as an argument for raising prices.

Illustration :-

During the 1980s and 1990s, U.S. businesses increasingly emphasized quality control. One of the arguments in favor of quality-control programs is that quality control can increase productivity. Suppose a company implemented a quality-control program and has been operating under it for 2 years. The company's president wants to determine whether worker productivity significantly increased since installation of the program. Company records contain the figures for items produced per worker during a sample of production runs 2 years ago. Productivity figures on the same workers are gathered now and compared to the previous figures. The following data represent items produced per hour. The company's statistical analyst uses the Wilcoxon matched-pairs signed rank test to determine whether there is a significant increase in per worker production for $\alpha = .01$.

<u>Worker</u>	<u>Before</u>	<u>d</u>
1	5	
2	4	
3	9	
4	6	
5	3	
6	8	
7	7	
8	10	
9	3	
10	7	

Solution

HYPOTHESIZE:

STEP 1. The hypotheses are as follows.

$$H_0: M_d = 0$$

$$H_a: M_d < 0$$

TEST:

STEP 2. The analyst applies a Wilcoxon matched-pairs signed rank test to the data to test the difference in productivity from before to after. He assumes the underlying distributions are symmetrical.

STEP 3. Use $\alpha = .01$.

STEP 4. This test is one tailed. The critical value is $z = -2.33$. If the observed value of the test statistic is less than -2.33 , the null hypothesis is rejected.

STEP 5. The sample data are as already given.

STEP 6. The analyst computes the difference values, and, because zero differences are to be eliminated, deletes worker 3 from the study. This reduces n from 20 to 19. He then ranks the differences regardless of sign. The differences that are the same (ties) receive the average rank for those values. For example, the differences for workers 4, 5, 7, 10, and 14 are the same. The ranks for these five are 7, 8, 9, 10, and 11, so each worker receives the rank of 9, the average of these five ranks.

Worker	Before	After	d	Rank
1	5	11	-6	-19
2	4	9	-5	-17
3	9	9	0	delete
4	6	8	-2	-9
5	3	5	-2	-9
6	8	7	+1	+3.5
7	7	9	-2	-9
8	10	9	+1	+3.5
9	3	7	-4	-14.5
10	7	9	-2	-9
11	2	6	-4	-14.5
12	5	10	-5	-17
13	4	9	-5	-17
14	5	7	-2	-9
15	8	9	-1	-3.5
16	7	6	+1	+3.5
17	9	10	-1	-3.5
18	5	8	-3	-12.5
19	4	5	-1	-3.5
20	3	6	-3	-12.5

The analyst determines the values of T_+ , T_- , and T to be

$$T_+ = 3.5 + 3.5 + 3.5 = 10.5$$

$$T_- = 19 + 17 + 9 + 9 + 9 + 14.5 + 9 + 14.5 + 17 + 17 \\ + 9 + 3.5 + 3.5 + 12.5 + 3.5 + 12.5 = 179.5$$

$$T = \text{minimum of } (10.5, 179.5) = 10.5$$

The mean and standard deviation of T are

$$\mu_T = \frac{(n)(n+1)}{4} = \frac{(19)(20)}{4} = 95$$

$$\sigma_T = \sqrt{\frac{(n)(n+1)(2n+1)}{24}} = \sqrt{\frac{(19)(20)(39)}{24}} = 24.8$$

The observed z value is

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{10.5 - 95}{24.8} = -3.41$$

ACTION:

STEP 7. The observed z value (-3.41) is in the rejection region, so the analyst rejects the null hypothesis. The productivity is significantly greater after the implementation of quality control at this company.

BUSINESS IMPLICATIONS:

STEP 8. Managers, the quality team, and any consultants can point to the figures as validation of the efficacy of the quality program. Such results could be used to justify further activity in the area of quality

Kruskal-Wallis test

The *nonparametric alternative to the one-way analysis of variance* is the **Kruskal-Wallis test**, developed in 1952 by William H. Kruskal and W. Allen Wallis. Like the one-way analysis of variance, the Kruskal-Wallis test is used to determine whether c 3 samples come from the same or different populations. Whereas the one-way ANOVA is based on the assumptions of normally distributed populations, independent groups, at least interval level data, and equal population variances, the Kruskal-Wallis test can be used to analyze ordinal data and is not based on any assumption about population shape. The Kruskal-Wallis test is based on the assumption that the c groups are independent and that individual items are selected randomly.

The hypotheses tested by the Kruskal-Wallis test follow.

H_0 : The c populations are identical.

H_a : At least one of the c populations is different.

This test determines whether all of the groups come from the same or equal populations or whether at least one group comes from a different population. The process of computing a Kruskal-Wallis K statistic begins with ranking the data in all the groups together, as though they were from one group. The smallest value is awarded a 1. As usual, for ties, each value is given the average rank for those tied values. Unlike one way ANOVA, in which the raw data are analyzed, the Kruskal-Wallis test analyzes the ranks of the data.

Formula 17.3 is used to compute a Kruskal-Wallis K statistic.

KRUSKAL-WALLIS TEST (17.3)

$$K = \frac{12}{n(n+1)} \left(\sum_{j=1}^c \frac{T_j^2}{n_j} \right) - 3(n+1)$$

where

c = number of groups

n = total number of items

T_j = total of ranks in a group

n_j = number of items in a group

$K \approx \chi^2$, with $df = c - 1$

Two Partners	Three or More Partners	HMO
13	24	26
15	16	22
20	19	31
18	22	27
23	25	28
	14	33
	17	

Kruskal-Wallis Analysis of Physicians' Patients

Two Partners	Three or More Partners	HMO	
1	12	14	
3	4	9.5	
8	7	17	
6	9.5	15	
11	13	16	
	2	18	
	5		
$T_1 = 29$	$T_2 = 52.5$	$T_3 = 89.5$	
$n_1 = 5$	$n_2 = 7$	$n_3 = 6$	$n = 18$
$\sum_{j=1}^3 \frac{T_j^2}{n_j} = \frac{(29)^2}{5} + \frac{(52.5)^2}{7} + \frac{(89.5)^2}{6} = 1,897$			

The K value is approximately chi-square distributed, with $c - 1$ degrees of freedom as long as n_j is not less than 5 for any group.

Suppose a researcher wants to determine whether the number of physicians in an office produces significant differences in the number of office patients seen by each physician per day. She takes a random sample of physicians from practices in which (1) there are only two partners, (2) there are three or more partners, or (3) the office is a health maintenance organization (HMO). Table 17.4 shows the data she obtained.

Three groups are targeted in this study, so $c = 3$, and $n = 18$ physicians, with the numbers of patients ranked for these physicians. The researcher sums the ranks within each column to obtain T_j , as shown in Table 17.5.

The Kruskal-Wallis K is

$$K = \frac{12}{18(18 + 1)}(1,897) - 3(18 + 1) = 9.56$$

The critical chi-square value is $2\alpha, df$. If $\alpha = .05$ and df for $c - 1 = 3 - 1 = 2$, $2.05, 2 = 5.9915$. This test is always one-tailed, and the rejection region is always in the right tail of the distribution. Because $K = 9.56$ is larger than the critical 2 value, the researcher rejects the null hypothesis. The number of patients seen in the office by a physician is not the same in these three sizes of offices. Examination of the values in each group reveals that physicians in two-partner offices see fewer patients per physician in the office, and HMO physicians see more patients per physician in the office.

Figure 17.7 is the Minitab computer output for this example. The statistic H printed in the output is equivalent to the K statistic calculated here (both K and H are 9.56).

Agribusiness researchers are interested in determining the conditions under which Christmas trees grow fastest. A random sample of equivalent-size seedlings is divided into four groups. The trees are all grown in the same field. One group is left to grow naturally, one group is given extra water, one group is given fertilizer spikes, and one group is given fertilizer spikes and extra water. At the end of one year, the seedlings are measured for growth (in height). These measurements are shown for each group. Use the Kruskal-Wallis test to determine whether there is a significant difference in the growth of trees in these groups. Use $\alpha = .01$.

Group 1 (native)	Group 2 (+ water)	Group 3 (+ fertilizer)	Group 4 (+ water and fertilizer)
8 in.	10 in.	11 in.	18 in.
5	12	14	20
7	11	10	16
11	9	16	15
9	13	17	14
6	12	12	22

Solution

Here, $n = 24$, and $n_j = 6$ in each group.

HYPOTHESIZE:

STEP 1. The hypotheses follow.

H₀: group 1 = group 2 = group 3 = group 4

H_a: At least one group is different.

TEST:

STEP 2. The Kruskal-Wallis K is the appropriate test statistic.

STEP 3. Alpha is .01.

STEP 4. The degrees of freedom are $c - 1 = 4 - 1 = 3$. The critical value of chi square is $.01, 3 = 11.3449$. If the observed value of K is greater than 11.3449, the decision is to reject the null hypothesis.

STEP 5. The data are as shown previously.

STEP 6. Ranking all group values yields the following.

