



# Statistics

# A Scenario



Here are scores of two Students Shweta and Sushma in the last five tests:

<b>Shweta:</b>	56	40	45	78	89
<b>Sushma:</b>	35	92	78	55	48

**Who is a better performer?**

- Does the averages of their respective scores help in identifying the better performer?

Average score for Shweta: 61.6

Average score for Sushma: 61.6

- Let us compute one more statistic - Range

Score Range for Shweta:  $89-40 = 49$

Score Range for Sushma:  $92-35 = 57$

**Now, who is a better performer?**

# What is Statistics



**Statistics** is a Mathematical Science pertaining to:

- a) Collection
- b) Analysis
- c) Interpretation or explanation
- d) Presentation of data

# Which of the following are Statistical Questions



a. How many days are in March?

Not statistical. This question is answered by counting the number of days in March. This produces a single number. This question is not answered by collecting data that vary.

b. How old is your dog?

Not statistical. This question is answered by a single number. It is not answered by collecting data that vary.

c. On average, how old are the dogs that live on this street?

Statistical. This question would be answered by collecting data, and there would be variability in that data.

d. What proportion of the students at your school like watermelons?

Statistical. This question would be answered by collecting data, and there would be variability in that data.

e. Do you like watermelons?

Not statistical. This question is answered by a single response. It is not answered by collecting data that vary.

f. How many bricks are in this wall?

Not statistical. This question would be answered by counting the bricks. This produces a single number. This question is not answered by collecting data that vary.

g. What was the temperature at noon today at City Hall?

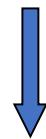
Non-statistical (there is one temperature).

# Types of Statistics



- Statistics is broadly classified into two types.

- Descriptive.
- Inferential.



## **Descriptive Statistics**

Collecting, summarizing, and  
describing data



## **Inferential Statistics**

Drawing conclusions and/or making  
decisions concerning a population based  
only on sample data

# Descriptive Statistics



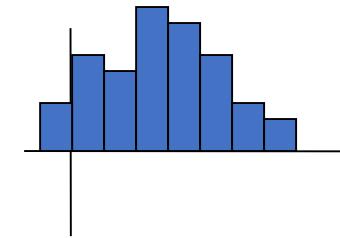
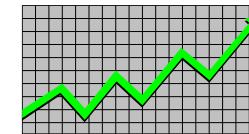
- Collect data

- e.g., Survey



- Present data

- e.g., Tables and graphs



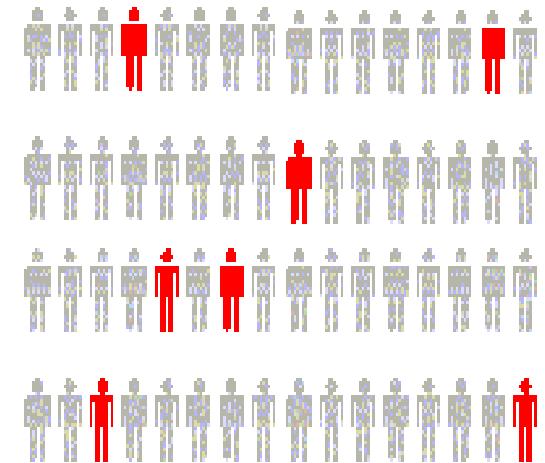
- Characterize data

- e.g., Sample mean =  $\frac{\sum X_i}{n}$

# Inferential Statistics



- Estimation
  - e.g., Estimate the population mean weight using the sample mean weight
- Hypothesis testing
  - e.g., Test the claim that the population mean weight is 120 pounds



**Drawing conclusions about a large group of individuals  
based on a subset of the large group.**

# Following questions are Inferential or Descriptive?



A recent study examined the math and verbal SAT scores of high school seniors across the country. Which of the following statements are descriptive in nature and which are inferential.

- a. The mean math SAT score was 492.

Descriptive

- b. The mean verbal SAT score was 475.

Descriptive

- c. Students in the Northeast scored higher in math but lower in verbal.

Inferential

- d. 80% of all students taking the exam were headed for college.

Inferential

- e. 32% of the students scored above 610 on the verbal SAT.

Inferential

- f. The math SAT scores are higher than they were 10 years ago.

Inferential

# Terminologies Used



## VARIABLE

- A **variable** is any characteristics, number, or quantity that can be measured or counted.

## DATA

- **Data** are the different values associated with a variable.

## OPERATIONAL DEFINITIONS

- Data values are meaningless unless their variables have **operational definitions**, universally accepted meanings that are clear to all associated with an analysis.

## POPULATION

- A **population** consists of all the items or individuals about which you want to draw a conclusion.

## SAMPLE

- A **sample** is the portion of a population selected for analysis.

## PARAMETER

- A **parameter** is a numerical measure that describes a characteristic of a population.

## STATISTIC

- A **statistic** is a numerical measure that describes a characteristic of a sample.

# Example



A college dean is interested in learning about the average age of faculty. Identify the basic terms in this situation.

**Variable** - “age” of each faculty member.

**Population** - Total number of faculty members at the college.

**Sample** - Subset of that population. E.g. Select 10 faculty members and determine their age.

**Experiment** – A method used to select the ages forming the sample and determining the actual age of each faculty member in the sample.

**Parameter** - “average” age of all faculty at the college(Population).

**Statistic** - “average” age for all faculty in the sample.

# Types of Variables

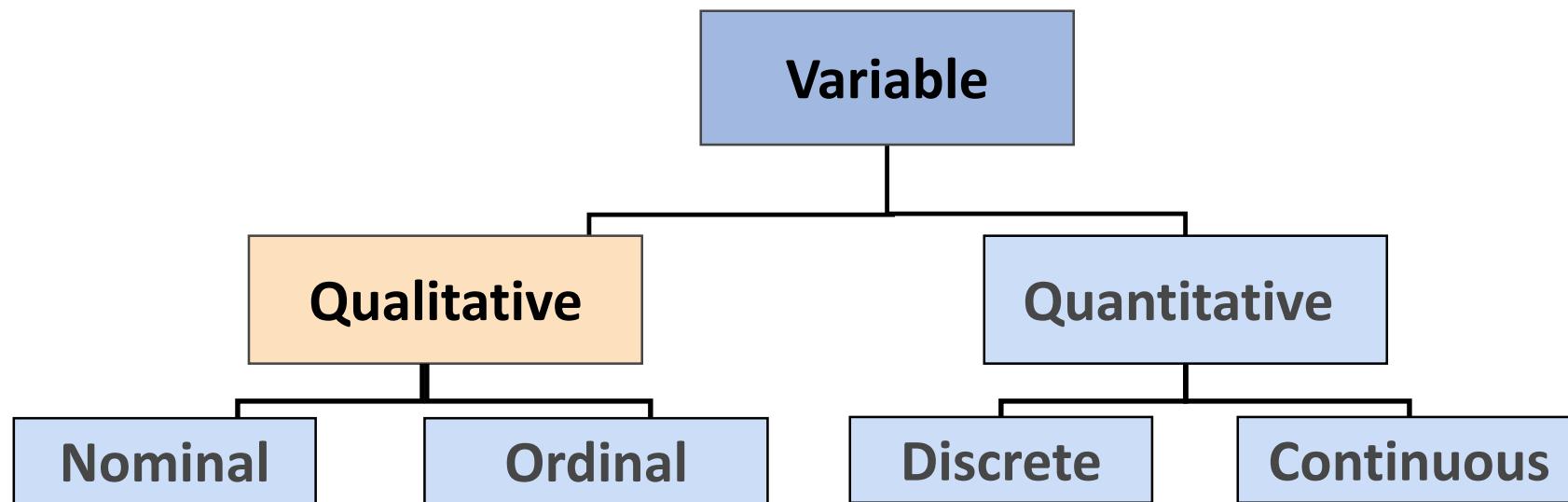


**Qualitative or Attribute or Categorical, Variable:** A variable that categorizes or describes an element of a population.

**Note:** Arithmetic operations, such as addition and averaging, are not meaningful for data resulting from a qualitative variable.

**Quantitative, or Numerical, Variable:** A variable that quantifies an element of a population.

**Note:** Arithmetic operations such as addition and averaging, are meaningful for data resulting from a quantitative variable.



# Identify each of the following as Qualitative or Quantitative

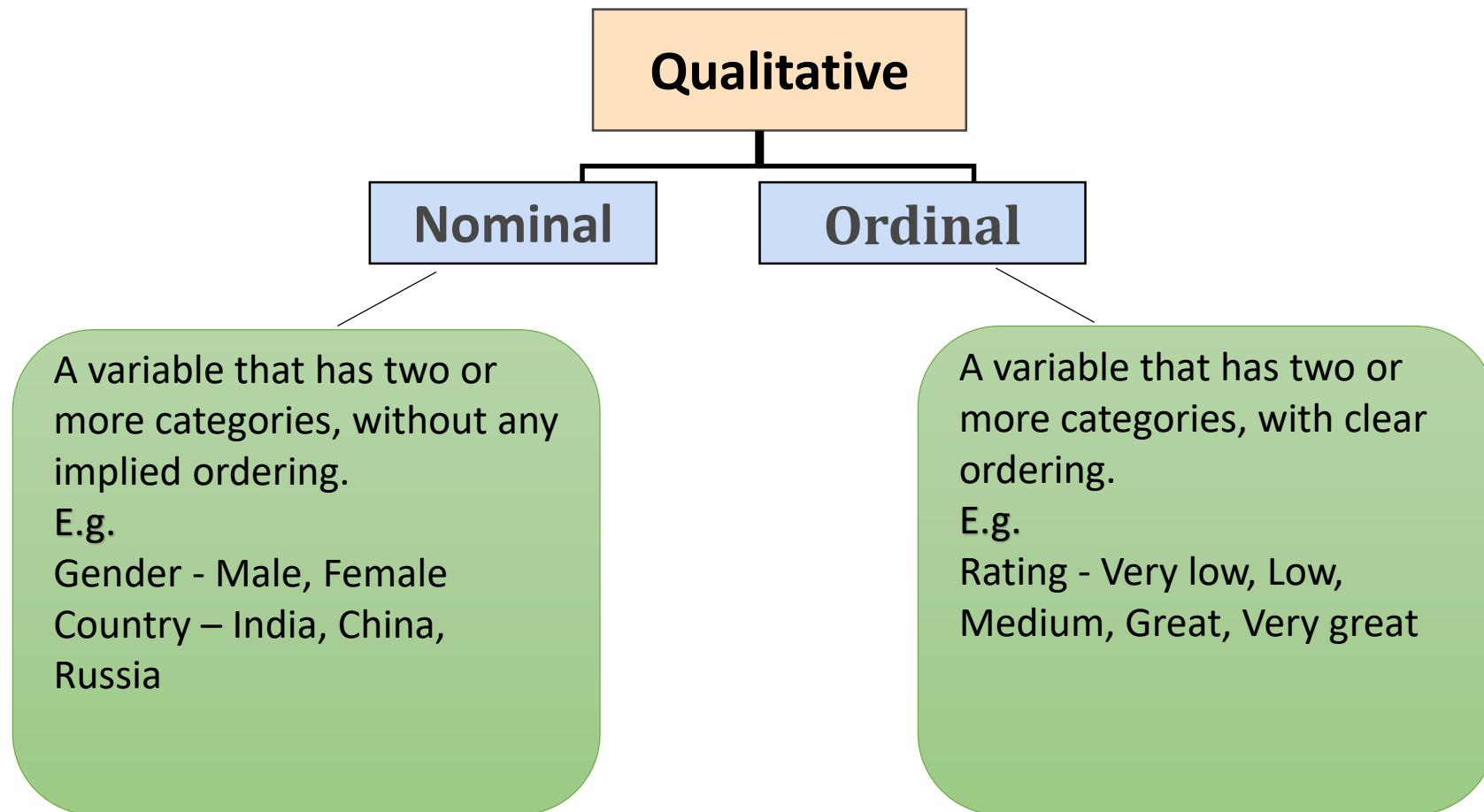


1. The residence hall for each student in a statistics class.  
**(Qualitative)**
2. The amount of gasoline pumped by the next 10 customers at the local Unimart.  
**(Quantitative)**
3. The amount of radon in the basement of each of 25 homes in a new development.  
**(Quantitative)**
4. The color of the baseball cap worn by each of 20 students.  
**(Qualitative)**
5. The length of time to complete a mathematics homework assignment.  
**(Quantitative)**
6. The state in which each truck is registered when stopped and inspected at a weigh station.  
**(Qualitative)**

# Qualitative Variable



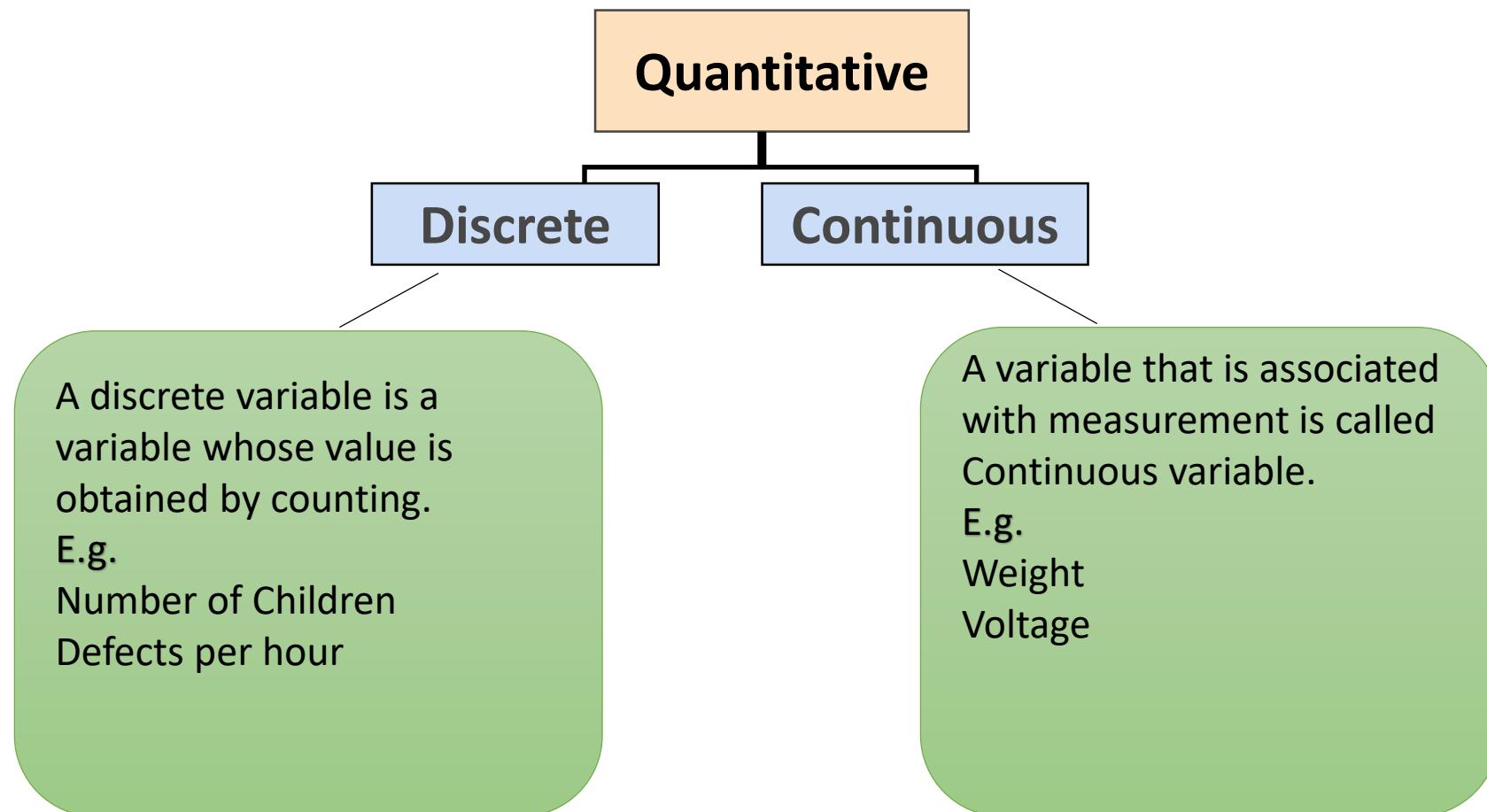
Data that can be added into categories



# Quantitative Variable



Data that quantifies an element of the population.