1	2	3	4	
4	7.5	10	22	
1	13	16.5	23	
3	10	7.5	19.5	
10	5.5	19.5	18	
5.5	15	21	16.5	
2	13	13	24	
$T_1 = 25.5$	$T_2 = 64.0$	$T_3 = 87.5$	$T_4 = 123.0$	
$n_1 = 6$	$n_2 = 6$	$n_3 = 6$	$n_4 = 6$	$n = 24$

$$\sum_{j=1}^c \frac{T_j^2}{n_j} = \frac{(25.5)^2}{6} + \frac{(64)^2}{6} + \frac{(87.5)^2}{6} + \frac{(123)^2}{6} = 4,588.6$$

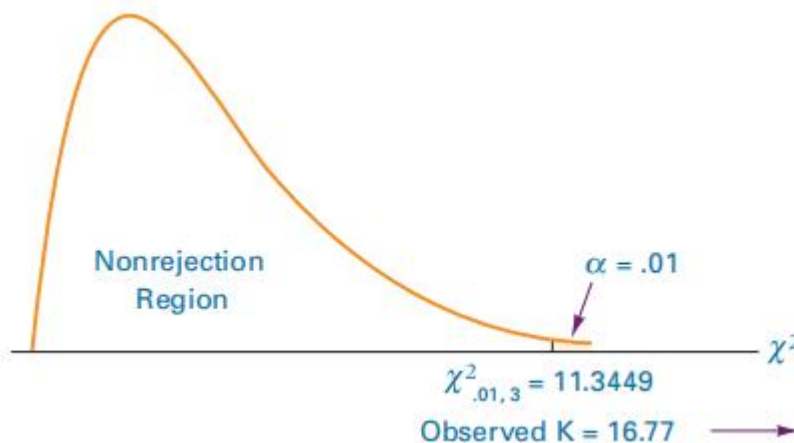
$$K = \frac{12}{24(24 + 1)} (4,588.6) - 3(24 + 1) = 16.77$$

ACTION:

STEP 7. The observed K value is 16.77 and the critical $\chi^2_{.01,3} = 11.3449$. Because the observed value is greater than the table value, the null hypothesis is rejected. There is a significant difference in the way the trees grow.

BUSINESS IMPLICATIONS:

STEP 8. From the increased heights in the original data, the trees with both water and fertilizer seem to be doing the best. However, these are sample data; without analyzing the pairs of samples with nonparametric multiple comparisons (not included in this text), it is difficult to conclude whether the water/fertilizer group is actually growing faster than the others. It appears that the trees under natural conditions are growing more slowly than the others. The following diagram shows the relationship of the observed K value and the critical chi-square value.



Kruskal-Wallis test

The *nonparametric alternative to the one-way analysis of variance* is the **Kruskal-Wallis test**, developed in 1952 by William H. Kruskal and W. Allen Wallis. Like the one-way analysis of variance, the Kruskal-Wallis test is used to determine whether c samples come from the same or different populations. Whereas the one-way ANOVA is based on the assumptions of normally distributed populations, independent groups, at least interval level data, and equal population variances, the Kruskal-Wallis test can be used to analyze ordinal data and is not based on any assumption about population shape. The Kruskal-Wallis test is based on the assumption that the c groups are independent and that individual items are selected randomly.

The hypotheses tested by the Kruskal-Wallis test follow.

H_0 : The c populations are identical.

H_a : At least one of the c populations is different.

This test determines whether all of the groups come from the same or equal populations or whether at least one group comes from a different population. The process of computing a Kruskal-Wallis K statistic begins with ranking the data in all the groups together, as though they were from one group. The smallest value is awarded a 1. As usual, for ties, each value is given the average rank for those tied values. Unlike one way ANOVA, in which the raw data are analyzed, the Kruskal-Wallis test analyzes the ranks of the data.

Formula 17.3 is used to compute a Kruskal-Wallis K statistic.

KRUSKAL-WALLIS TEST (17.3)

$$K = \frac{12}{n(n+1)} \left(\sum_{j=1}^c \frac{T_j^2}{n_j} \right) - 3(n+1)$$

where

c = number of groups

n = total number of items

T_j = total of ranks in a group

n_j = number of items in a group

$K \approx \chi^2$, with $df = c - 1$

Illustration :-

Agribusiness researchers are interested in determining the conditions under which Christmas trees grow fastest. A random sample of equivalent-size seedlings is divided into four groups. The trees are all grown in the same field. One group is left to grow naturally, one group is given extra water, one group is given fertilizer spikes, and one group is given fertilizer spikes and extra water. At the end of one year, the seedlings are measured for growth (in height). These measurements are shown for each group. Use the Kruskal-Wallis test to determine whether there is a significant difference in the growth of trees in these groups. Use $\alpha = .01$.

Group 1 (native)	Group 2 (+ water)	Group 3 (+ fertilizer)	Group 4 (+ water and fertilizer)
8 in.	10 in.	11 in.	18 in.
5	12	14	20
7	11	10	16
11	9	16	15
9	13	17	14
6	12	12	22

Solution

Here, $n = 24$, and $n_j = 6$ in each group.

HYPOTHESIZE:

STEP 1. The hypotheses follow.

H_0 : group 1 = group 2 = group 3 = group 4

H_a : At least one group is different.

TEST:

STEP 2. The Kruskal-Wallis K is the appropriate test statistic.

STEP 3. Alpha is .01.

STEP 4. The degrees of freedom are $c - 1 = 4 - 1 = 3$. The critical value of chi square is $.01, 3 = 11.3449$. If the observed value of K is greater than 11.3449, the decision is to reject the null hypothesis.

STEP 5. The data are as shown previously.

STEP 6. Ranking all group values yields the following.

1	2	3	4	
4	7.5	10	22	
1	13	16.5	23	
3	10	7.5	19.5	
10	5.5	19.5	18	
5.5	15	21	16.5	
2	13	13	24	
$T_1 = 25.5$	$T_2 = 64.0$	$T_3 = 87.5$	$T_4 = 123.0$	
$n_1 = 6$	$n_2 = 6$	$n_3 = 6$	$n_4 = 6$	$n = 24$

$$\sum_{j=1}^c \frac{T_j^2}{n_j} = \frac{(25.5)^2}{6} + \frac{(64)^2}{6} + \frac{(87.5)^2}{6} + \frac{(123)^2}{6} = 4,588.6$$

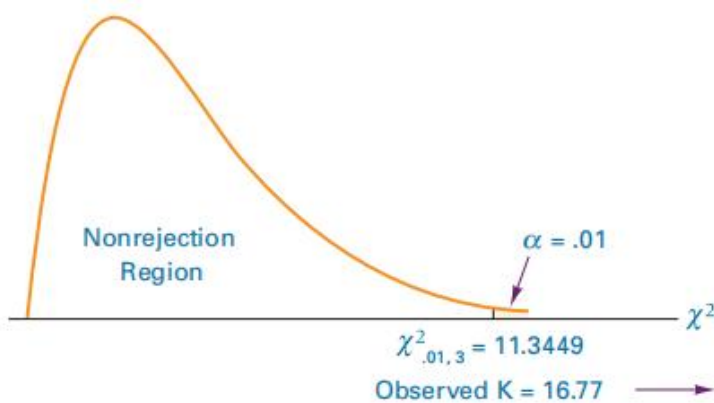
$$K = \frac{12}{24(24 + 1)} (4,588.6) - 3(24 + 1) = 16.77$$

ACTION:

STEP 7. The observed K value is 16.77 and the critical $2.01,3 = 11.3449$. Because the observed value is greater than the table value, the null hypothesis is rejected. There is a significant difference in the way the trees grow.

BUSINESS IMPLICATIONS:

STEP 8. From the increased heights in the original data, the trees with both water and fertilizer seem to be doing the best. However, these are sample data; without analyzing the pairs of samples with nonparametric multiple comparisons (not included in this text), it is difficult to conclude whether the water/fertilizer group is actually growing faster than the others. It appears that the trees under natural conditions are growing more slowly than the others. The following diagram shows the relationship of the observed K value and the critical chi-square value.



Friedman test

The **Friedman test**, developed by M. Friedman in 1937, is a *nonparametric alternative to the randomized block design* discussed in Chapter 11. The randomized block design has the same assumptions as other ANOVA procedures, including observations are drawn from normally distributed populations. When this assumption cannot be met or when the researcher has ranked data, the Friedman test provides a nonparametric alternative.

Three assumptions underlie the Friedman test.

1. The blocks are independent.
2. No interaction is present between blocks and treatments.
3. Observations within each block can be ranked.

The hypotheses being tested are as follows.

H₀: The treatment populations are equal.

H_a: At least one treatment population yields larger values than at least one other treatment population.

The first step in computing a Friedman test is to convert all raw data to ranks (unless the data are already ranked). However, unlike the Kruskal-Wallis test where all data are ranked together, the data in a Friedman test are ranked *within* each block from smallest (1) to largest (c). Each block contains c ranks, where c is the number of treatment levels. Using these ranks, the Friedman test will test to determine whether it is likely that the different treatment levels (columns) came from the same population. Formula 17.4 is used to calculate the test statistic, which is approximately chi-square distributed with df = c - 1 if c > 4 or when c = 3 and b > 9, or when c = 4 and b > 4.

FRIEDMAN TEST (17.4)

$$\chi_r^2 = \frac{12}{bc(c+1)} \sum_{j=1}^c R_j^2 - 3b(c+1)$$

where

c = number of treatment levels (columns)

b = number of blocks (rows)

R_j = total of ranks for a particular treatment level (column)

j = particular treatment level (column)

$\chi_r^2 \approx \chi^2$, with df = c - 1

As an example, suppose a manufacturing company assembles microcircuits that contain a plastic housing. Managers are _____ concerned about an

unacceptably high number of the products that sustained housing damage during shipment. The housing component is made by four different suppliers. Managers have decided to conduct a study of the plastic housing by randomly selecting five housings made by each of the four suppliers. To determine whether a supplier is consistent during the production week, one housing is selected for each day of the week. That is, for each supplier, a housing made on Monday is selected, one made on Tuesday is selected, and so on.

In analyzing the data, the treatment variable is supplier and the treatment levels are the four suppliers. The blocking effect is day of the week with each day representing a block level. The quality control team wants to determine whether there is any significant difference in the tensile strength of the plastic housing by supplier. The data are given here (in pounds per inch).

Day	Supplier 1	Supplier 2	Supplier 3	Supplier 4
Monday	62	63	57	61
Tuesday	63	61	59	65
Wednesday	61	62	56	63
Thursday	62	60	57	64
Friday	64	63	58	66

HYPOTHESIZE:

STEP 1. The hypotheses follow.

H₀: The supplier populations are equal.

H_a: At least one supplier population yields larger values than at least one other supplier population.

TEST:

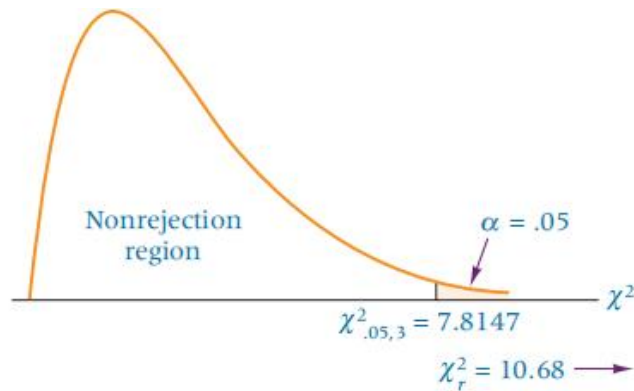
STEP 2. The quality researchers do not feel they have enough evidence to conclude that the observations come from normally distributed populations. Because they are analyzing a randomized block design, the Friedman test is appropriate.

STEP 3. Let $\alpha = .05$.

STEP 4. For four treatment levels (suppliers), $c = 4$ and $df = 4 - 1 = 3$. The critical value is $2.05, 3 = 7.8147$. If the observed chi-square is greater than 7.8147, the decision is to reject the null hypothesis.

STEP 5. The sample data are as given.

STEP 6. The calculations begin by ranking the observations in each row with 1 designating the rank of the smallest observation. The ranks are then summed for each column, producing R_j . The values of R_j are squared and then summed. Because the study is concerned with five days of the week, five blocking levels are used and $b = 5$. The value of R_j is computed as shown in the following table.



Day	Supplier 1	Supplier 2	Supplier 3	Supplier 4
Monday	3	4	1	2
Tuesday	3	2	1	4
Wednesday	2	3	1	4
Thursday	3	2	1	4
Friday	<u>3</u>	<u>2</u>	<u>1</u>	<u>4</u>
R_j	14	13	5	18
R_j^2	196	169	25	324

$$\sum_{j=1}^4 R_j^2 = (196 + 169 + 25 + 324) = 714$$

$$\chi_r^2 = \frac{12}{bc(c+1)} \sum_{j=1}^c R_j^2 - 3b(c+1) = \frac{12}{5(4)(4+1)} (714) - 3(5)(4+1) = 10.68$$

ACTION:

STEP 7. Because the observed value of is greater than the critical value,

$\chi_{.05,3}^2 = 7.8147$, the decision is to reject the null hypothesis.

BUSINESS IMPLICATIONS:

STEP 8. Statistically, there is a significant difference in the tensile strength of housings made by different suppliers. The sample data indicate that supplier 3 is producing housings with a lower tensile strength than those made by other suppliers and that supplier 4 is producing housings with higher tensile strength. Further study by managers and a quality team may result in attempts to bring supplier 3 up to standard on tensile strength or perhaps cancellation of the contract.

Illustration :-

A market research company wants to determine brand preference for refrigerators. Five companies contracted with the research company to have their products be included in the study. As part of the study, the research company randomly selects 10 potential refrigerator buyers and shows them one of each of the five brands. Each survey participant is then asked to rank the refrigerator brands from 1 to 5. The results of these rankings are given in the table. Use the Friedman test and $\alpha = .01$ to determine whether there are any significant differences between the rankings of these brands.

Solution

HYPOTHESIZE:

STEP 1. The hypotheses are as follows.

H_0 : The brand populations are equal.

H_a : At least one brand population yields larger values than at least one other brand population.

TEST:

STEP 2. The market researchers collected ranked data that are ordinal in level. The Friedman test is the appropriate test.

STEP 3. Let $\alpha = .01$.

STEP 4. Because the study uses five treatment levels (brands), $c = 5$ and $df = 5 - 1 = 4$. The critical value is $2.01, 4 = 13.2767$. If the observed chi-square is greater than 13.2767, the decision is to reject the null hypothesis.

STEP 5. The sample data follow.

STEP 6. The ranks are totaled for each column, squared, and then summed across the column totals. The results are shown in the table.

Individual	Brand A	Brand B	Brand C	Brand D	Brand E
1	3	5	2	4	1
2	1	3	2	4	5
3	3	4	5	2	1
4	2	3	1	4	5
5	5	4	2	1	3
6	1	5	3	4	2
7	4	1	3	2	5
8	2	3	4	5	1
9	2	4	5	3	1
10	3	5	4	2	1
R_j	26	37	31	31	25
R_j^2	676	1,369	961	961	625
$\Sigma R_j^2 = 4,592$					

The value of χ_r^2 is

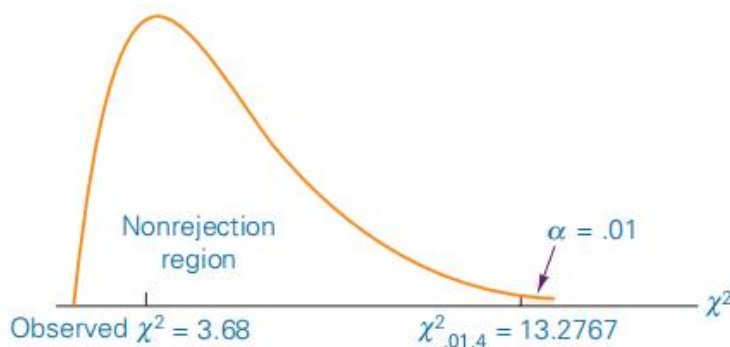
$$\chi_r^2 = \frac{12}{bc(c+1)} \sum_{j=1}^c R_j^2 - 3b(c+1) = \frac{12}{10(5)(5+1)} (4,592) - 3(10)(5+1) = 3.68$$

ACTION:

STEP 7. Because the observed value of = 3.68 is not greater than the critical value, 2.01,4 = 13.2767, the researchers fail to reject the null hypothesis.

BUSINESS IMPLICATIONS:

STEP 8. Potential refrigerator purchasers appear to have no significant brand preference. Marketing managers for the various companies might want to develop strategies for positively distinguishing their product from the others.



Spearman's rank correlation

The Pearson product-moment correlation coefficient, r , was presented and discussed as a technique to measure the amount or degree of association between two variables. The Pearson r requires at least interval level of measurement for the data. When only ordinal-level data or ranked data are available, **Spearman's rank correlation**, r_s , can be used to analyze the degree of association of two variables. Charles E. Spearman (1863–1945) developed this correlation coefficient.

The formula for calculating a Spearman's rank correlation is as follows:

**SPEARMAN'S RANK
CORRELATION (17.7)**

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

where

n = number of pairs being correlated

d = the difference in the ranks of each pair

Wilcoxon matched-pairs signed rank test

The Mann-Whitney U test presented in Section 17.2 is a nonparametric alternative to the t test for two *independent* samples. If the two samples are *related*, the U test is not applicable. A test that does handle related data is the **Wilcoxon matched-pairs signed rank test**, which serves as a *nonparametric alternative to the t test for two related samples*. Developed by Frank Wilcoxon in 1945, the Wilcoxon test, like the t test for two related samples, is used to analyze several different types of studies when the data of one group are related to the data in the other group, including before-and-after studies, studies in which measures are taken on the same person or object under two different conditions, and studies of twins or other relatives.