# Identify each of the following as examples of Nominal, Ordinal, Discrete or Continuous Variable



1. The length of time until a pain reliever begins to work.

**Continuous**

2. The number of chocolate chips in a cookie.

**Discrete**

3. The number of colors used in a statistics textbook.

**Discrete**

4. The brand of refrigerator in a home.

**Nominal**

5. The overall satisfaction rating of a new car.

**Ordinal**

6. The number of files on a computer's hard disk.

**Discrete**

7. The pH level of the water in a swimming pool.

**Continuous**

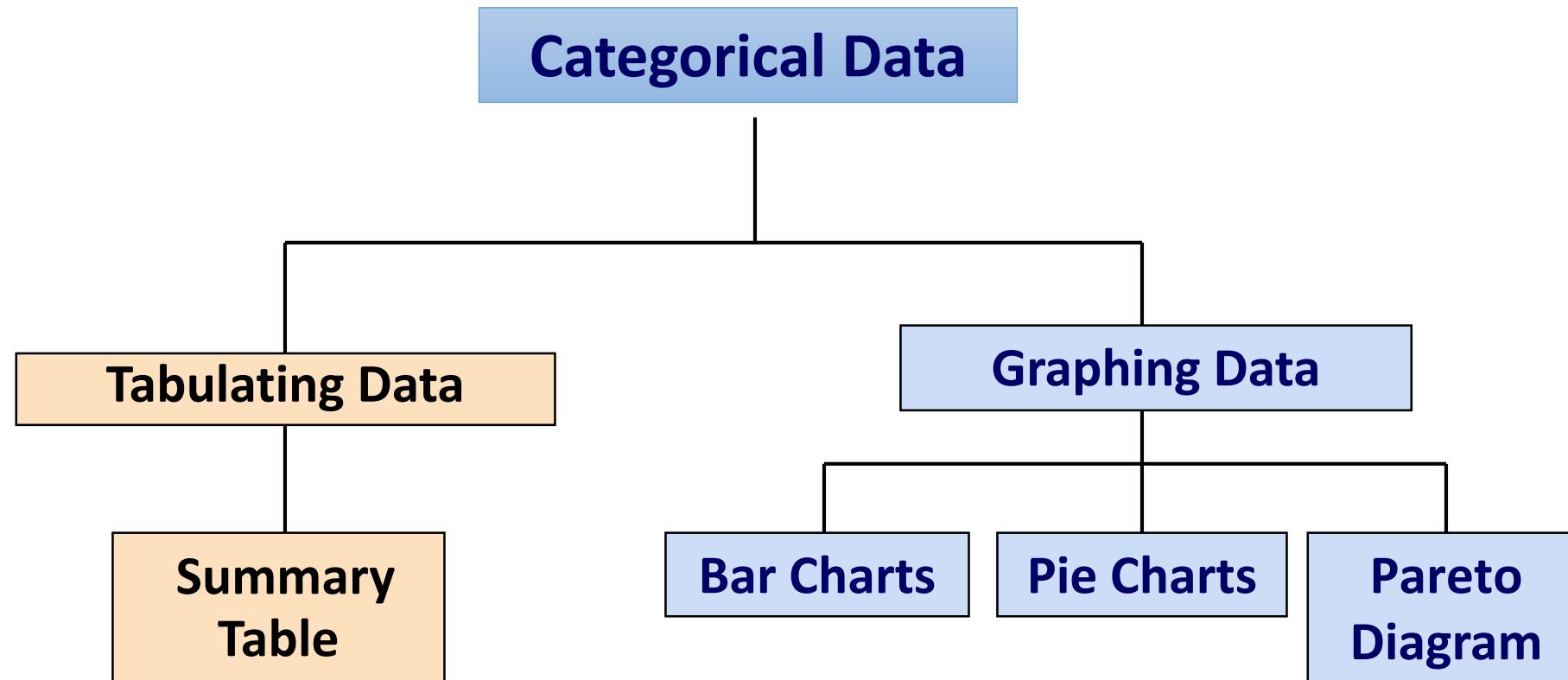
8. The number of staples in a stapler.

**Discrete**

# Presenting Categorical Data in Tables & Graphs



- **Categorical** data are summarized by Tables & Graphs.



# Summary Table



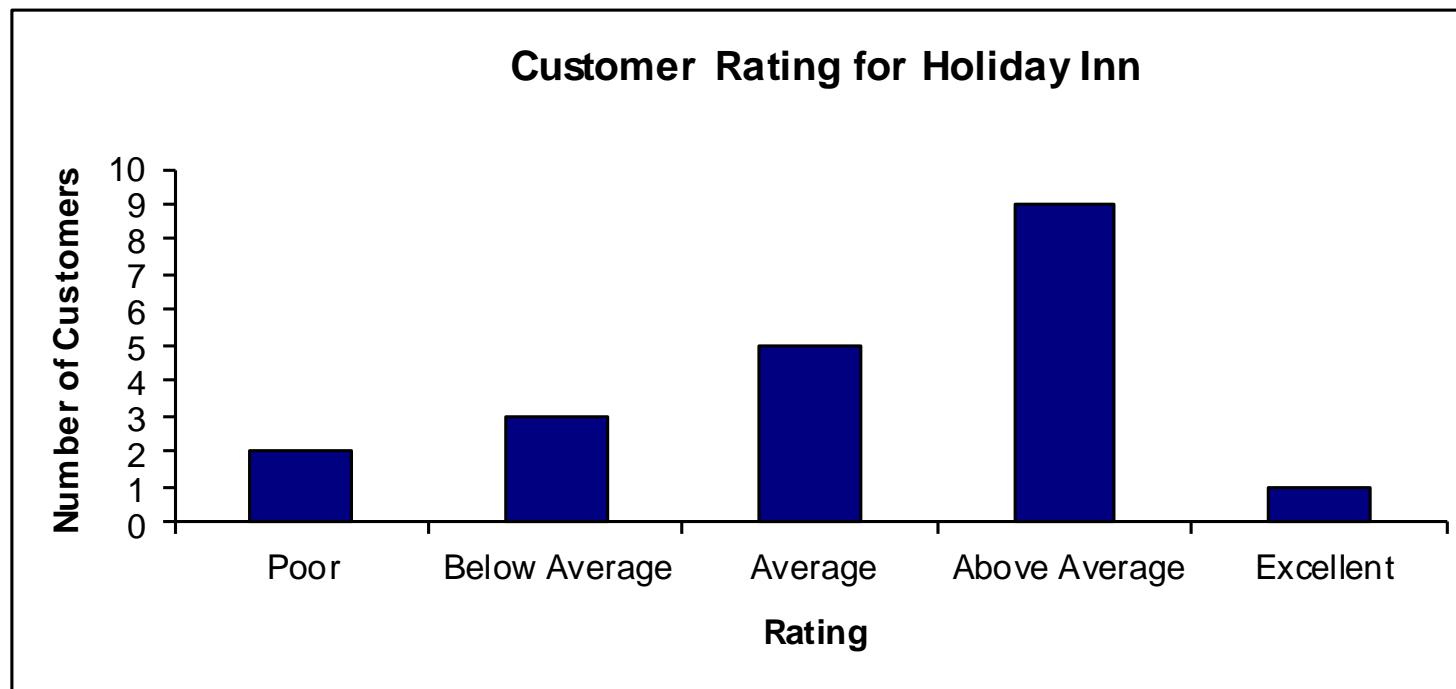
A **summary table** indicates the frequency, amount, or percentage of items in a set of categories so that you can see differences between categories.

Banking Preference?	Percent
ATM	16%
Automated or live telephone	2%
Drive-through service at branch	17%
In person at branch	41%
Internet	24%

# Bar Chart



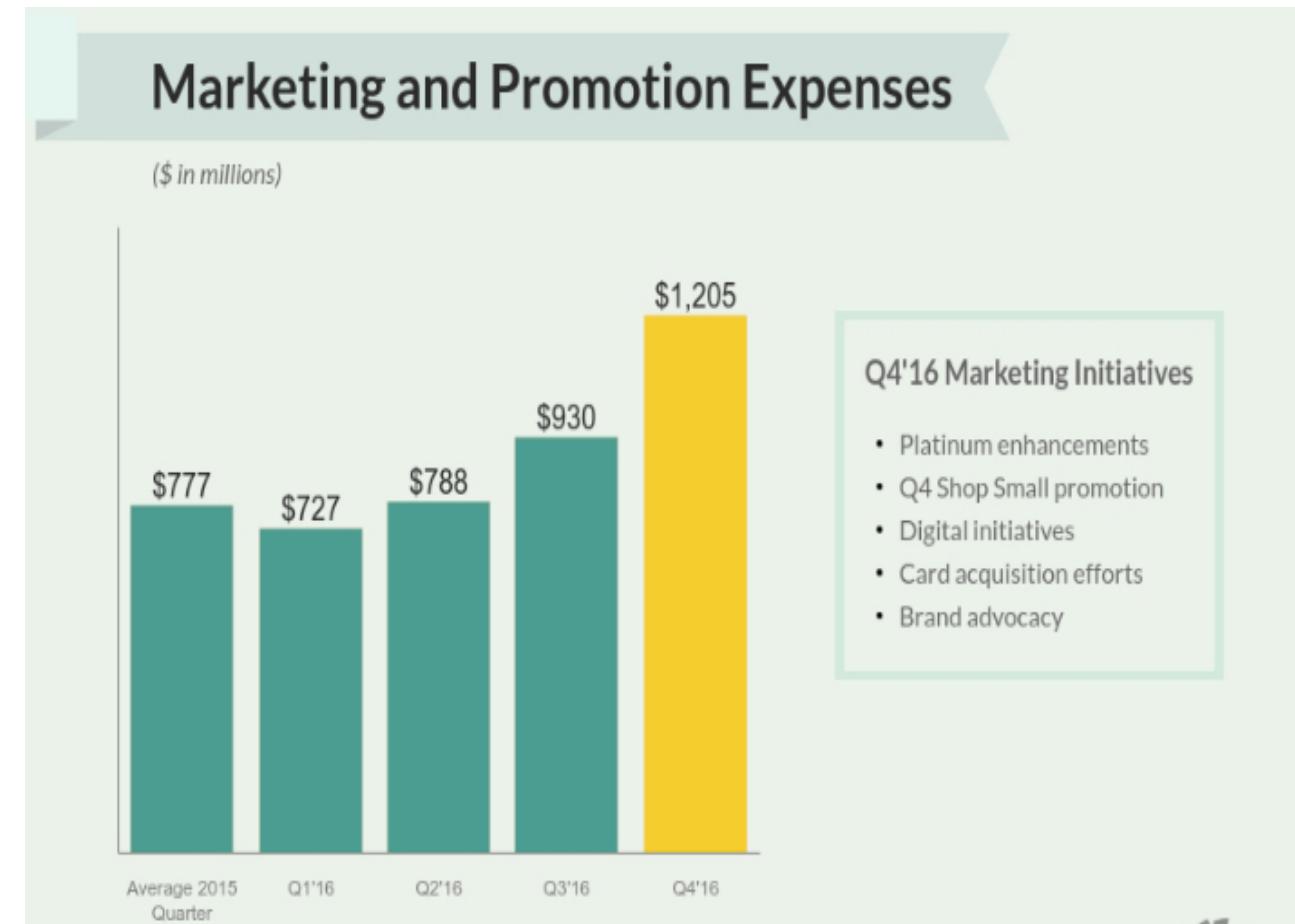
- A bar graph is a tool for depicting categorical data and show comparison between the categories
- One axis of the chart specifies categories being compared and other axis represents the discreet value or frequency.
- The length of the bars are proportional to the values that they represent.



# Common Use



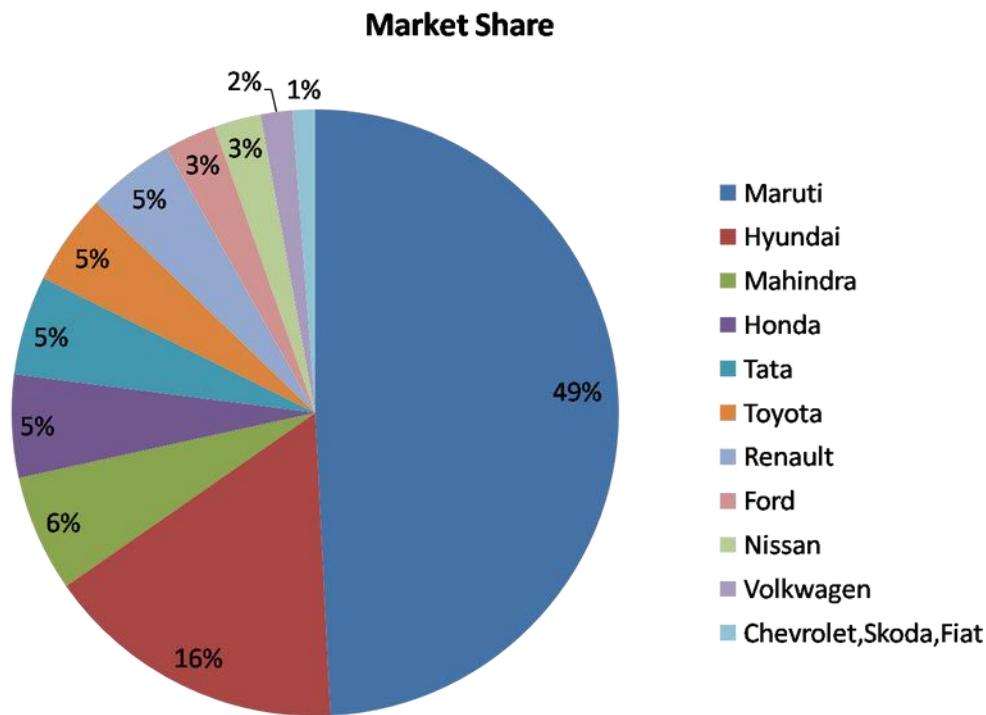
- Holds particular power in the marketing industry.
- The charts are commonly used to present **financial forecasts and outcomes**,
- Ideal for comparing any sort of numeric value, including group sizes, inventories, ratings and survey responses.



# Pie Chart



- The **pie chart** is a circle broken up into slices that represent categories.
- The size of each slice of the pie varies according to the percentage in each category.