The Wilcoxon test utilizes the differences of the scores of the two matched groups in a manner similar to that of the t test for two related samples. After the difference scores have been computed, the Wilcoxon test ranks all differences regardless of whether the difference is positive or negative. The values are ranked from smallest to largest, with a rank of 1 assigned to the smallest difference. If a difference is negative, the rank is given a negative sign. The sum of the positive ranks is tallied along with the sum of the negative ranks. Zero differences representing ties between scores from the two groups are ignored, and the value of n is reduced accordingly. When ties occur between ranks, the ranks are averaged over the values. The smallest sum of ranks (either + or

-) is used in the analysis and is represented by T . The Wilcoxon matched-pairs signed rank test procedure for determining statistical significance differs with sample size. When the number of matched pairs, n , is greater than 15, the value of T is approximately normally distributed and a z score is computed to test the null hypothesis. When sample size is small, $n \leq 15$, a different procedure is followed.

Two assumptions underlie the use of this technique.

1. The paired data are selected randomly.
2. The underlying distributions are symmetrical.

For two-tailed tests:

$$H_0: M_d = 0 \quad H_a: M_d \neq 0$$

For one-tailed tests:

$$H_0: M_d = 0 \quad H_a: M_d > 0$$

or

$$H_0: M_d = 0 \quad H_a: M_d < 0$$

where M_d is the median.

Small-Sample Case ($n \leq 15$)

When sample size is small, a critical value against which to compare T can be found in Table A.14 to determine whether the null hypothesis should be rejected. The critical value is located by using n and α . Critical values are given in the table for $\alpha = .05, .025, .01, \text{ and } .005$ for two-tailed tests and $\alpha = .10, .05, .02, \text{ and } .01$ for one-tailed tests. If the observed value of T is less than or equal to the critical value of T , the decision is to reject the null hypothesis.

As an example, consider the survey by American Demographics that estimated the average annual household spending on healthcare. The U.S. metropolitan average was \$1,800. Suppose six families in Pittsburgh, Pennsylvania, are matched demographically with six families in Oakland, California, and their amounts of household spending on healthcare for last year are obtained. The data follow on the next page.

Family Pair	Pittsburgh	Oakland
1	\$1,950	\$1,760
2	1,840	1,870
3	2,015	1,810
4	1,580	1,660
5	1,790	1,340
6	1,925	1,765

A healthcare analyst uses $\alpha = .05$ to test to determine whether there is a significant difference in annual household healthcare spending between these two cities.

HYPOTHESIZE:

STEP 1. The following hypotheses are being tested.

$$H_0: M_d = 0$$

$$H_a: M_d \neq 0$$

TEST:

STEP 2. Because the sample size of pairs is six, the small-sample Wilcoxon matched pairs signed ranks test is appropriate if the underlying distributions are assumed to be symmetrical.

STEP 3. Alpha is .05.

STEP 4. From Table A.14, if the observed value of T is less than or equal to 1, the decision is to reject the null hypothesis.

STEP 5. The sample data were listed earlier.

STEP 6.

Family Pair	Pittsburgh	Oakland	d	Rank
1	\$1,950	\$1,760	+190	+4
2	1,840	1,870	-30	-1
3	2,015	1,810	+205	+5
4	1,580	1,660	-80	-2
5	1,790	1,340	+450	+6
6	1,925	1,765	+160	+3

$$T = \text{minimum of } (T_+, T_-)$$

$$T_+ = 4 + 5 + 6 + 3 = 18$$

$$T_- = 1 + 2 = 3$$

$$T = \text{minimum of } (18, 3) = 3$$

ACTION:

STEP 7. Because $T = 3$ is greater than critical $T = 1$, the decision is not to reject the null hypothesis.

BUSINESS IMPLICATIONS:

STEP 8. Not enough evidence is provided to declare that Pittsburgh and Oakland differ in annual household spending on healthcare. This information may be useful to healthcare providers and employers in the two cities and particularly to businesses that either operate in both cities or are planning to move from one to the other. Rates can be established on the notion that healthcare costs are about the same in both cities. In addition, employees considering transfers from one city to the other can expect their annual healthcare costs to remain about the same.

Large-Sample Case ($n > 15$)

For large samples, the T statistic is approximately normally distributed and a z score can be used as the test statistic. Formula 17.2 contains the necessary formulas to complete this procedure.

WILCOXON MATCHED-PAIRS SIGNED RANK TEST (17.2)

$$\mu_T = \frac{(n)(n + 1)}{4}$$

$$\sigma_T = \sqrt{\frac{(n)(n + 1)(2n + 1)}{24}}$$

$$z = \frac{T - \mu_T}{\sigma_T}$$

where

n = number of pairs

T = total ranks for either + or - differences, whichever is less in magnitude

This technique can be applied to the airline industry, where an analyst might want to determine whether there is a difference in the cost per mile of airfares in the United States between 1979 and 2009 for various cities. The data in Table 17.3 represent the costs per mile of airline tickets for a sample of 17 cities for both 1979 and 2009.

HYPOTHESIZE:

STEP 1. The analyst states the hypotheses as follows.

TEST:

STEP 2. The analyst applies a Wilcoxon matched-pairs signed rank test to the data to test the difference in cents per mile for the two periods of time. She assumes the underlying distributions are symmetrical.

STEP 3. Use $\alpha = .05$.

City	1979	2009
1	20.3	22.8
2	19.5	12.7
3	18.6	14.1
4	20.9	16.1
5	19.9	25.2
6	18.6	20.2
7	19.6	14.9
8	23.2	21.3
9	21.8	18.7
10	20.3	20.9
11	19.2	22.6
12	19.5	16.9
13	18.7	20.6
14	17.7	18.5
15	21.6	23.4
16	22.4	21.3
17	20.8	17.4

STEP 4. Because this test is two-tailed, $\alpha = .025$ and the critical values are $z = 1.96$.

If the observed value of the test statistic is greater than 1.96 or less than -1.96, the null hypothesis is rejected.

STEP 5. The sample data are given in Table 17.3.

STEP 6. The analyst begins the process by computing a difference score, d . Which year's data are subtracted from the other does not matter as long as consistency in direction is maintained. For the data in Table 17.3, the analyst subtracted the 2009 figures from the 1979 figures. The sign of the difference is left on the difference score. Next, she ranks the differences without regard to sign, but the sign is left on the rank as an identifier. Note the tie for ranks 6 and 7; each is given a rank of 6.5, the average of the two ranks. The same applies to ranks 11 and 12.

After the analyst ranks all difference values regardless of sign, she sums the positive ranks (T_+) and the negative ranks (T_-). She then determines the T value from these two sums as the smallest T_+ or T_- .

$$\begin{aligned}T &= \text{minimum of } (T_+, T_-) \\T_+ &= 17 + 13 + 15 + 14 + 6.5 + 10 + 9 + 3 + 11.5 = 99 \\T_- &= 8 + 16 + 4 + 1 + 11.5 + 6.5 + 2 + 5 = 54 \\T &= \text{minimum of } (99, 54) = 54\end{aligned}$$

The T value is normally distributed for large sample sizes, with a mean and standard deviation of

$$\begin{aligned}\mu_T &= \frac{(n)(n+1)}{4} = \frac{(17)(18)}{4} = 76.5 \\ \sigma_T &= \sqrt{\frac{(n)(n+1)(2n+1)}{24}} = \sqrt{\frac{(17)(18)(35)}{24}} = 21.1\end{aligned}$$

The observed z value is

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{54 - 76.5}{21.1} = -1.07$$

$$\begin{aligned}T &= \text{minimum of } (T_+, T_-) \\T_+ &= 17 + 13 + 15 + 14 + 6.5 + 10 + 9 + 3 + 11.5 = 99 \\T_- &= 8 + 16 + 4 + 1 + 11.5 + 6.5 + 2 + 5 = 54 \\T &= \text{minimum of } (99, 54) = 54\end{aligned}$$

The T value is normally distributed for large sample sizes, with a mean and standard deviation of

$$\begin{aligned}\mu_T &= \frac{(n)(n+1)}{4} = \frac{(17)(18)}{4} = 76.5 \\ \sigma_T &= \sqrt{\frac{(n)(n+1)(2n+1)}{24}} = \sqrt{\frac{(17)(18)(35)}{24}} = 21.1\end{aligned}$$

The observed z value is

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{54 - 76.5}{21.1} = -1.07$$

ACTION:

STEP 7. The critical z value for this two-tailed test is $z_{0.025} = 1.96$. The observed $z = -1.07$, so the analyst fails to reject the null hypothesis. There is no significant difference in the cost of airline tickets between 1979 and 2009.

BUSINESS IMPLICATIONS:

STEP 8. Promoters in the airline industry can use this type of information (the fact that ticket prices have not increased significantly in 30 years) to sell their product as a good buy. In addition, industry managers could use it as an argument for raising prices.

Illustration :-

During the 1980s and 1990s, U.S. businesses increasingly emphasized quality control. One of the arguments in favor of quality-control programs is that quality control can increase productivity. Suppose a company implemented a quality-control program and has been operating under it for 2 years. The company's president wants to determine whether worker productivity significantly increased since installation of the program. Company records contain the figures for items produced per worker during a sample of production runs 2 years ago. Productivity figures on the same workers are gathered now and compared to the previous figures. The following data

represent items produced per hour. The company's statistical analyst uses the Wilcoxon matched-pairs signed rank test to determine whether there is a significant increase in per worker production for $\alpha = .01$.