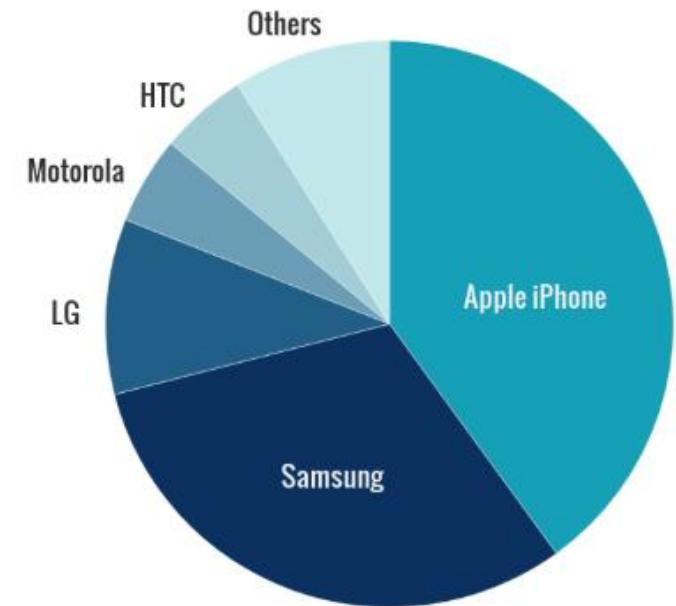
# Common Use



- Compare the size of market segments.
- A simple pie graph can clearly illustrate how the most popular mobile-phone manufacturers compare based on the sizes of their user-bases.
- Audiences can quickly understand that Apple and Samsung hold almost 75-percent of the mobile-communication market, with Apple slightly ahead.

**Smartphone Brand Market Share**

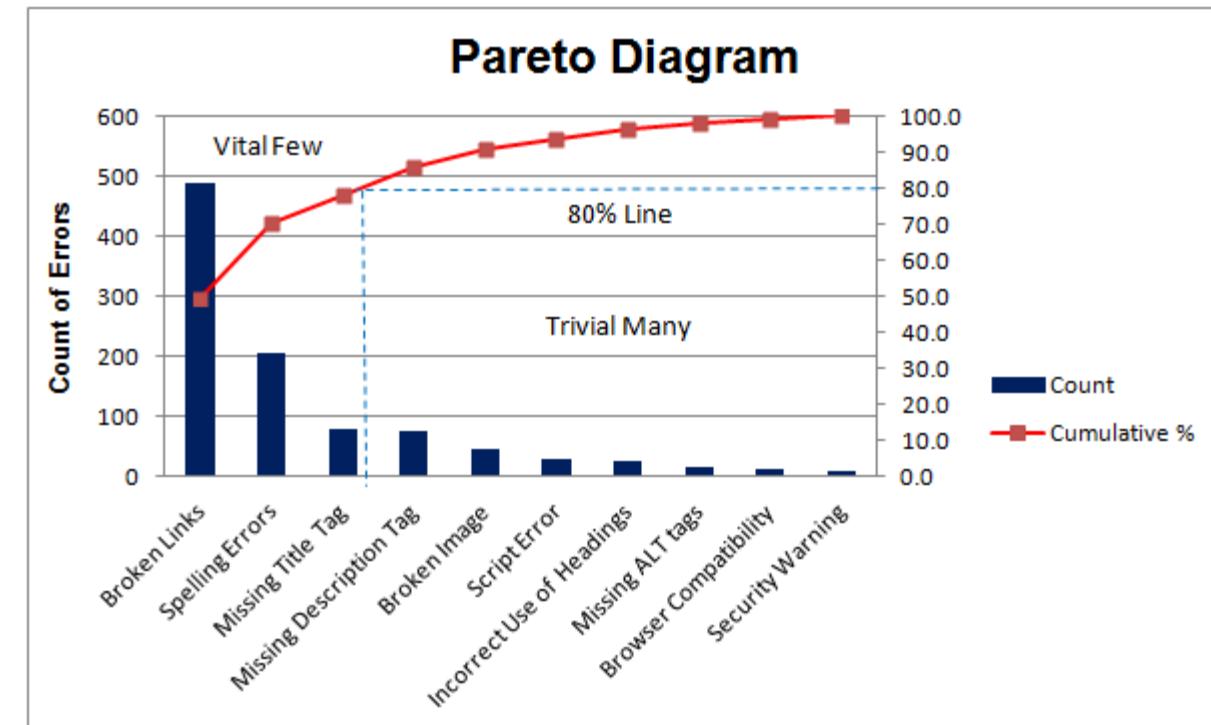
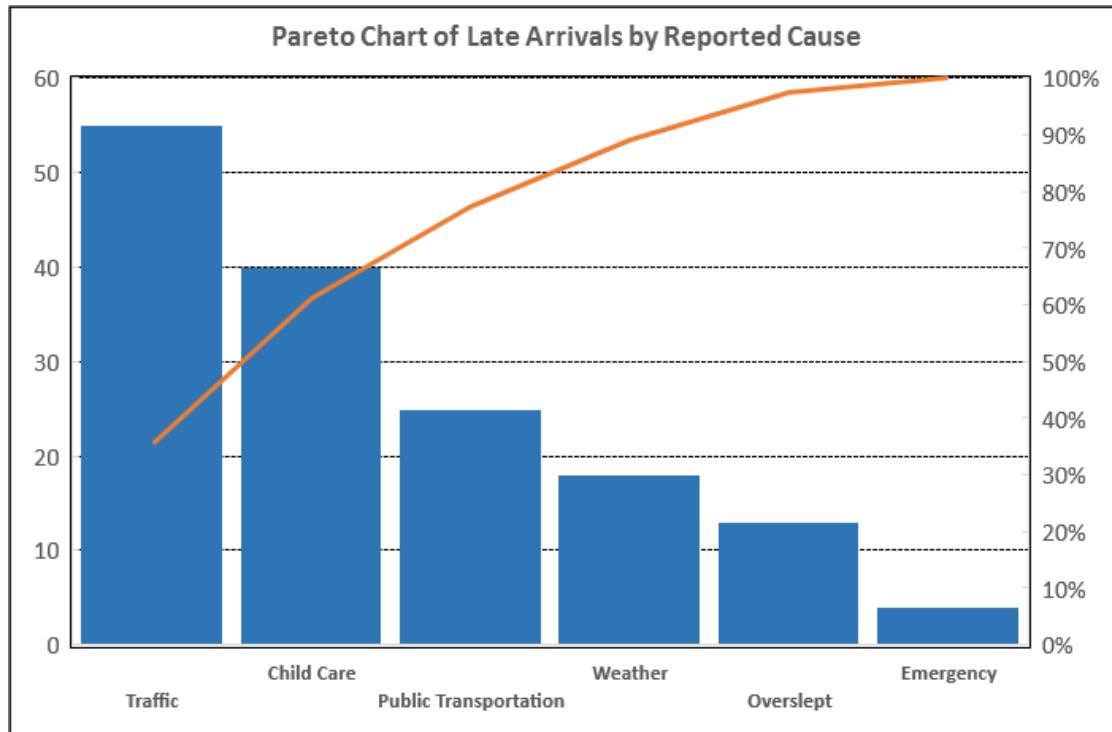
Smartphone Owners in U.S.  
Broadband Households



# Pareto Diagram



- Used to portray categorical data (nominal scale).
- A vertical bar chart, where categories are shown in descending order of frequency.
- A cumulative polygon is shown in the same graph.
- Used to separate the “vital few” from the “trivial many”.

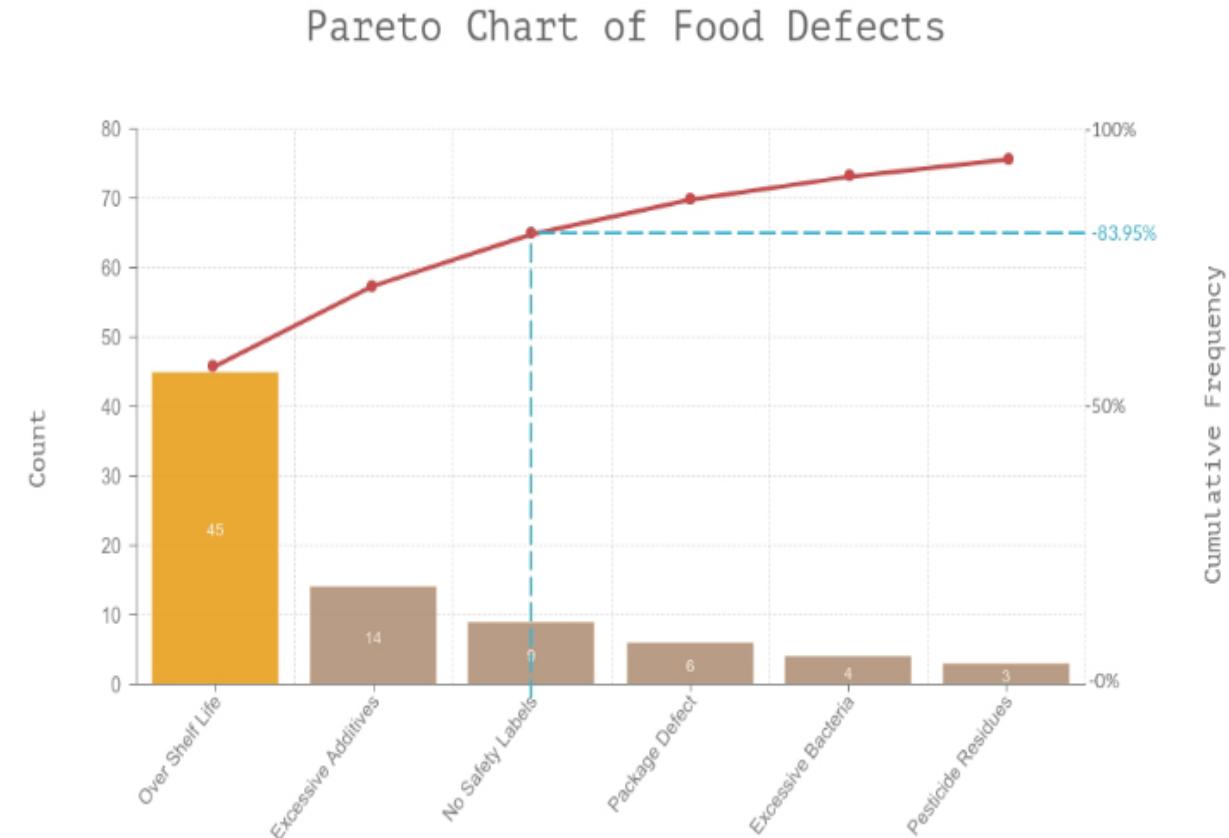


# Common Use

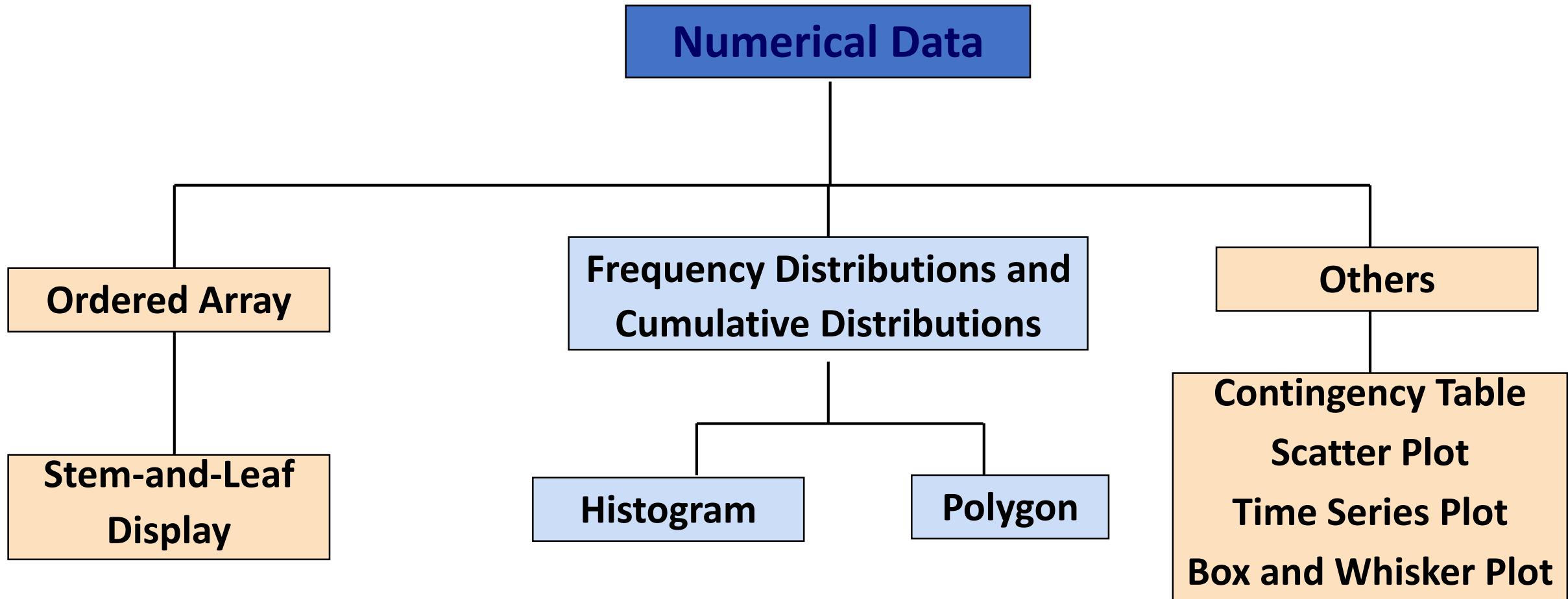


Designed to highlight the most important of a set of factors.

**For example,** in a Pareto chart that tracks the type and frequency of food defects, the bars illustrate each type of defects' total occurrences — as reported on one of the charts' axes, while the line charts the cumulative frequency of all categories, from most to least prevalent. The result is a graph that clearly reflects the most common food defects and what percentage of the whole each represents.



# Presenting Numerical Data in Tables & Graphs



# Ordered Array



- An ordered array is a sequence of data, in rank order, from the smallest value to the largest value.
- Shows range (minimum value to maximum value)
- May help identify outliers (unusual observations)

## Ordered Array

98	80
94	80
92	78
90	75
86	72
84	70
83	68
82	62
82	58
82	36

Scores are listed in order, typically from highest to lowest.

# Stem-and-Leaf Display



- A simple way to see how the data are distributed and where concentrations of data exist.
- A **stem-and-leaf display** organizes data into groups (called stems) so that the values within each group (the leaves) branch out to the right on each row.
- The display is achieved by separating the sorted data series into leading digits (the stems) and the trailing digits (the leaves).

Age of Surveyed College Students	Day Students					
	16	17	17	18	18	18
	19	19	20	20	21	22
	22	25	27	32	38	42
	Night Students					
	18	18	19	19	20	21
	23	28	32	33	41	45

Age of College Students						
	Day Students			Night Students		
	Stem	Leaf	Stem	Leaf		
	1	67788899				
	2	0012257				
	3	28				
	4	2				

# Frequency Distribution



- The frequency distribution is a summary table in which the data are arranged into numerically ordered classes.
- You must give attention to selecting the appropriate number of class groupings for the table, determining a suitable width of a class grouping, and establishing the boundaries of each class grouping to avoid overlapping.
- The number of classes depends on the number of values in the data. With a larger number of values, typically there are more classes. In general, a frequency distribution should have at least 5 but no more than 15 classes.
- To determine the width of a class interval, you divide the range (Highest value–Lowest value) of the data by the number of class groupings desired.
- Example: A manufacturer of insulation randomly selects 20 winter days and records the daily high temperature

**24, 35, 17, 21, 24, 37, 26, 46, 58, 30, 32, 13, 12, 38, 41, 43, 44, 27, 53, 27**

# Frequency Distribution - Example



- Sort raw data in ascending order:  
12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35,  
37, 38, 41, 43, 44, 46, 53, 58
- Find range:  $58 - 12 = 46$
- Select number of classes: 5 (usually between 5 and 15)
- Compute class interval (width): 10 ( $46/5$  then round up)
- Determine class boundaries (limits):
  - Class 1: 10 to less than 20
  - Class 2: 20 to less than 30
  - Class 3: 30 to less than 40
  - Class 4: 40 to less than 50
  - Class 5: 50 to less than 60
- Compute class midpoints: 15, 25, 35, 45, 55
- Count observations & assign to classes

**Data in ordered array:**

**12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58**

Class	Frequency	Relative Frequency	Percentage
<b>10 but less than 20</b>	3	.15	15
<b>20 but less than 30</b>	6	.30	30
<b>30 but less than 40</b>	5	.25	25
<b>40 but less than 50</b>	4	.20	20
<b>50 but less than 60</b>	2	.10	10
<b>Total</b>	<b>20</b>	<b>1.00</b>	<b>100</b>

# Frequency Distribution – Benefits & Tips



- **Benefits:**
  - It condenses the raw data into a more useful form.
  - It allows for a quick visual interpretation of the data.
  - It enables the determination of the major characteristics of the data set including where the data are concentrated / clustered.
- **Tips:**
  - Different class boundaries may provide different pictures for the same data (especially for smaller data sets).
  - Shifts in data concentration may show up when different class boundaries are chosen.
  - As the size of the data set increases, the impact of alterations in the selection of class boundaries is greatly reduced.
  - When comparing two or more groups with different sample sizes, you must use either a relative frequency or a percentage distribution.

# Cumulative Frequency Distribution



Data in ordered array:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Class	Frequency	Percentage	Cumulative Frequency	Cumulative Percentage
10 but less than 20	3	15	3	15
20 but less than 30	6	30	9	45
30 but less than 40	5	25	14	70
40 but less than 50	4	20	18	90
50 but less than 60	2	10	20	100
Total	20	100		

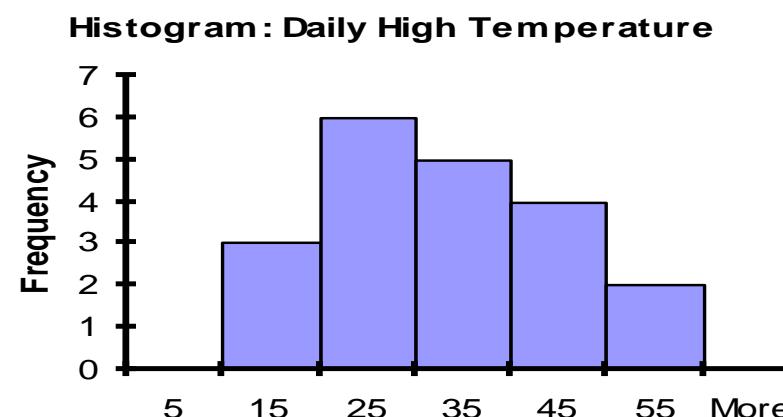
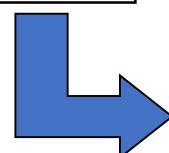
# Histogram



- A vertical bar chart of the data in a frequency distribution is called a histogram.
- It differs from a bar graph, in the sense that a bar graph relates two variables, but a histogram relates only one.
- The vertical axis is either frequency, relative frequency, or percentage.
- The height of the bars represent the frequency, relative frequency, or percentage.

Class	Frequency	Relative Frequency	Percentage
10 but less than 20	3	.15	15
20 but less than 30	6	.30	30
30 but less than 40	5	.25	25
40 but less than 50	4	.20	20
50 but less than 60	2	.10	10
Total	20	1.00	100

(In a percentage histogram  
the vertical axis would be  
defined to show the  
percentage of observations  
per class)



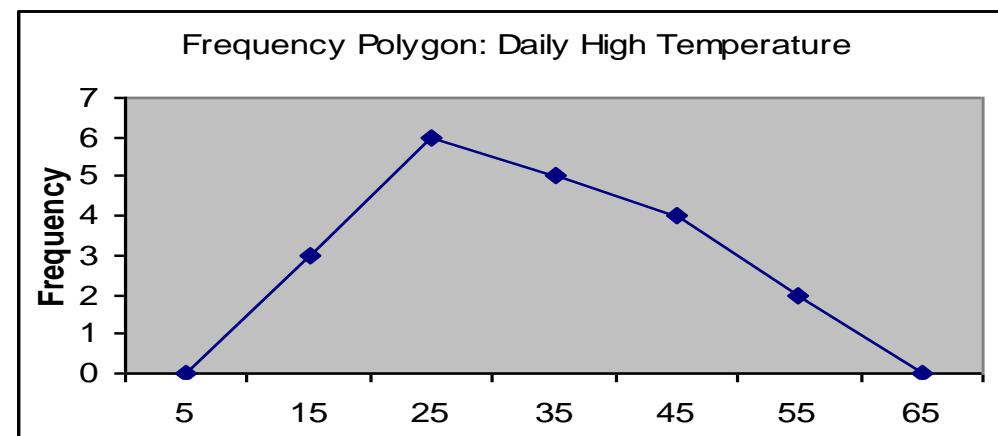
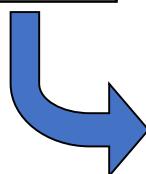
# Polygon



- A percentage polygon is formed by having the midpoint of each class represent the data in that class and then connecting the sequence of midpoints at their respective class percentages.
- The cumulative percentage polygon, or ogive, displays the variable of interest along the X axis, and the cumulative percentages along the Y axis.
- Useful when there are two or more groups to compare.

Class	Class Midpoint	Frequency
10 but less than 20	15	3
20 but less than 30	25	6
30 but less than 40	35	5
40 but less than 50	45	4
50 but less than 60	55	2

(In a percentage polygon the vertical axis would be defined to show the percentage of observations per class)



# Cross Tabulation – Contingency Table



- Used to study patterns that may exist between two or more categorical variables.
- Cross tabulations can be presented in:
  - Tabular form -- Contingency Tables
  - A cross-classification (or contingency) table presents the results of two categorical variables. The joint responses are classified so that the categories of one variable are located in the rows and the categories of the other variable are located in the columns.
  - The cell is the intersection of the row and column and the value in the cell represents the data corresponding to that specific pairing of row and column categories.
  - A useful way to visually display the results of cross-classification data is by constructing a side-by-side bar chart.

# Cross Tabulation – Contingency Table



A survey was conducted to study the importance of brand name to consumers as compared to a few years ago. The results, classified by gender, were as follows:

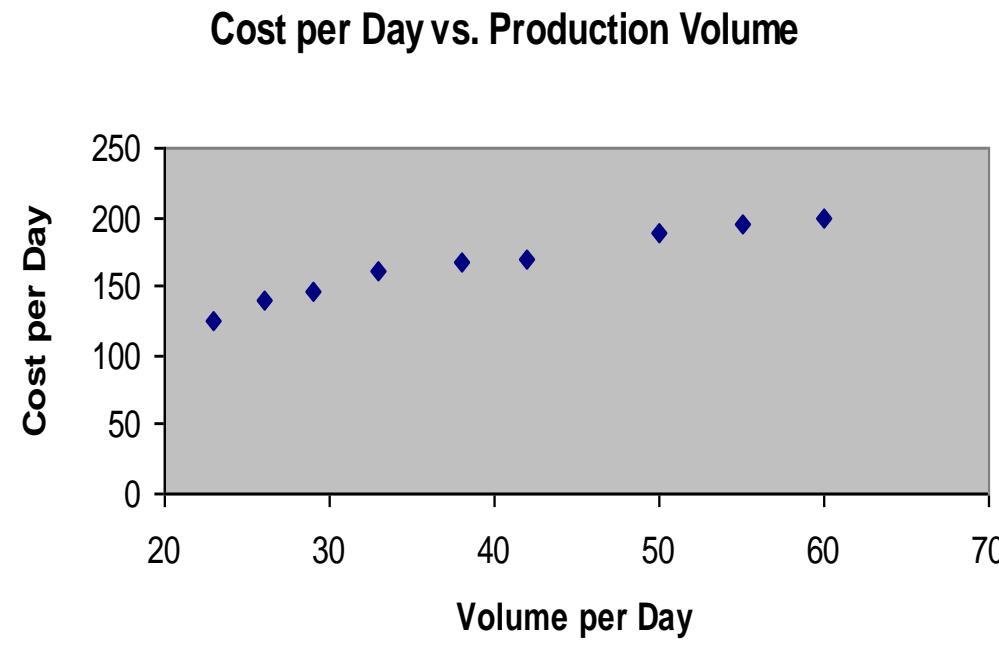
<b>Importance of Brand Name</b>	<b>Male</b>	<b>Female</b>	<b>Total</b>
<b>More</b>	450	300	750
<b>Equal or Less</b>	3300	3450	6750
<b>Total</b>	<b>3750</b>	<b>3750</b>	<b>7500</b>

# Scatter Plots



- Scatter plots are used for numerical data consisting of paired observations taken from two numerical variables
- One variable is measured on the vertical axis and the other variable is measured on the horizontal axis
- Scatter plots are used to examine possible relationships between two numerical variables

Volume per day	Cost per day
23	125
26	140
29	146
33	160
38	167
42	170
50	188
55	195
60	200

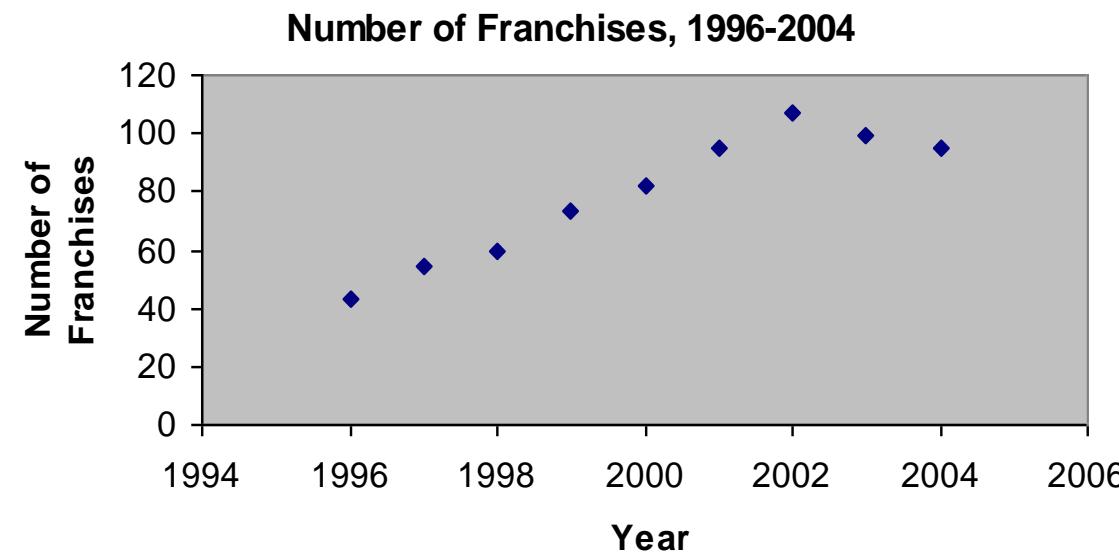


# Time Series Plots



- A Time Series Plot is used to study patterns in the values of a numeric variable over time.
- Numeric variable is measured on the vertical axis and the time period is measured on the horizontal axis

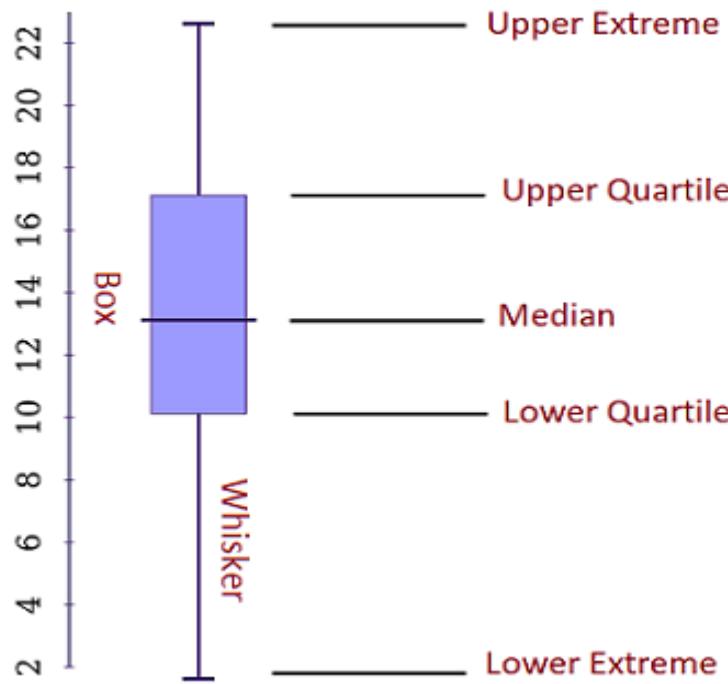
Year	Number of Franchises
1996	43
1997	54
1998	60
1999	73
2000	82
2001	95
2002	107
2003	99
2004	95



# Box and Whisker Plot



- It is a graphical method of displaying variation in a set of data.
- A box and whisker plot (also known as a box plot) is a graph that represents visually data from a five-number summary.



**Upper Extreme** – the highest value in a given dataset.

**Upper Quartile** – above that value, the upper 25% of the data are contained.

**Median value** – the middle number in the set.

**Lower Quartile** – below that value, the lower 25% of the data are contained.

**Lower Extreme** – the smallest value in a given dataset.

**Whiskers** – the lines that extend from the boxes. They are used to indicate variability out of the upper and lower quartiles.

# Example 1



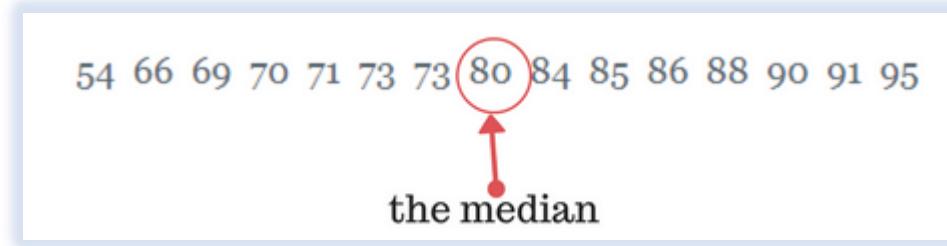
Suppose you have the math test results for a class of 15 students.

- Here are the results: 91 95 54 69 80 85 88 73 71 70 66 90 86 84 73
- It is hard to say what is the middle point (the median) because the value points are not ordered.