Worker	Before	After	Worker	Before	After
1	5	11	11	2	6
2	4	9	12	5	10
3	9	9	13	4	9
4	6	8	14	5	7
5	3	5	15	8	9
6	8	7	16	7	6
7	7	9	17	9	10
8	10	9	18	5	8
9	3	7	19	4	5
10	7	9	20	3	6

Solution

HYPOTHESIZE:

STEP 1. The hypotheses are as follows.

TEST:

STEP 2. The analyst applies a Wilcoxon matched-pairs signed rank test to the data to test the difference in productivity from before to after. He assumes the underlying distributions are symmetrical.

STEP 3. Use $\alpha = .01$.

STEP 4. This test is one tailed. The critical value is $z = -2.33$. If the observed value of the test statistic is less than -2.33 , the null hypothesis is rejected.

STEP 5. The sample data are as already given.

STEP 6. The analyst computes the difference values, and, because zero differences are to be eliminated, deletes worker 3 from the study. This reduces n from 20 to 19. He then ranks the differences regardless of sign. The differences that are the same (ties) receive the

average rank for those values. For example, the differences for workers 4, 5, 7, 10, and 14 are the same. The ranks for these five are 7, 8, 9, 10, and 11, so each worker receives the rank of 9, the average of these five ranks.

Worker	Before	After	<i>d</i>	Rank
1	5	11	-6	-19
2	4	9	-5	-17
3	9	9	0	delete
4	6	8	-2	-9
5	3	5	-2	-9
6	8	7	+1	+3.5
7	7	9	-2	-9
8	10	9	+1	+3.5
9	3	7	-4	-14.5
10	7	9	-2	-9
11	2	6	-4	-14.5
12	5	10	-5	-17
13	4	9	-5	-17
14	5	7	-2	-9
15	8	9	-1	-3.5
16	7	6	+1	+3.5
17	9	10	-1	-3.5
18	5	8	-3	-12.5
19	4	5	-1	-3.5
20	3	6	-3	-12.5

The analyst determines the values of T_+ , T_- , and T

The analyst determines the values of T_+ , T_- , and T to be

$$T_+ = 3.5 + 3.5 + 3.5 = 10.5$$

$$T_- = 19 + 17 + 9 + 9 + 9 + 14.5 + 9 + 14.5 + 17 + 17 \\ + 9 + 3.5 + 3.5 + 12.5 + 3.5 + 12.5 = 179.5$$

$$T = \text{minimum of } (10.5, 179.5) = 10.5$$

The mean and standard deviation of T are

$$\mu_T = \frac{(n)(n+1)}{4} = \frac{(19)(20)}{4} = 95$$

$$\sigma_T = \sqrt{\frac{(n)(n+1)(2n+1)}{24}} = \sqrt{\frac{(19)(20)(39)}{24}} = 24.8$$

The observed z value is

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{10.5 - 95}{24.8} = -3.41$$

ACTION:

STEP 7. The observed z value (-3.41) is in the rejection region, so the analyst rejects the null hypothesis. The productivity is significantly greater after the implementation of quality control at this company.

BUSINESS IMPLICATIONS:

STEP 8. Managers, the quality team, and any consultants can point to the figures as validation of the efficacy of the quality program. Such results could be used to justify further activity in the area of quality.

Kruskal-Wallis test

The *nonparametric alternative to the one-way analysis of variance* is the **Kruskal-Wallis test**, developed in 1952 by William H. Kruskal and W. Allen Wallis. Like the one-way analysis of variance, the Kruskal-Wallis test is used to determine whether c 3 samples come from the same or different populations. Whereas the one-way ANOVA is based on the assumptions of normally distributed populations, independent groups, at least interval level data, and equal population variances, the Kruskal-Wallis test can be used to analyze ordinal data and is not based on any assumption about population shape. The Kruskal-Wallis test is based on the assumption that the c groups are independent and that individual items are selected randomly.

The hypotheses tested by the Kruskal-Wallis test follow.

H_0 : The c populations are identical.

H_a : At least one of the c populations is different.

This test determines whether all of the groups come from the same or equal populations or whether at least one group comes from a different population. The process of computing a Kruskal-Wallis K statistic begins with ranking the data in all the groups together, as though they were from one group. The smallest value is awarded a 1. As usual, for ties, each value is given the average rank for those tied values. Unlike one way ANOVA, in which the raw data are analyzed, the Kruskal-Wallis test analyzes the ranks of the data.

Formula 17.3 is used to compute a Kruskal-Wallis K statistic.

KRUSKAL-WALLIS TEST
(17.3)

$$K = \frac{12}{n(n+1)} \left(\sum_{j=1}^c \frac{T_j^2}{n_j} \right) - 3(n+1)$$

where

c = number of groups

n = total number of items

T_j = total of ranks in a group

n_j = number of items in a group

$K \approx \chi^2$, with $df = c - 1$

**Number of Office Patients
per Doctor**

	Two Partners	Three or More Partners	HMO
	13	24	26
	15	16	22
	20	19	31
	18	22	27
	23	25	28
		14	33
		17	

Kruskal-Wallis Analy

Two Partners	Three or Partn
1	12
3	4
8	7
6	9.5
11	13
	2
	5
$T_1 = 29$	$T_2 = 52$
$n_1 = 5$	$n_2 = 7$
$\sum_{j=1}^3 \frac{T_j^2}{n_j} = \frac{(29)^2}{5} + \dots$	

The K value is approximately chi-square distributed, with $c - 1$ degrees of freedom as long as n_j is not less than 5 for any group. Suppose a researcher wants to determine whether the number of physicians in an office produces significant differences in the number of office patients seen by each physician per day. She takes a random sample of physicians from practices in which (1) there are only two partners, (2) there are three or more partners, or (3) the office is a health maintenance organization (HMO). Table 17.4 shows the data she obtained.

Three groups are targeted in this study, so $c = 3$, and $n = 18$ physicians, with the numbers of patients ranked for these physicians. The researcher sums the ranks within each column to obtain T_j , as shown in Table 17.5.

The Kruskal-Wallis K is

$$K = \frac{12}{18(18 + 1)}(1,897) - 3(18 + 1) = 9.56$$

The critical chi-square value is $2\alpha, df$. If $\alpha = .05$ and df for $c - 1 = 3 - 1 = 2$, $2.05, 2 =$

5.9915. This test is always one-tailed, and the rejection region is always in the right tail of the distribution. Because $K = 9.56$ is larger than the critical 2 value, the researcher rejects the null hypothesis. The number of patients seen in the office by a physician is not the same in these three sizes of offices. Examination of the values in each group reveals that physicians in two-partner offices see fewer patients per physician in the office, and HMO physicians see more patients per physician in the office.

Figure 17.7 is the Minitab computer output for this example. The statistic H printed in the output is equivalent to the K statistic calculated here (both K and H are 9.56).

Illustration :-

Agribusiness researchers are interested in determining the conditions under which Christmas trees grow fastest. A random sample of equivalent-size seedlings is divided into four groups. The trees are all grown in the same field. One group is left to grow naturally, one group is given extra water, one group is given fertilizer spikes, and one group is given fertilizer spikes and extra water. At the end of one

year, the seedlings are measured for growth (in height). These measurements are shown for each group. Use the Kruskal-Wallis test to determine whether there is a significant difference in the growth of trees in these groups. Use $\alpha = .01$.

Group 1 (native)	Group 2 (+ water)	Group 3 (+ fertilizer)	Group 4 (+ water and fertilizer)
8 in.	10 in.	11 in.	18 in.
5	12	14	20
7	11	10	16
11	9	16	15
9	13	17	14
6	12	12	22

Solution

Here, $n = 24$, and $n_j = 6$ in each group.

HYPOTHESIZE:

STEP 1. The hypotheses follow.

H_0 : group 1 = group 2 = group 3 = group 4

H_a : At least one group is different.

TEST:

STEP 2. The Kruskal-Wallis K is the appropriate test statistic.

STEP 3. Alpha is .01.

STEP 4. The degrees of freedom are $c - 1 = 4 - 1 = 3$. The critical value of chi square is $.01, 3 = 11.3449$. If the observed value of K is greater than 11.3449, the decision is to reject the null hypothesis.

STEP 5. The data are as shown previously.

STEP 6. Ranking all group values yields the following.