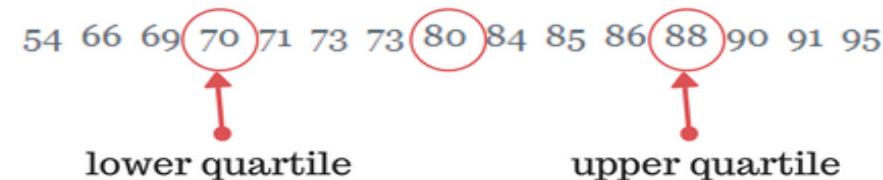
**Step 1:** Order the data points from least to greatest.

54 66 69 70 71 73 73 80 84 85 86 88 90 91 95

**Step 2:** Find the median of the data:



**Step 3:** Find the middle points of the two halves divided by the median (find the upper and lower quartiles).



# Example 1 (Continu..)



**Step 4:** Find the extreme values.

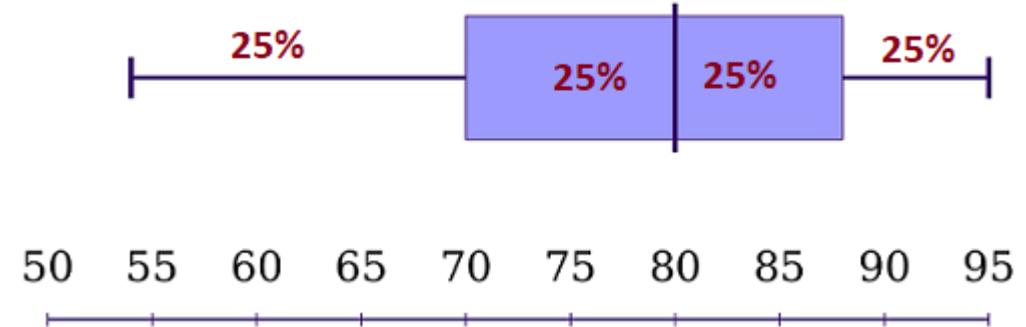
- Extreme values = 54 and 95.

So, we can determine that the five-number summary for the class of students is 54, 70, 80, 88, 95.

As you see, the plot is divided into four groups:

- A lower whisker
- A lower box half
- An upper box half
- An upper whisker

Each of those groups shows 25% of the data because we have an equal amount of data in each group.



## Interpreting the box and whisker plot results:

50% of the students have scores between 70 and 88 points.

75% scored lower than 88 points

50% have test results above 80.

# Example 2



## Comparative double box and whisker plot

Suppose an IT company has two stores that sell computers. The company recorded the number of sales each store made each month. In the past 12 months, we have the following numbers of sold computers.

**Store 1:** 350, 460, 20, 160, 580, 250, 210, 120, 200, 510, 290, 380.

**Store 2:** 520, 180, 260, 380, 80, 500, 630, 420, 210, 70, 440, 140.

### Store 1:

First, we put the data points in ascending order.

20, 120, 160, 200, 210, 250, 290, 350, 380, 460, 510 580.

- **The median** is  $(250 + 290) / 2 = 270$
- There are six numbers below the median: 20, 120, 160, 200, 210, 250.

**Lower quartile** is the median of these six items, so

$$= (\text{third} + \text{fourth data point}) / 2$$

$$= (160 + 200) / 2$$

$$= 180$$

There are also six numbers above the median: 290, 350, 380, 460, 510 580.

# Example 2 (Continu...)



**Upper quartile** is the median of these six data points

$$= (\text{third} + \text{fourth data points}) / 2$$

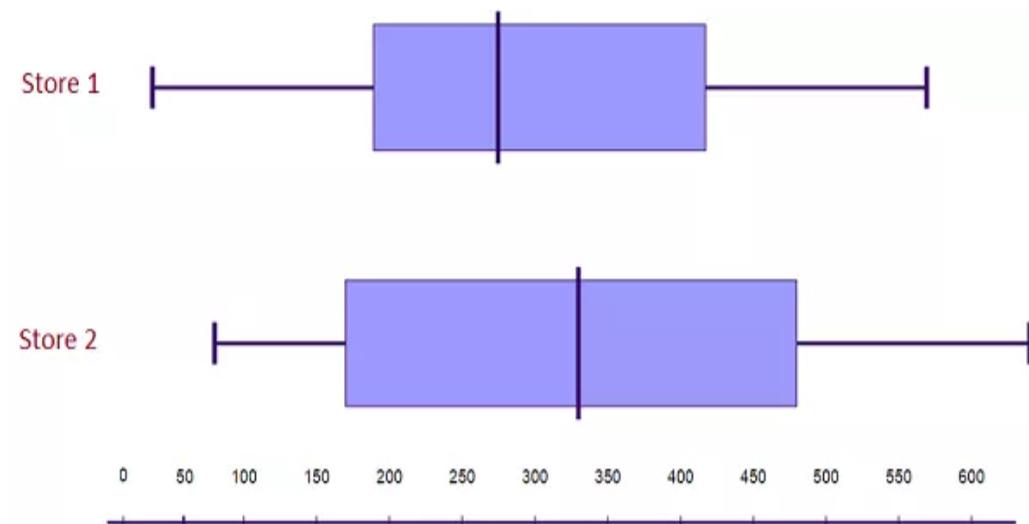
$$= 420$$

Finally, the five-number summary for Store 1's sales is 20, 180, 270, 420, 580.

Using the same calculations, we can find that the five-number summary for Store 2 is 70, 160, 320, 470, 630

## Interpreting the results:

- Store 2's highest and lowest sales are both higher than Store 1's relevant sales.
- Store 2's median sales value is higher than Store 1's.
- Store 2's interquartile range is larger.



**Final:** Store 2 consistently sells more computers than Store 1.

# Principles of Excellent Graphs

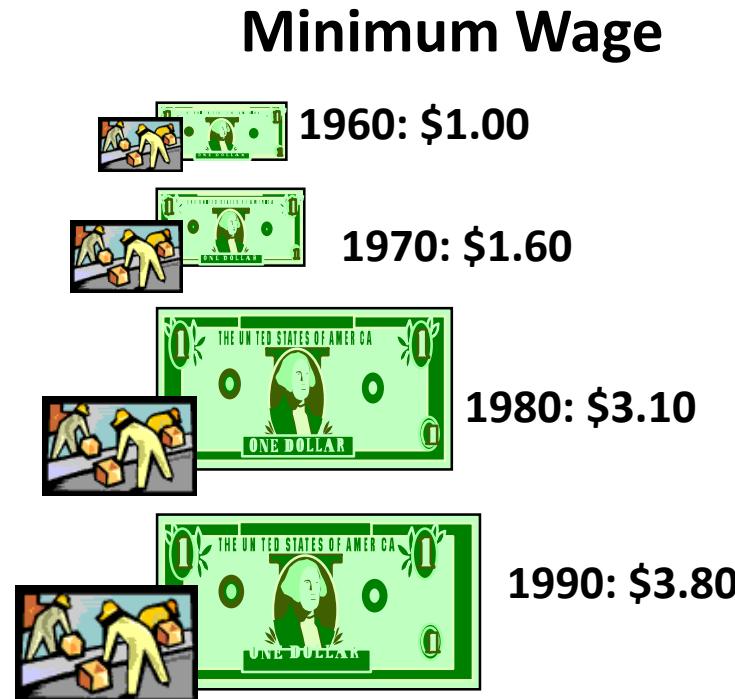


- The graph should not distort the data.
- The graph should not contain unnecessary adornments (sometimes referred to as chart junk).
- The scale on the vertical axis should begin at zero.
- All axes should be properly labeled.
- The graph should contain a title.
- The simplest possible graph should be used for a given set of data.

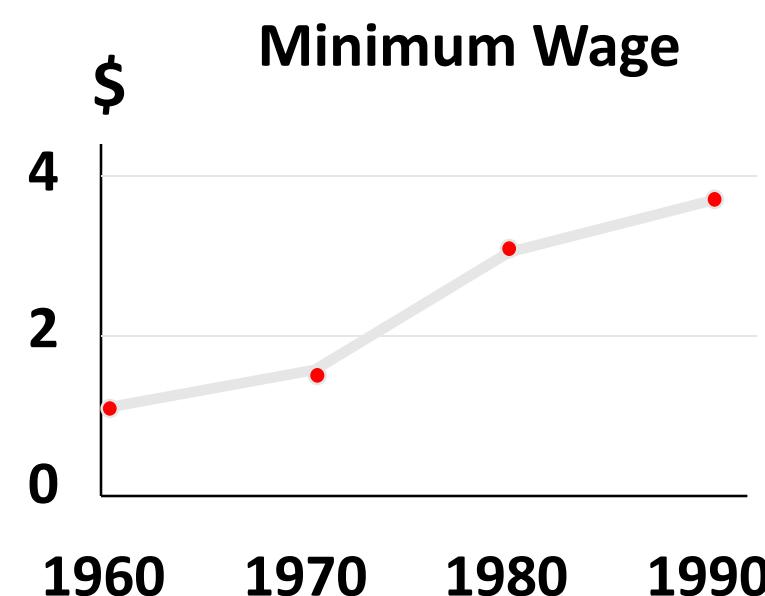
# Graphical Errors: Chart Junk



Bad Presentation



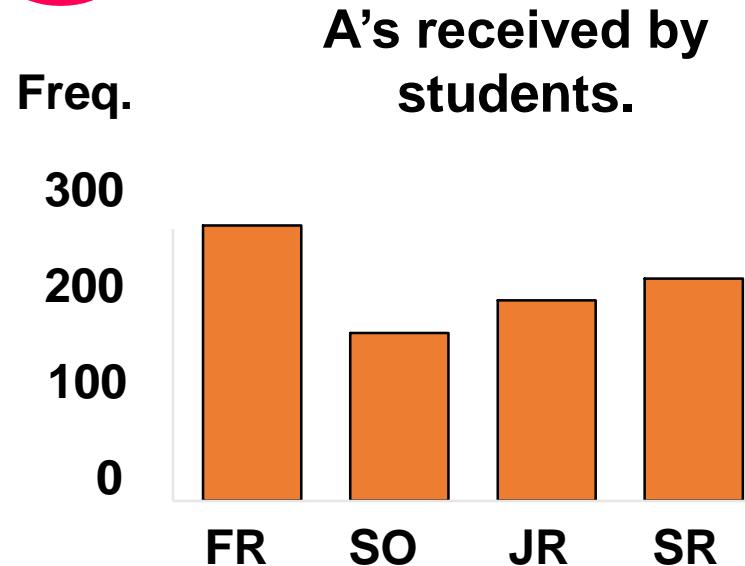
Good Presentation



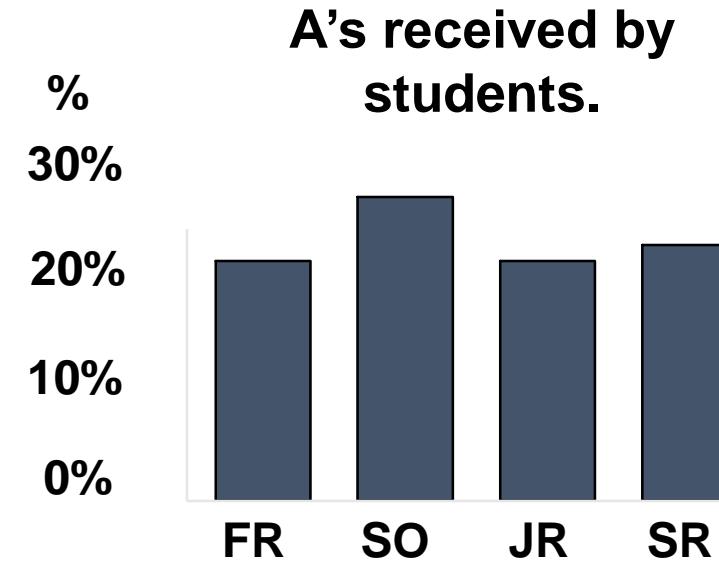
# Graphical Errors: No Relative Basis



## Bad Presentation



## Good Presentation

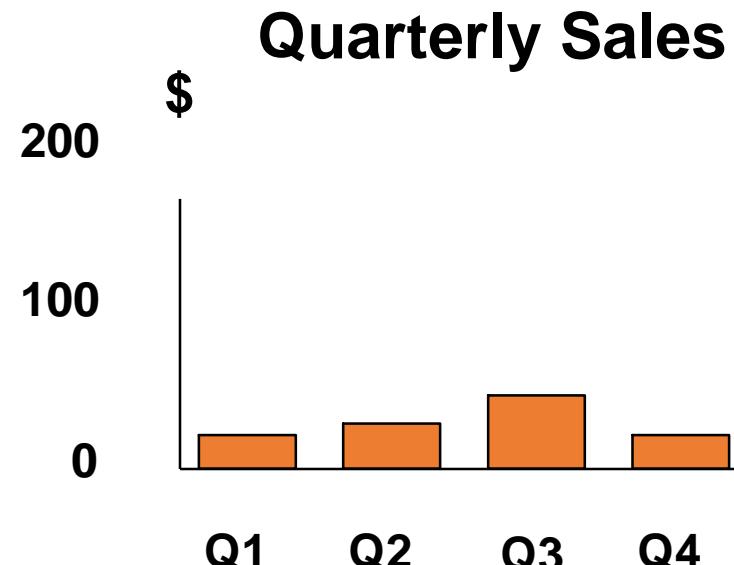


FR = Freshmen, SO = Sophomore, JR = Junior, SR = Senior

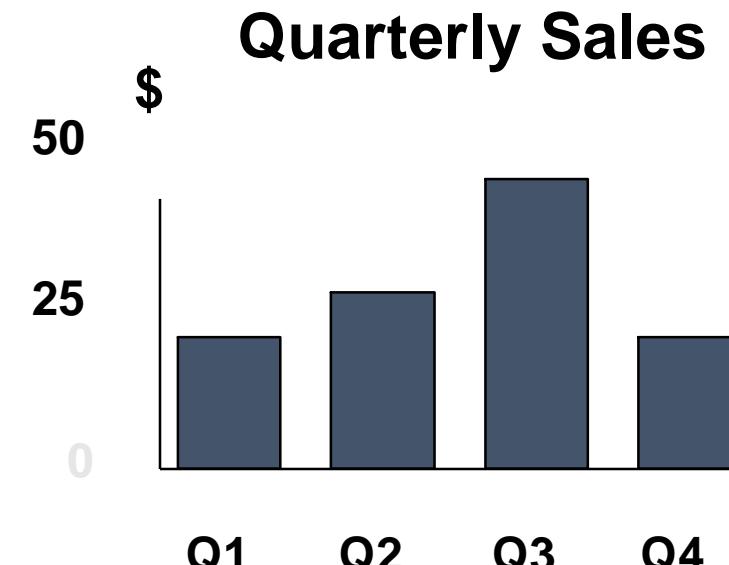
# Graphical Errors: Compressing the Vertical Axis



**Bad Presentation**



**Good Presentation**



# Graphical Errors

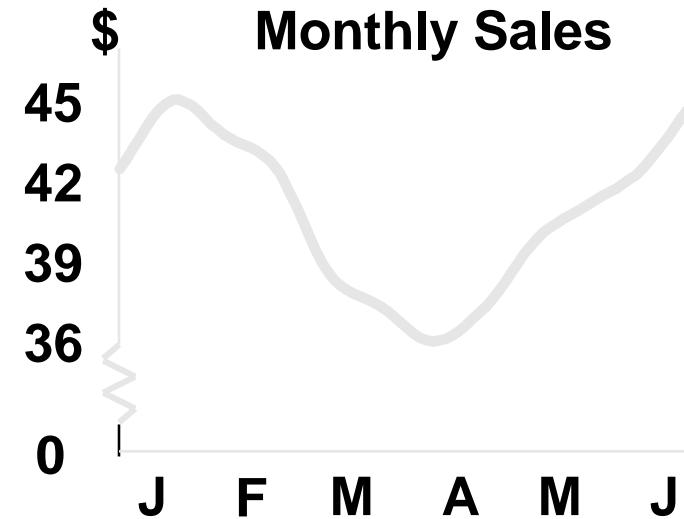
## No Zero Point on the Vertical Axis



**Bad Presentation**



**Good Presentations**



Graphing the first six months of sales

# Central Tendency



- The **central tendency** is the extent to which all the data values group around a typical or central value.
- It aims to provide an accurate description of the entire data.

## Central Tendency

MODE	MEDIAN	MEAN
<ul style="list-style-type: none"><li>• most frequent data point</li><li>• mode exists as a data point</li><li>• unaffected by extreme values</li><li>• useful for qualitative data</li><li>• may have more than 1 value</li></ul>	<ul style="list-style-type: none"><li>• value that divides ranked data points into halves: 50% larger than it, 50% smaller</li><li>• may not exist as a data point in the set</li><li>• influenced by position of items, but not their values</li></ul>	$\bar{x} = \frac{\sum x}{N}$ <ul style="list-style-type: none"><li>• most stable measure</li><li>• affected by extreme values</li><li>• may not exist as a data point in the set</li></ul>

# Mean



- The arithmetic mean (often just called “mean”) is the most common measure of central tendency.
- To be used for data that is normally distributed.
  - For a sample of size  $n$ :

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

Diagram illustrating the components of the arithmetic mean formula:

- Pronounced x-bar: Points to the symbol  $\bar{X}$ .
- The i<sup>th</sup> value: Points to the term  $X_i$  in the summation.
- Sample size: Points to the denominator  $n$ .
- Observed values: Points to the terms  $X_1, X_2, \dots, X_n$  in the summation.

# Mean Example



On his first three quizzes, Patrick earned a 15, 18, and 16. Find the mean.

**Answer:**

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

Here  $n=3$ ;  $x_1=15$ ,  $x_2= 18$  and  $x_3=16$ .

$$\text{Mean} = (15+18+16)/3 = 19$$

Mean is 19 points

**Types:**

- **Geometric mean:**
  - Type of average , usually used for growth rates, like population growth or interest rates.
  - **Geometric mean** multiplies items.

# Median

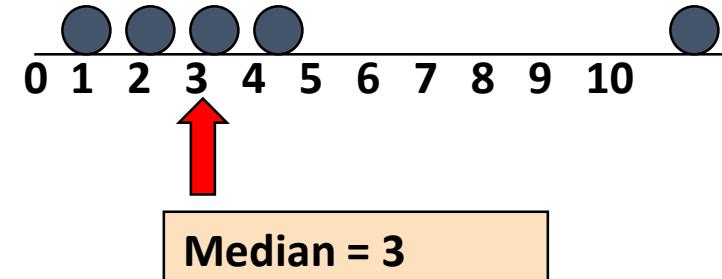
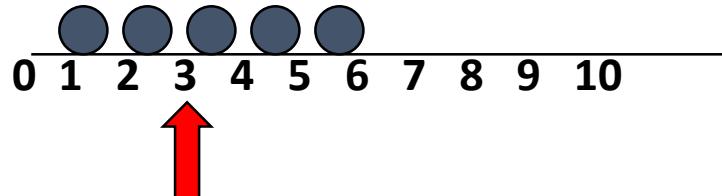


- The location of the median when the values are in numerical order (smallest to largest):

$$\text{Median position} = \frac{n+1}{2} \text{ position in the ordered data}$$

- If the number of values is odd, the median is the middle number
- If the number of values is even, the median is the average of the two middle numbers

Note: That  $\frac{n+1}{2}$  is not the *value* of the median, only the *position* of the median in the ranked data



# Median Example



Example: Birthday Activities (continued)

List the ages in order:

1, 1, 1, 1, 1, 13, 13, 13, 13, 13, 13

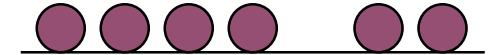
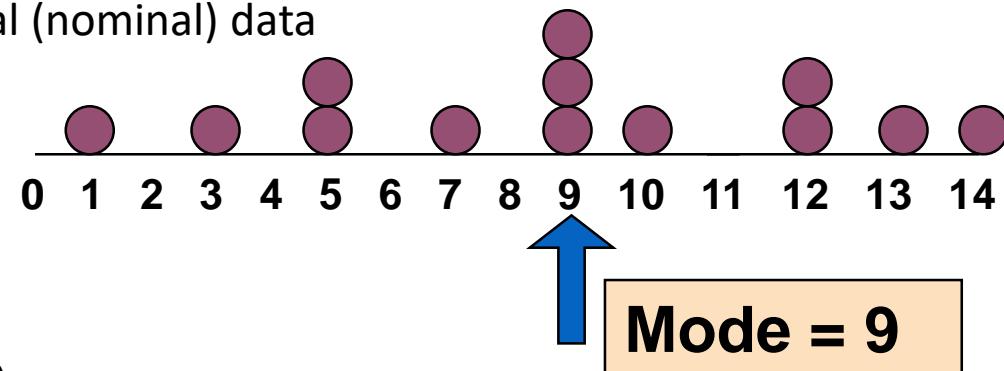
Choose the middle number:

1, 1, 1, 1, 1, **13**, 13, 13, 13, 13, 13

# Mode



- Value that occurs most often
- Not affected by extreme values
- Used for either numerical or categorical (nominal) data
- There may be no mode
- There may be several modes.



No Mode

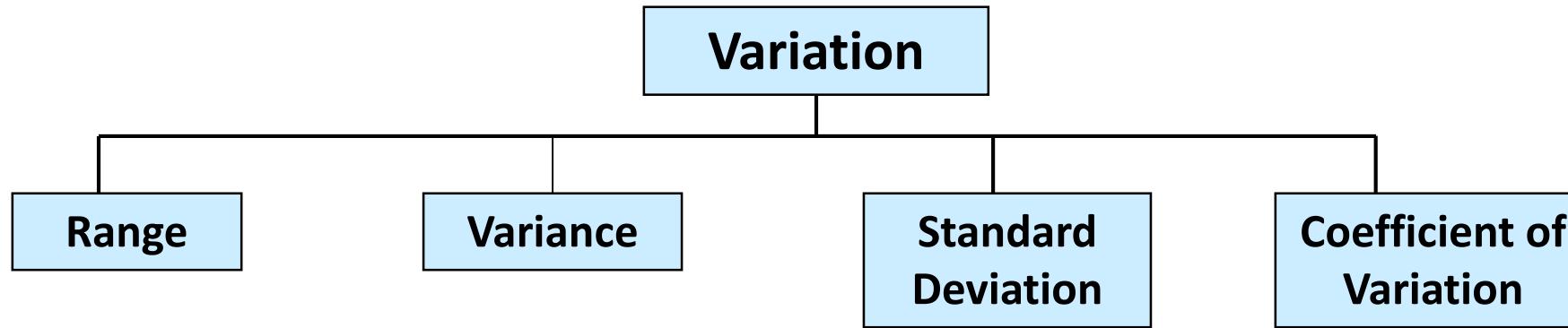
Example: Birthday Activities (continued)

Group the numbers so we can count them:

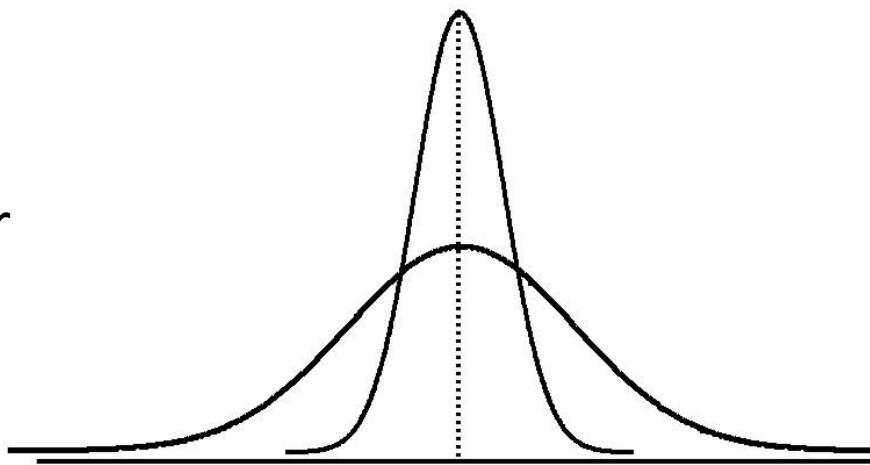
1, 1, 1, 1, 1, 13, 13, 13, 13, 13

"13" occurs 6 times, "1" occurs only 5 times, so the mode is **13**.

# Measures of Variation



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



Same center,  
different variation

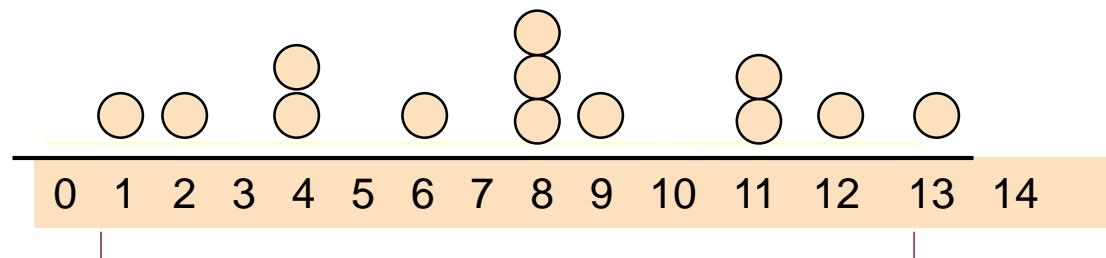
# Range



- Simplest measure of variation
- Difference between the largest and the smallest values:

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

Example:



$$\text{Range} = 13 - 1 = 12$$

# Variance



- Average (approximately) of squared deviations of values from the mean

- Sample variance:

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

Where  $\bar{X}$  = arithmetic mean

$n$  = sample size

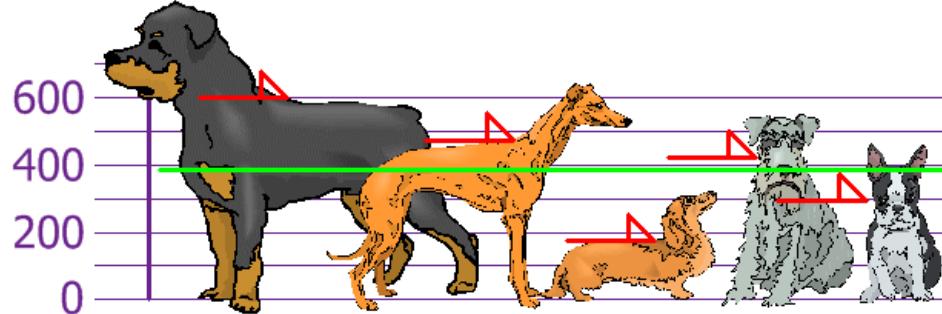
$X_i$  =  $i^{th}$  value of the variable  $X$

# Standard Deviation

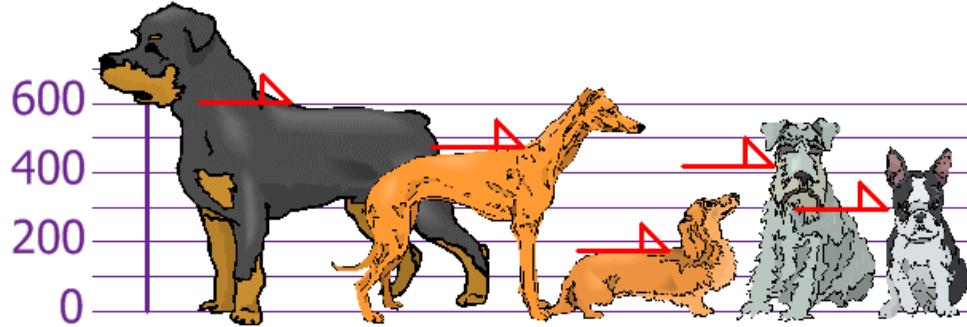


- Most commonly used measure of variation
- Shows variation about the mean
- Is the square root of the variance
- Has the **same units as the original data**
  - Sample standard deviation:

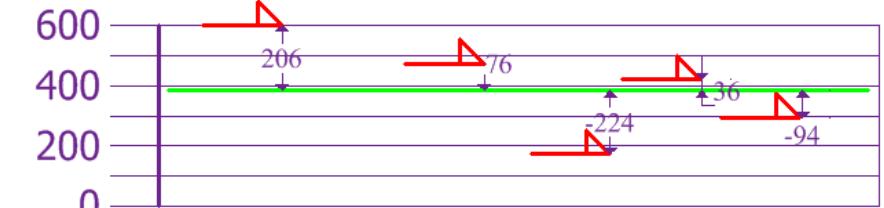
$$S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}}$$



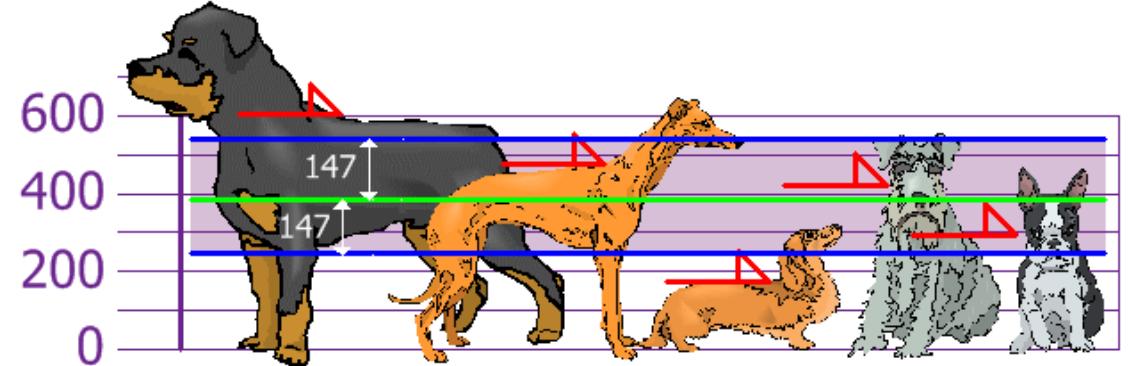
You and your friends have just measured the heights of your dogs (in millimeters):



- The heights (at the shoulders) are: 600mm, 470mm, 170mm, 430mm and 300mm



SD tell us which is normal , biggest and smallest.  
From this example with in blue line is normal and above blue line are biggest



# Coefficient of Variation



- Measure of relative variability.
  - It is the ratio of the standard deviation to the mean (average).  
For example, the expression “The standard deviation is 15% of the mean” is a Coefficient of Variation.(CV)
  - CV is particularly useful when you want to compare results from two different surveys or tests that have different measures or values.
- 
- Example:  
A researcher is comparing two multiple-choice tests with different conditions.  
In the first test, a typical multiple-choice test is administered. In the second test, alternative choices (i.e. incorrect answers) are randomly assigned to test takers. The results from the two tests are

	Regular Test	Randomized Answers
Mean	59.9	44.8
SD	10.2	12.7

# Coefficient of Variation



$$CV = (SD / \text{Mean}) * 10$$

	Regular Test	Randomized Answers
MEAN	59.9	44.8
SD	10.2	12.7
CV	17.03	28.35

- Looking at the standard deviations of 10.2 and 12.7, you might think that the tests have similar results.
- However, when you adjust for the difference in the means, the results have more significance.