1	2	3	4		
4	7.5	10	22		
1	13	16.5	23		
3	10	7.5	19.5		
10	5.5	19.5	18		
5.5	15	21	16.5		
<u>2</u>	<u>13</u>	<u>13</u>	<u>24</u>		
$T_1 = 25.5$	$T_2 = 64.0$	$T_3 = 87.5$	$T_4 = 123.0$		
$n_1 = 6$	$n_2 = 6$	$n_3 = 6$	$n_4 = 6$	$n = 24$	

$$\sum_{j=1}^c \frac{T_j^2}{n_j} = \frac{(25.5)^2}{6} + \frac{(64)^2}{6} + \frac{(87.5)^2}{6} + \frac{(123)^2}{6} = 4,588.6$$

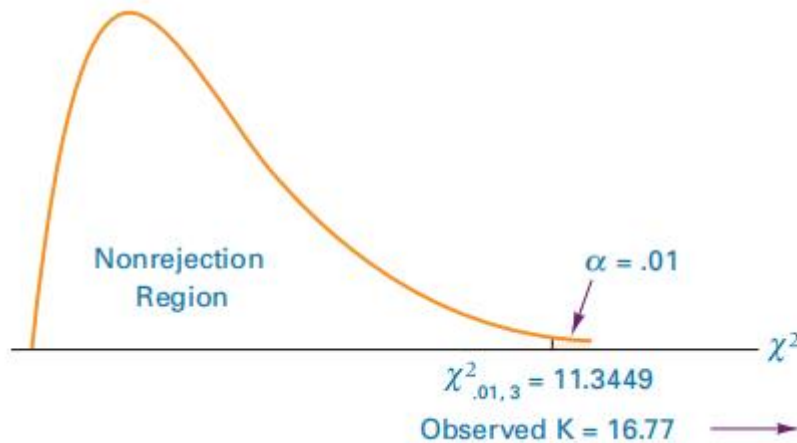
$$K = \frac{12}{24(24 + 1)} (4,588.6) - 3(24 + 1) = 16.77$$

ACTION:

STEP 7. The observed K value is 16.77 and the critical $2.01,3 = 11.3449$. Because the observed value is greater than the table value, the null hypothesis is rejected. There is a significant difference in the way the trees grow.

BUSINESS IMPLICATIONS:

STEP 8. From the increased heights in the original data, the trees with both water and fertilizer seem to be doing the best. However, these are sample data; without analyzing the pairs of samples with nonparametric multiple comparisons (not included in this text), it is difficult to conclude whether the water/fertilizer group is actually growing faster than the others. It appears that the trees under natural conditions are growing more slowly than the others. The following diagram shows the relationship of the observed K value and the critical chi-square value.



Friedman test

The **Friedman test**, developed by M. Friedman in 1937, is a *nonparametric alternative to the randomized block design* discussed in Chapter 11. The randomized block design has the same assumptions as other ANOVA procedures, including observations are drawn from normally distributed populations. When this assumption cannot be met or when the researcher has ranked data, the Friedman test provides a nonparametric alternative.

Three assumptions underlie the Friedman test.

1. The blocks are independent.
2. No interaction is present between blocks and treatments.
3. Observations within each block can be ranked.

The hypotheses being tested are as follows.

H₀: The treatment populations are equal.

H_a: At least one treatment population yields larger values than at least one other treatment population.

The first step in computing a Friedman test is to convert all raw data to ranks (unless the data are already ranked). However, unlike the Kruskal-Wallis test where all data are ranked together, the data in a Friedman test are ranked *within* each block from smallest (1) to largest (*c*). Each block contains *c* ranks, where *c* is the number of treatment levels. Using these ranks, the Friedman test will test to determine whether it is likely that the different treatment levels (columns) came from the same population. Formula 17.4 is used to calculate the test statistic, which is approximately chi-square distributed with $df = c - 1$ if $c > 4$ or when $c = 3$ and $b > 9$, or when $c = 4$ and $b > 4$.

FRIEDMAN TEST (17.4)

$$\chi_r^2 = \frac{12}{bc(c+1)} \sum_{j=1}^c R_j^2 - 3b(c+1)$$

where

c = number of treatment levels (columns)

b = number of blocks (rows)

R_j = total of ranks for a particular treatment level (column)

j = particular treatment level (column)

$\chi_r^2 \approx \chi^2$, with $df = c - 1$

As an example, suppose a manufacturing company assembles microcircuits that contain a plastic housing. Managers are concerned about an unacceptably high number of the products that sustained housing damage during shipment. The housing component is made by four different suppliers. Managers have decided to conduct a study of the plastic housing by randomly selecting five housings made by each of the four suppliers. To determine whether a supplier is consistent during the production week, one housing is selected for each day of the week. That is, for each supplier, a housing made on Monday is selected, one made on Tuesday is selected, and so on.

In analyzing the data, the treatment variable is supplier and the treatment levels are the four suppliers. The blocking effect is day of the week with each day representing a block level. The quality control team wants to determine whether there is any significant difference in the tensile strength of the plastic housing by supplier. The data are given here (in pounds per inch).

Day	Supplier 1	Supplier 2	Supplier 3	Supplier 4
Monday	62	63	57	61
Tuesday	63	61	59	65
Wednesday	61	62	56	63
Thursday	62	60	57	64
Friday	64	63	58	66

HYPOTHESIZE:

STEP 1. The hypotheses follow.

H₀: The supplier populations are equal.

H_a: At least one supplier population yields larger values than at least one other supplier population.

TEST:

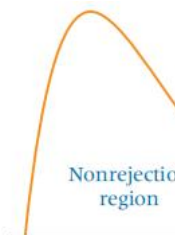
STEP 2. The quality researchers do not feel they have enough evidence to conclude that the observations come from normally distributed populations. Because they are analyzing a randomized block design, the Friedman test is appropriate.

STEP 3. Let $\alpha = .05$.

STEP 4. For four treatment levels (suppliers), $c = 4$ and $df = 4 - 1 = 3$. The critical value is $\chi^2_{.05,3} = 7.8147$. If the observed chi-square is greater than 7.8147, the decision is to reject the null hypothesis.

STEP 5. The sample data are as given.

STEP 6. The calculations begin by ranking the observations in each row with 1 designating the rank of the smallest observation. The ranks are then summed for each column, producing R_j . The values of R_j are squared and then summed. Because the study is concerned with five days of the week, five blocking levels are used and $b = 5$. The value of R_j is computed as shown in the following table.



Day	Supplier 1	Supplier 2	Supplier 3	Supplier 4
Monday	3	4	1	2
Tuesday	3	1	2	4
Wednesday	2	3	4	1
Thursday	3	1	2	4
Friday	3	4	1	2
R_j	14	15	10	13
R_j^2	196	225	100	169

$$\sum_{j=1}^4 R_j^2 = (196 + 225 + 100 + 169)$$

$$\chi_r^2 = \frac{12}{bc(c+1)} \sum_{j=1}^c R_j^2 - 3b(c+1)$$

ACTION:

STEP 7. Because the observed value of is greater than the critical value,
 $2.05,3 = 7.8147$, the decision is to reject the null hypothesis.

BUSINESS IMPLICATIONS:

STEP 8. Statistically, there is a significant difference in the tensile strength of housings made by different suppliers. The sample data indicate that supplier 3 is producing housings with a lower tensile strength than those made by other suppliers and that supplier 4 is producing housings with higher tensile strength. Further study by managers and a quality team may result in attempts to bring supplier 3 up to standard on tensile strength or perhaps cancellation of the contract.

Illustration :-

A market research company wants to determine brand preference for refrigerators. Five companies contracted with the research company to have their products be included in the study. As part of the study, the research company randomly selects 10 potential refrigerator buyers and shows them one of each of the five brands. Each survey participant is then asked to rank the refrigerator brands from 1 to 5. The results of these rankings are given in the table. Use the Friedman test and $\alpha = .01$ to determine whether there are any significant differences between the rankings of these brands.

Solution

HYPOTHESIZE:

STEP 1. The hypotheses are as follows.

H_0 : The brand populations are equal.

H_a : At least one brand population yields larger values than at least one other brand population.

TEST:

STEP 2. The market researchers collected ranked data that are ordinal in level. The Friedman test is the appropriate test.

STEP 3. Let $\alpha = .01$.

STEP 4. Because the study uses five treatment levels (brands), $c = 5$ and $df = 5 - 1 = 4$. The critical value is $2.01,4 = 13.2767$. If the observed chi-square is greater than 13.2767, the decision is to reject the null hypothesis.

STEP 5. The sample data follow.

STEP 6. The ranks are totaled for each column, squared, and then summed across the column totals. The results are shown in the table.

Individual	Brand A	Brand B	Brand C	Brand D	Brand E
1	3	5	2	4	1
2	1	3	2	4	5
3	3	4	5	2	1
4	2	3	1	4	5
5	5	4	2	1	3
6	1	5	3	4	2
7	4	1	3	2	5
8	2	3	4	5	1
9	2	4	5	3	1
10	3	5	4	2	1
R_j	26	37	31	31	25
R_j^2	676	1,369	961	961	625
					$\Sigma R_j^2 = 4,592$

The value of χ_r^2 is

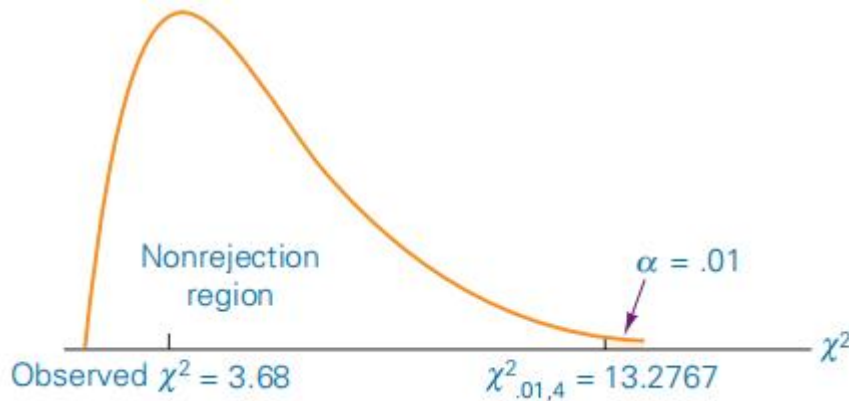
$$\chi_r^2 = \frac{12}{bc(c+1)} \sum_{j=1}^c R_j^2 - 3b(c+1) = \frac{12}{10(5)(5+1)} (4,592) - 3(10)(5+1) = 3.68$$

ACTION:

STEP 7. Because the observed value is 3.68 is not greater than the critical value, 2.01,4 = 13.2767, the researchers fail to reject the null hypothesis.