Regular test: CV = 17.03

Randomized answers: CV = 28.35

# Process of Data Collection



1. Define the objectives of the survey or experiment.

Example: Estimate the average life of an electronic component.

2. Define the variable and population of interest.

Example: Length of time for anesthesia to wear off after surgery.

3. Defining the data-collection and data-measuring schemes. This includes sampling procedures, sample size, and the data-measuring device (questionnaire, scale, ruler, etc.).

4. Determine the appropriate descriptive or inferential data-analysis techniques.

## Methods used to Collect Data:

**Experiment:** The investigator controls or modifies the environment and observes the effect on the variable under study.

**Survey:** Data are obtained by sampling some of the population of interest. The investigator does not modify the environment

**Census:** A 100% survey. Every element of the population is listed. Seldom used: difficult and time-consuming to compile, and expensive.

# Sampling



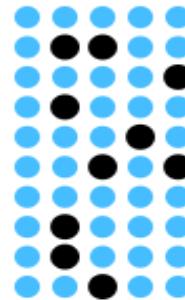
**Population:** The whole group we are interested in.

**Census:** A collection of data from the whole population.

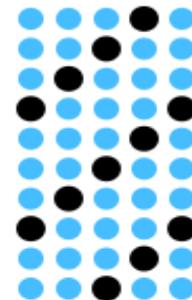
**Sample:** A collection of data from **part** of the population.

**Sampling:** Process used in statistical analysis in which a predetermined number of observations are taken from a larger population.

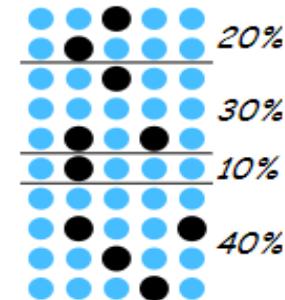
**Types:**



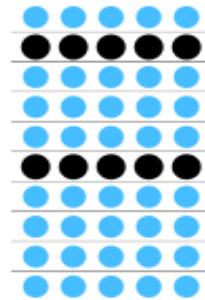
**Random Sample**  
(pick randomly  
from list)



**Systematic  
Sample**  
(such as every 4th)



**Stratified Sample**  
(randomly, but in  
ratio to group size)



**Cluster Sample**  
(choose whole  
groups randomly)

# Sampling Types



## 1. Random Sampling:

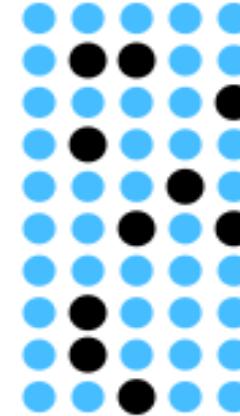
Pick Randomly from list.

**Example:** You want to know the favorite colors for people at your school, but don't have the time to ask everyone.

Somehow get a full list of students printed out, then

- place all pages on the ground, drop a pencil and note down the student's name.
- repeat until you have 50 names.

Now survey those 50.

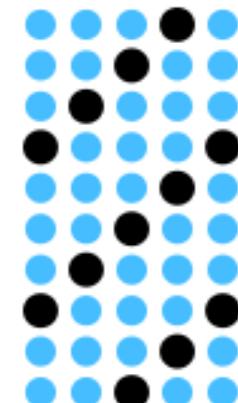


## 2. Systematic Sampling:

Sample is selected according to a random starting point and a fixed, periodic interval.

**Example:** You want to know the favorite colors for people at your school, but don't have the time to ask everyone.

**Solution:** stand at the gate and choose "every 4th person to arrive."



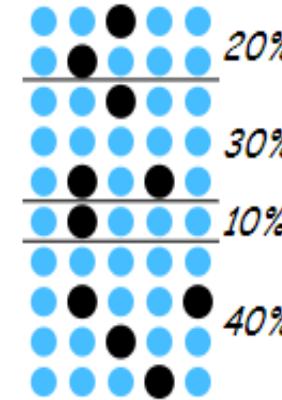
# Sampling Types (Continu...)



## 3. Stratified Sampling:

This is where we divide the population into groups by some characteristic such as age or occupation or gender.

Then make sure our survey includes people from each group in proportion to how many there are in the whole population.



**Example:** We want to survey 300 people in the USA

This is the population breakdown for the USA in 2010:

We want to survey 300 people, so we choose:

Age Range	Percent	Age Range	Percent	People
0-4	6.5%	0-4	6.5%	20
5-17	17.5%	5-17	17.5%	52
18-23	9.9%	18-23	9.9%	30
24-44	26.6%	24-44	26.7%	80
45-64	26.4%	45-64	26.4%	79
65+	13.0%	65+	13.0%	39
<b>100%</b>		<b>100%</b>		<b>300</b>

# Sampling Types



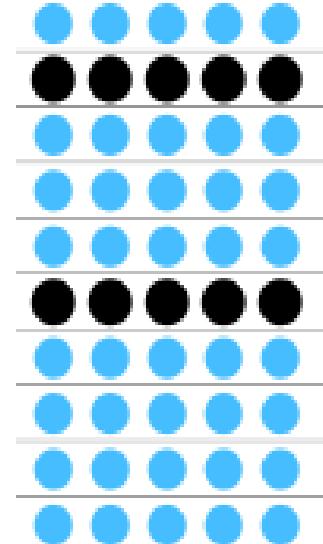
## 4. Cluster Sampling:

We break the population into many groups, then randomly choose whole groups.

**Example:** we divide the town into many different zones, then randomly choose 5 zones and survey everyone in those zones.

Cluster sampling works best when the clusters are similar in character to each other.

**For Instance:** If the town has rich and poor zones then try to create a new way of dividing the town into fairer regions.





# Probability Theory

# Uncertainties



Managers often base their decisions on an analysis of uncertainties such as the following:

What are the *chances* that sales will decrease if we increase prices?

What is the *likelihood* a new assembly method will increase productivity?

What are the *odds* that a new investment will be profitable?

# Probability



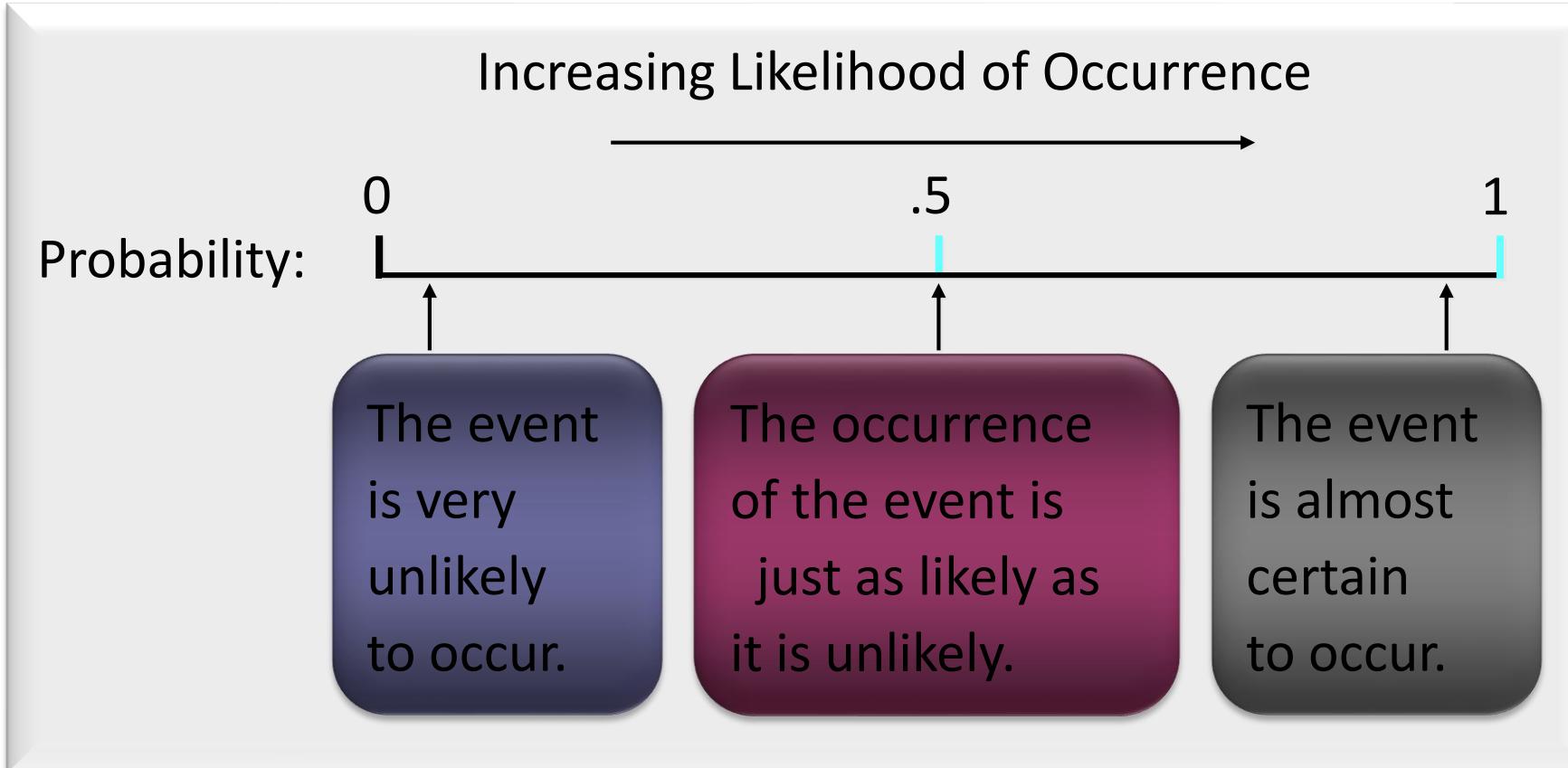
Probability is a numerical measure of the likelihood that an event will occur.

Probability values are always assigned on a scale from 0 to 1.

A probability near zero indicates an event is quite unlikely to occur.

A probability near one indicates an event is almost certain to occur.

# Probability as a Numerical Measure of the Likelihood of Occurrence



# Statistical Experiment



In statistics, the notion of an experiment differs somewhat from that of an experiment in the physical sciences.

In statistical experiments, probability determines outcomes.

Even though the experiment is repeated in exactly the same way, an entirely different outcome may occur.

For this reason, statistical experiments are sometimes called *random experiments*.

# Terminologies



An experiment is any process that generates well-defined outcomes.

The sample space for an experiment is the set of all experimental outcomes.

An experimental outcome is also called a sample point.

# An Experiment and its sample space



## Experiment

Toss a coin

Inspection a part

Conduct a sales call

Roll a die

Play a football game

## Experiment Outcomes

Head, tail

Defective, non-defective

Purchase, no purchase

1, 2, 3, 4, 5, 6

Win, lose, tie

# An Experiment and its sample space



## Example: Bradley Investments

Bradley has invested in two stocks, Markley Oil and Collins Mining. Bradley has determined that the possible outcomes of these investments three months from now are as follows.

Investment Gain or Loss in 3 Months (in \$000)	
<u>Markley Oil</u>	<u>Collins Mining</u>
10	8
5	-2
0	
-20	

# A Counting Rule for Multiple-Step Experiment



If an experiment consists of a sequence of  $k$  steps in which there are  $n_1$  possible results for the first step,  $n_2$  possible results for the second step, and so on, then the total number of experimental outcomes is given by  $(n_1)(n_2) \dots (n_k)$ .

A helpful graphical representation of a multiple-step experiment is a tree diagram.

# A Counting Rule for Multiple-Step Experiment



## Example: Bradley Investments

Bradley Investments can be viewed as a two-step experiment. It involves two stocks, each with a set of experimental outcomes.

Markley Oil:  $n_1 = 4$

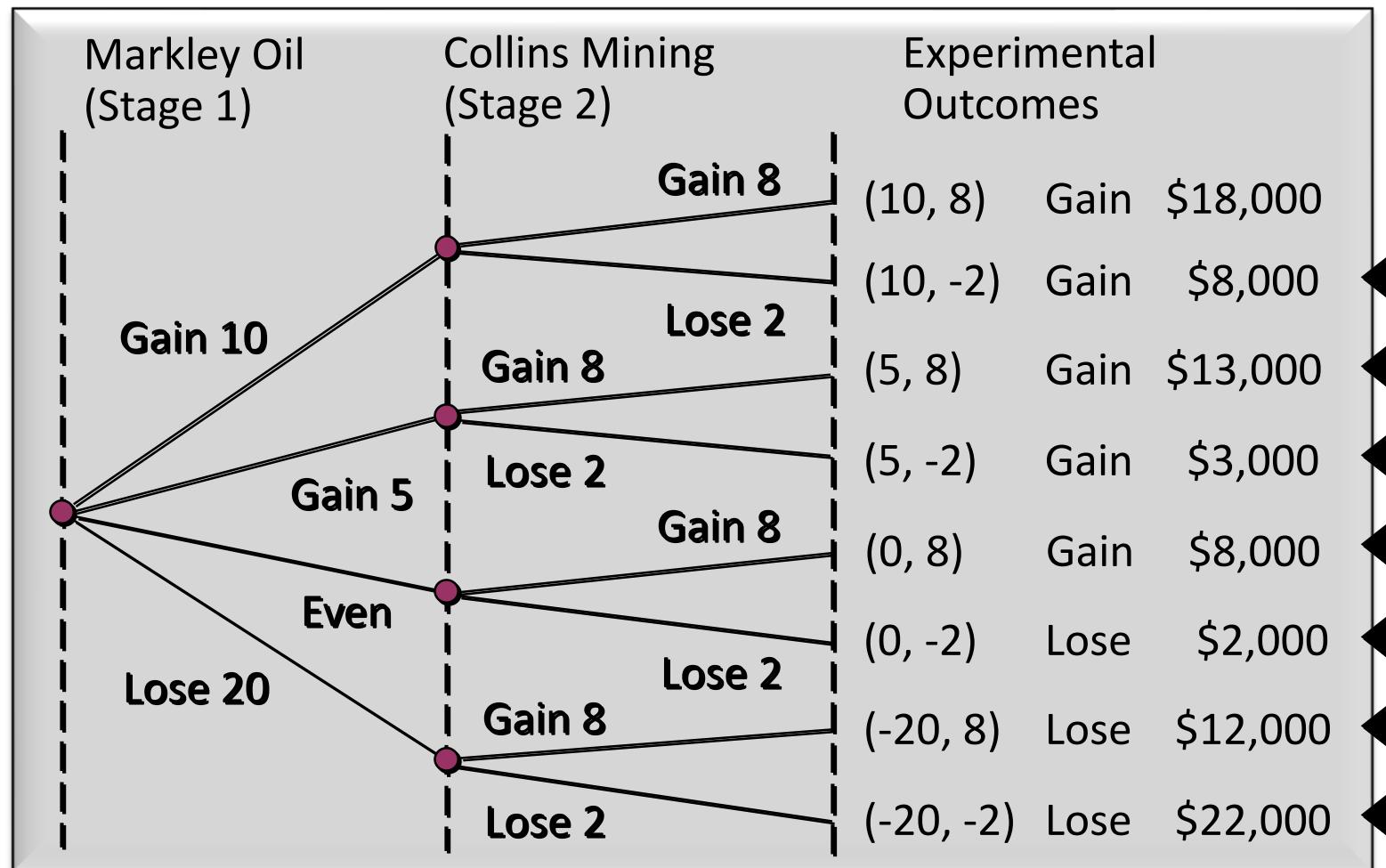
Collins Mining:  $n_2 = 2$

Total Number of  
Experimental Outcomes:  $n_1 n_2 = (4)(2) = 8$

# Tree Diagram



Example: Bradley Investments



# Counting Rule for Combinations



Number of Combinations of  $N$  Objects  
Taken  $n$  at a Time

A second useful counting rule enables us to count the number of experimental outcomes when  $n$  objects are to be selected from a set of  $N$  objects.

$$C_n^N = \binom{N}{n} = \frac{N!}{n!(N-n)!}$$

where:  $N! = N(N - 1)(N - 2) \dots (2)(1)$   
 $n! = n(n - 1)(n - 2) \dots (2)(1)$   
 $0! = 1$

For a fruit salad, how many different combinations of 2 ingredients can you make with apple, banana and cherry?