BUSINESS IMPLICATIONS:

STEP 8. Potential refrigerator purchasers appear to have no significant brand preference. Marketing managers for the various companies might want to develop strategies for positively distinguishing their product from the others.



Spearman's rank correlation

In Chapter 12, the Pearson product-moment correlation coefficient, r , was presented and discussed as a technique to measure the amount or degree of association between two variables. The Pearson r requires at least interval level of measurement for the data. When only ordinal-level data or ranked data are available, **Spearman's rank correlation**, r_s , can be used to analyze the degree of association of two variables. Charles E. Spearman (1863–1945) developed this correlation coefficient.

The formula for calculating a Spearman's rank correlation is as follows:

SPEARMAN'S RANK CORRELATION (17.7)

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

where

n = number of pairs being correlated

d = the difference in the ranks of each pair

Illustration :-

How strong is the correlation between crude oil prices and prices of gasoline at the pump? In an effort to estimate this association, an oil company analyst gathered the data shown over a period of several months. She lets crude oil prices be represented by the market value of a barrel of West Texas intermediate crude and gasoline prices be the estimated average price of regular unleaded gasoline in a certain city. She computes a Spearman's rank correlation for these data.

Crude Oil	Gasoline
\$14.60	\$3.25
10.50	3.26
12.30	3.28
15.10	3.26
18.35	3.32
22.60	3.44
28.90	3.56
31.40	3.60
26.75	3.54

Solution

Here, $n = 9$. When the analyst ranks the values within each group and computes the values of d and d^2 , she obtains the following.

Crude Oil	Gasoline	d	d^2
3	1	+2	4
1	2.5	-1.5	2.25
2	4	-2	4
4	2.5	+1.5	2.25
5	5	0	0
6	6	0	0
8	8	0	0
9	9	0	0
7	7	0	0
		$\Sigma d^2 = 12.5$	

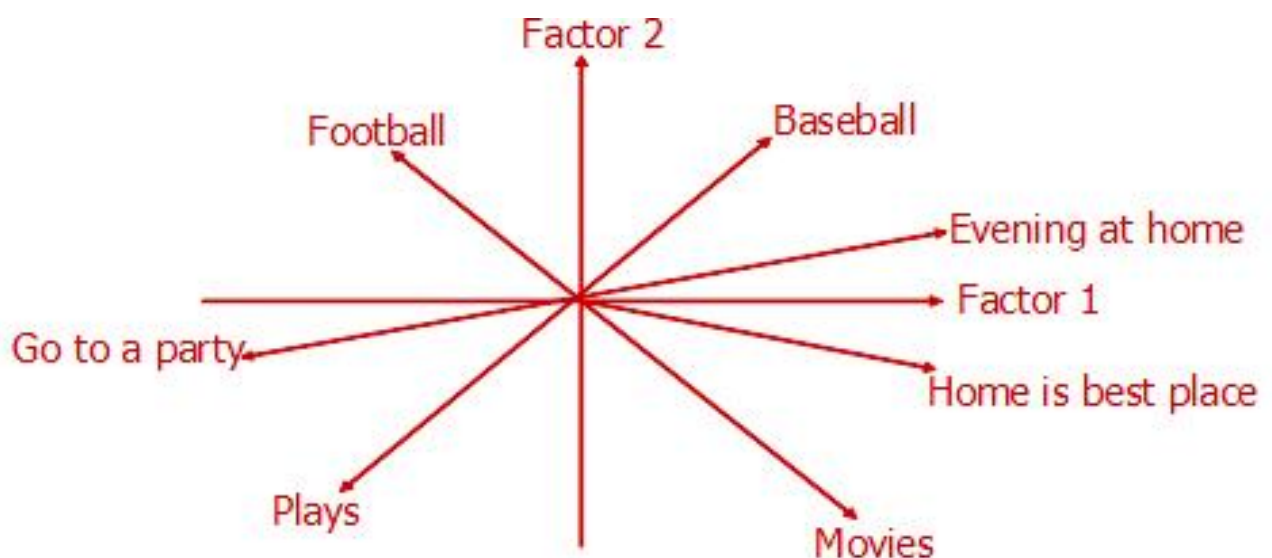
$$r_s = 1 - \frac{6 \Sigma d^2}{n(n^2 - 1)} = 1 - \frac{6(12.5)}{9(9^2 - 1)} = +.896$$

A high positive correlation is computed between the price of a barrel of West Texas intermediate crude and a gallon of regular unleaded gasoline.

Factor Analysis

- Factor analysis is a class of procedures used for data reduction and summarization.
- It is an interdependence technique: no distinction between dependent and independent variables.
- Factor analysis is used:
 - To identify underlying dimensions, or factors, that explain the correlations among a set of variables.
 - To identify a new, smaller, set of uncorrelated variables to replace the original set of correlated variables.

Factors Underlying Selected Psychographics and Lifestyles



Factor Analysis Model

Each variable is expressed as a linear combination of factors. The factors are some common factors plus a unique factor. The factor model is represented as:

$$X_i = A_{i1}F_1 + A_{i2}F_2 + A_{i3}F_3 + \dots + A_{im}F_m + V_iU_i$$

where

X_i = i th standardized variable

A_{ij} = standardized mult reg coeff of var i on common factor j

F_j = common factor j

V_i = standardized reg coeff of var i on unique factor i

U_i = the unique factor for variable i

m = number of common factors

Factor Analysis Model

The first set of weights (factor score coefficients) are chosen so that the first factor explains the largest portion of the total variance.

Then a second set of weights can be selected, so that the second factor explains most of the residual variance, subject to being uncorrelated with the first factor.

This same principle applies for selecting additional weights for the additional factors.

The common factors themselves can be expressed as linear combinations of the observed variables.

$$F_i = W_{i1}X_1 + W_{i2}X_2 + W_{i3}X_3 + \dots + W_{ik}X_k$$

Where:

F_i = estimate of i th factor

W_i = weight or factor score coefficient

k = number of variables

Statistics Associated with Factor Analysis

Bartlett's test of sphericity. Bartlett's test of sphericity is used to test the hypothesis that the variables are uncorrelated in the population (i.e., the population corr matrix is an identity matrix)

Correlation matrix. A correlation matrix is a lower triangle matrix showing the simple correlations, r , between all possible pairs of variables included in the analysis. The diagonal elements are all 1.

Communality. Amount of variance a variable shares with all the other variables. This is the proportion of variance explained by the common factors.

Eigenvalue. Represents the total variance explained by each factor.

Factor loadings. Correlations between the variables and the factors.

Factor matrix. A factor matrix contains the factor loadings of all the variables on all the factors

Factor scores. Factor scores are

composite scores estimated for each respondent on the derived factors.

Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy. Used to examine the appropriateness of factor analysis. High values (between 0.5 and 1.0) indicate appropriateness. Values below 0.5 imply not.

Percentage of variance. The percentage of the total variance attributed to each factor.

Scree plot. A scree plot is a plot of the Eigenvalues against the number of factors in order of extraction.

Example: Factor Analysis

HATCO is a large industrial supplier

A marketing research firm surveyed 100 HATCO customers, to investigate the customers' perceptions of HATCO

The marketing research firm obtained data on 7 different variables from HATCO's customers

Before doing further analysis, the mkt res firm ran a Factor Analysis to see if the data could be reduced

Example: Factor Analysis

In a B2B situation, HATCO wanted to know the perceptions that its customers had about it

The mktg res firm gathered data on 7 variables

1. Delivery speed
2. Price level
3. Price flexibility
4. Manufacturer's image
5. Overall service

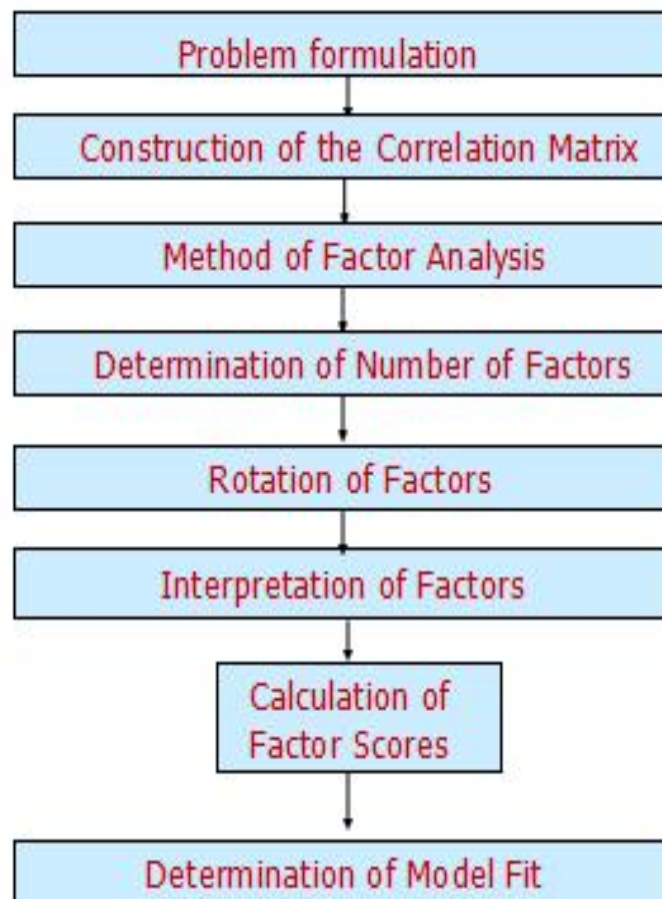
6. Salesforce image

7. Product quality

Each var was measured on a 10 cm graphic rating scale

Conducting Factor Analysis

Fig. 19.2



Formulate the Problem

The objectives of factor analysis should be identified.

The variables to be included in the factor analysis should be specified.
The variables should be measured on an interval or ratio scale.

An appropriate sample size should be used. As a rough guideline, there should be at least four or five times as many observations (sample size) as there are variables.

Construct the Correlation Matrix

The analytical process is based on a matrix of correlations between the variables.

If the Bartlett's test of sphericity is not rejected, then factor analysis is not appropriate.

If the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy is small, then the correlations between pairs of variables cannot be explained by other variables and factor analysis may not be appropriate.

Determine the Method of Factor Analysis

- In Principal components analysis, the total variance in the data is considered.

-Used to determine the min number of factors that will account for max variance in the data.

- In Common factor analysis, the factors are estimated based only on the common variance.

-Communalities are inserted in the diagonal of the correlation matrix.

-Used to identify the underlying dimensions and when the common variance is of interest.

Rotation of Factors

Through rotation the factor matrix is transformed into a simpler one that is easier to interpret.

After rotation each factor should have nonzero, or significant, loadings for only some of the variables. Each variable should have nonzero or significant loadings with only a few factors, if possible with only one.

The rotation is called orthogonal rotation if the axes are maintained at right angles.

Varimax procedure. Axes maintained at right angles

-Most common method for rotation.

-An orthogonal method of rotation that minimizes the number of variables with high loadings on a factor.

-Orthogonal rotation results in uncorrelated factors.

Oblique rotation. Axes not maintained at right angles

-Factors are correlated.

-Oblique rotation should be used when factors in the population are likely to be strongly correlated.

Interpret Factors

A factor can be interpreted in terms of the variables that load high on it.

Another useful aid in interpretation is to plot the variables, using the factor loadings as coordinates. Variables at the end of an axis are those that have high loadings on only that factor, and hence describe the factor.

Discriminant Analysis and Classification

Discriminant Analysis as a Type of MANOVA

The good news about DA is that it is a lot like MANOVA; in fact in the case of a factor with only two levels it is the same thing

Has the same assumptions as MANOVA; multivariate normality, independence of cases, homogeneity of group covariances

DA permits a multivariate analysis of variance hypothesis of the test that two or more groups (conditions, levels) differ significantly on a linear combination of discriminating variables. Another way to put this is: how well can the levels of the *grouping variable* be discriminated by scores on the *discriminating variables*?

In general it's good to use naturally occurring groups that are mutually exclusive groups that are exhaustive of the domain, rather than median splits or arbitrary divisions

The good news about DA is that it is a lot like MANOVA; in fact in the case of a factor with only two levels it is the same thing

Has the same assumptions as MANOVA; multivariate normality, independence of cases, homogeneity of group covariances

DA permits a multivariate analysis of variance hypothesis of the test that two or more groups (conditions, levels) differ significantly on a linear combination of discriminating variables. Another way to put this is: how well can the levels of the *grouping variable* be discriminated by scores on the *discriminating variables*?

In general it's good to use naturally occurring groups that are mutually exclusive groups that are exhaustive of the domain, rather than median splits or arbitrary divisions

In the case where there are more than two groups, DA permits you to test the hypothesis that there is **more than one significant way** of describing how the groups differ on a weighted linear combination of the discriminating variables, and you can think of these combinations, called *canonical variables*, as "dimensions" of difference. These variables will be uncorrelated with each other

This way of using DA is called *descriptive discriminant analysis*

Usually discriminant analysis is presented conceptually in an upside down sort of way, where what you would traditionally think of as dependent variables are actually the predictor variables, and group membership rather than being the levels of the IV are groups whose membership is being predicted

When it is used in this way, the hypothesis you are testing is that there is a linear combination of variables which when appropriately weighted (like beta weights) will maximally discriminate between members of two or more groups and permit new cases to be classified into the groups

In this mode, called *predictive discriminant analysis*, DA is used to develop a classification rule that will permit things like classifying people as potential Republican voters or not, or to predict their future status as able to complete four years of college or not, or to be able to pay their car loan

Discriminant analysis is part of the general linear model and combines some of the features familiar to you from multiple regression and some from MANOVA. It's basically multiple regression where the criterion variable is nominal rather than interval/ratio level

When DA is used in this predictive way it is usually followed up by classification procedures to classify new cases based on the obtained discriminant function(s)

Let's work through an example of discriminant analysis, and show how it can approach a question from two sides: testing a MANOVA hypothesis and predicting group membership

First let's consider the hypothesis that a nation's level of concentration of wealth (in the hands of a few, more widely distributed, or somewhere in between) has a significant impact on four dependent variables: human development score, political rights score, the gini (inequality) index, and civil liberties score

Note. In creating these three wealth concentration "groups" out of interval level data I am not advocating this practice but only creating "groups" for purposes of illustration. Naturally occurring, clearly separated groups, e.g., males and females, people who survived after five years of diagnosis and people who didn't) are preferred for the grouping variable

This sounds like a hypothesis that could be tested with MANOVA, and it is, but it can also be tested with discriminant analysis

First let's look at what MANOVA will tell us about this hypothesis

What is Multidimensional

Scaling [MDS] ?

- **MDS** (aka “Smallest Space Analysis”)
- Has origins in Psychometrics in 1920-’60s:
- Scale construction and dimensionality reduction
- Underwent major burst of development in 1960s due to “non-metric revolution”(Coombs) and computing developments allowing iterative estimation
- Originally designed for analysis of LTM of dis/similarities data , taking a range of measures (not just PM correlations):
- “*anything which, by an act of faith, can be considered a similarity*” (Shepard)
- Extended rapidly to deal with wide range of other types of data
- Rectangular matrices ; triads, pair-comparisons, free-sorting
- “stacks” of matrices (3-way scaling – INDSCAL)
- Given a map, it’s easy to calculate the distances between the points ...
- MDS operates the other way round:
- **Given the data** [interpreted as quasi “distances”] **it attempts to find the configuration** [location of points] **which generated the distances**
- This is “Classic MDS”: developed in 1930s – but imperfect, not robust, & works only if data are ratio.
- Whereas more recent MDS can work when **only the ordinal information** exists: “Non-metric” = ordinal MDS (Coombs / Kruskal “*non-metric revolution*”)
- What?? You can create an accurate map from only the *rank* –order of the distances???

A student’s definition:

If you are interested in how certain objects relate to each other ... and if you would like to present these relationships in the form of a map then MDS is the technique you need” (Mr Gawels, KUB) A good start!

MDS provides ...

- a useful and easily-assimilable graphic visualisation of all sorts of data
- Tukey: “*A picture is worth a thousand words*”
- In a user-chosen (small) # of dimensions
- providing a graphical representation of the structure underlying a complex data set
- And measure how well / badly the solution distances match the data dissimilarities (Stress)

MDS is a family of models differentiated by ...

- **(DATA)** the empirical inter-relationships between a set of “objects”/variables which are given in a set of dis/similarity data
- Basically, type of input data, defined by their “Way” and “Mode” [e.g. 2W1M]. (Cf observations vs data)
- **(FUNCTION)** data are then optimally re-scaled (according to permissible transformations for the data) in terms of ...
- Choice of level of measurement [e.g. ordinal]
- **(MODEL)** the assumptions of the model chosen to represent the data
- Usually (Euclidean) Distance model
- **MDS** can be used with a wide variety of **DATA**
- e.g.: **SORTS OF DATA**
- **direct** data (pair comparisons, ratings, rankings, triads, counts)
- **derived** data (profiles, co-occurrence matrices, textual data, aggregated data)
- **measures** of association etc derived from simpler data, and **tables** of data.

TYPES of DATA

- Described by WAY (2W=matrix; 3W=stack of matrices ...)
- And MODE (# sets of distinct objects – eg variables, subjects)
- E.G. 2W1M; 2W2M; 3W2M ... 7W4M