Answer: {apple, banana}, {apple, cherry} or {banana, cherry}

# Counting Rule for Permutations



Number of Permutations of  $N$  Objects  
Taken  $n$  at a Time

A third useful counting rule enables us to count the number of experimental outcomes when  $n$  objects are to be selected from a set of  $N$  objects, where the order of selection is important.

$$P_n^N = n! \binom{N}{n} = \frac{N!}{(N-n)!}$$

where:  $N! = N(N - 1)(N - 2) \dots (2)(1)$   
 $n! = n(n - 1)(n - 2) \dots (2)(1)$   
 $0! = 1$

You want to visit the homes of three friends Alex ("a"), Betty ("b") and Chandra ("c"), but haven't decided in what order. What choices do you have?

Answer: {a,b,c} {a,c,b} {b,a,c} {b,c,a} {c,a,b} {c,b,a}

# Assigning Probabilities



## Basic Requirements for Assigning Probabilities

1. The probability assigned to each experimental outcome must be between 0 and 1, inclusively.

$$0 \leq P(E_i) \leq 1 \text{ for all } i$$

where:

$E_i$  is the  $i$ th experimental outcome  
and  $P(E_i)$  is its probability

# Assigning Probabilities



## Basic Requirements for Assigning Probabilities

2. The sum of the probabilities for all experimental outcomes must equal 1.

$$P(E_1) + P(E_2) + \dots + P(E_n) = 1$$

where:

$n$  is the number of experimental outcomes

# Assigning Probabilities



## Classical Method

Assigning probabilities based on the assumption  
of equally likely outcomes

## Relative Frequency Method

Assigning probabilities based on experimentation  
or historical data

## Subjective Method

Assigning probabilities based on judgment

# Classical Method



Example: Rolling a Die

If an experiment has  $n$  possible outcomes, the classical method would assign a probability of  $1/n$  to each outcome.

Experiment: Rolling a die

Sample Space:  $S = \{1, 2, 3, 4, 5, 6\}$

Probabilities: Each sample point has a  
1/6 chance of occurring

# Relative Frequency Method



## Example: Lucas Tool Rental

Lucas Tool Rental would like to assign probabilities to the number of car polishers it rents each day.

Office records show the following frequencies of daily rentals for the last 40 days.

<u>Number of Polishers Rented</u>	<u>Number of Days</u>
0	4
1	6
2	18
3	10
4	2

# Relative Frequency Method



Example: Lucas Tool Rental

Each probability assignment is given by dividing the frequency (number of days) by the total frequency (total number of days).

<u>Number of Polishers Rented</u>	<u>Number of Days</u>	<u>Probability</u>
0	4	.10
1	6	.15
2	18	.45
3	10	.25
4	$\frac{2}{40}$	$\frac{.05}{1.00}$

A dark grey callout bubble with a white border and a black arrow points from the bottom right towards the probability value in the last row of the table. Inside the bubble, the fraction  $4/40$  is displayed.

# Subjective Method



When economic conditions and a company's circumstances change rapidly it might be inappropriate to assign probabilities based solely on historical data.

We can use any data available as well as our experience and intuition, but ultimately a probability value should express our degree of belief that the experimental outcome will occur.

The best probability estimates often are obtained by combining the estimates from the classical or relative frequency approach with the subjective estimate.

# Subjective Method



Example: Bradley Investments

An analyst made the following probability estimates.

<u>Exper. Outcome</u>	<u>Net Gain or Loss</u>	<u>Probability</u>
(10, 8)	\$18,000 Gain	.20
(10, -2)	\$8,000 Gain	.08
(5, 8)	\$13,000 Gain	.16
(5, -2)	\$3,000 Gain	.26
(0, 8)	\$8,000 Gain	.10
(0, -2)	\$2,000 Loss	.12
(-20, 8)	\$12,000 Loss	.02
(-20, -2)	\$22,000 Loss	.06

# Events and Their Probabilities



An event is a collection of sample points.

The probability of any event is equal to the sum of the probabilities of the sample points in the event.

If we can identify all the sample points of an experiment and assign a probability to each, we can compute the probability of an event.

# Events and Their Probabilities



Example: Bradley Investments

Event  $M$  = Markley Oil Profitable

$$M = \{(10, 8), (10, -2), (5, 8), (5, -2)\}$$

$$\begin{aligned}P(M) &= P(10, 8) + P(10, -2) + P(5, 8) + P(5, -2) \\&= .20 + .08 + .16 + .26 \\&= .70\end{aligned}$$

# Events and Their Probabilities



Example: Bradley Investments

Event C = Collins Mining Profitable

$$C = \{(10, 8), (5, 8), (0, 8), (-20, 8)\}$$

$$\begin{aligned}P(C) &= P(10, 8) + P(5, 8) + P(0, 8) + P(-20, 8) \\&= .20 + .16 + .10 + .02 \\&= .48\end{aligned}$$

# Some Basics Relationships Of Probability



There are some basic probability relationships that can be used to compute the probability of an event without knowledge of all the sample point probabilities.

Complement of an Event

Union of Two Events

Intersection of Two Events

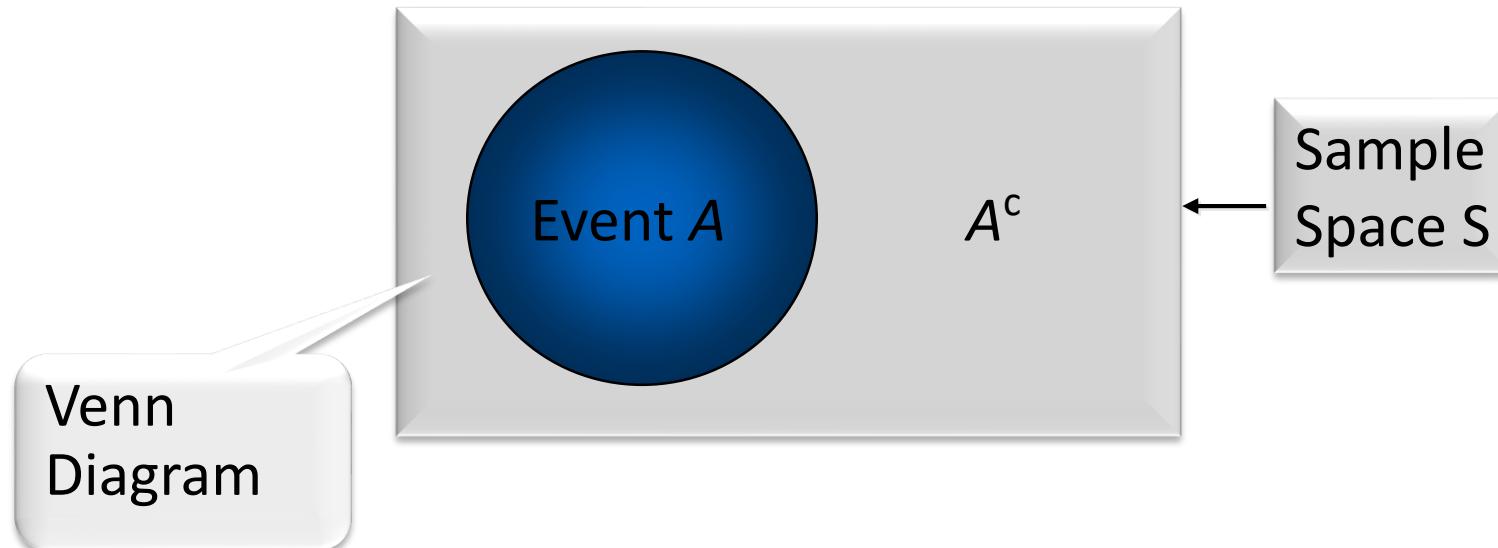
Mutually Exclusive Events

# Complement of an Event



The complement of event  $A$  is defined to be the event consisting of all sample points that are not in  $A$ .

The complement of  $A$  is denoted by  $A^c$ .



Consider the experiment of rolling a single die.

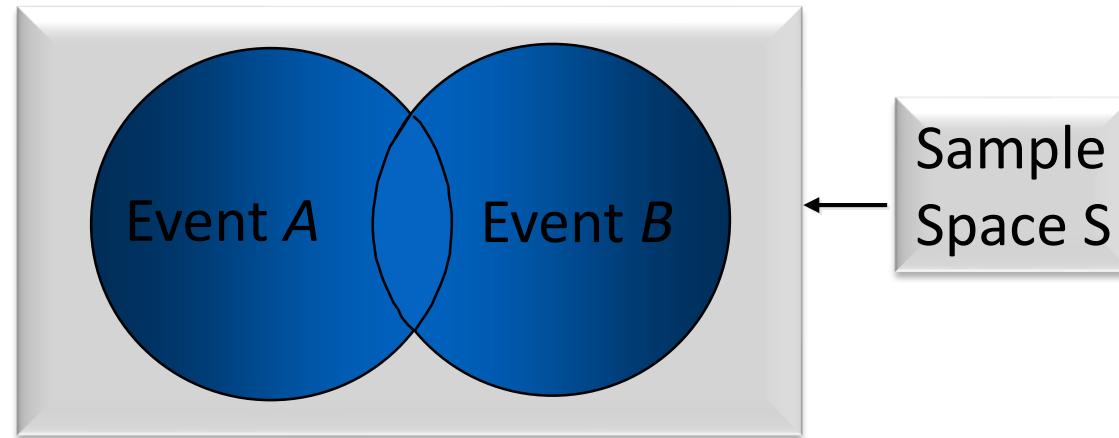
$S=\{1,2,3,4,5,6\}$ . Let  $A$  be the event that the roll yields a number greater than 4 .  $A=\{5,6\}$  . Then  $A^c=\{1,2,3,4\}$

# Union of Two Events



The union of events  $A$  and  $B$  is the event containing all sample points that are in  $A$  or  $B$  or both.

The union of events  $A$  and  $B$  is denoted by  $A \cup B$



# Union of Two Events



Example: Bradley Investments

Event  $M$  = Markley Oil Profitable

Event  $C$  = Collins Mining Profitable

$M \cup C$  = Markley Oil Profitable

or Collins Mining Profitable (or both)

$$M \cup C = \{(10, 8), (10, -2), (5, 8), (5, -2), (0, 8), (-20, 8)\}$$

$$P(M \cup C) = P(10, 8) + P(10, -2) + P(5, 8) + P(5, -2)$$

$$+ P(0, 8) + P(-20, 8)$$

$$= .20 + .08 + .16 + .26 + .10 + .02$$

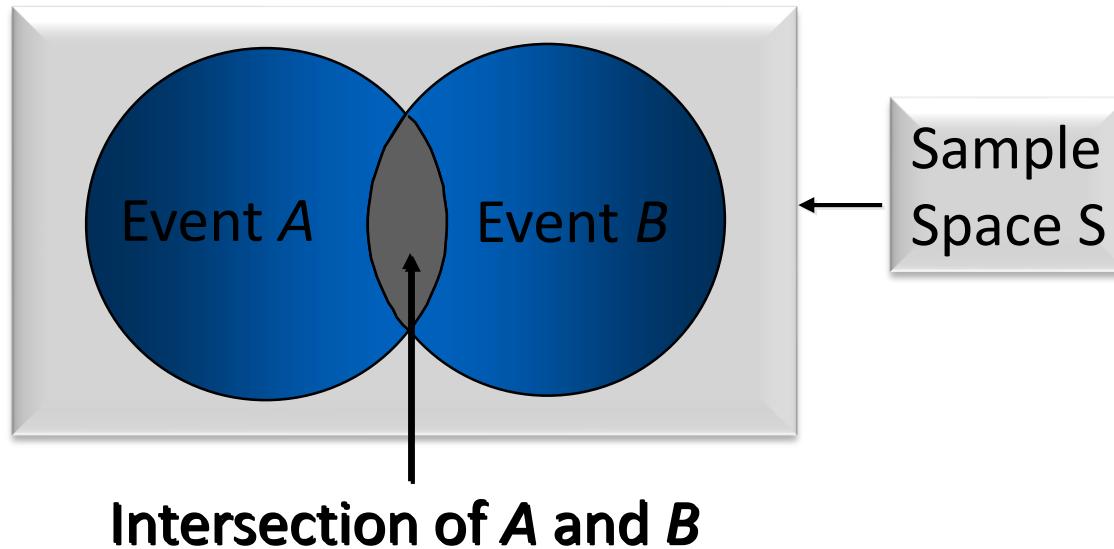
$$= .82$$

# Intersection of Two Events



The intersection of events  $A$  and  $B$  is the set of all sample points that are in both  $A$  and  $B$ .

The intersection of events  $A$  and  $B$  is denoted by  $A \cap B$



# Intersection of Two Events



Example: Bradley Investments

Event  $M$  = Markley Oil Profitable

Event  $C$  = Collins Mining Profitable

$M \cap C$  = Markley Oil Profitable

and Collins Mining Profitable

$$M \cap C = \{(10, 8), (5, 8)\}$$

$$P(M \cap C) = P(10, 8) + P(5, 8)$$

$$= .20 + .16$$

$$= \textcircled{36}$$

# Addition Law



The addition law provides a way to compute the probability of event  $A$ , or  $B$ , or both  $A$  and  $B$  occurring.

The law is written as:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

# Addition Law



Example: Bradley Investments

Event  $M$  = Markley Oil Profitable

Event  $C$  = Collins Mining Profitable

$M \cup C$  = Markley Oil Profitable

or Collins Mining Profitable

We know:  $P(M) = .70$ ,  $P(C) = .48$ ,  $P(M \cap C) = .36$

Thus:  $P(M \cup C) = P(M) + P(C) - P(M \cap C)$

$$= .70 + .48 - .36$$

$$= .82$$

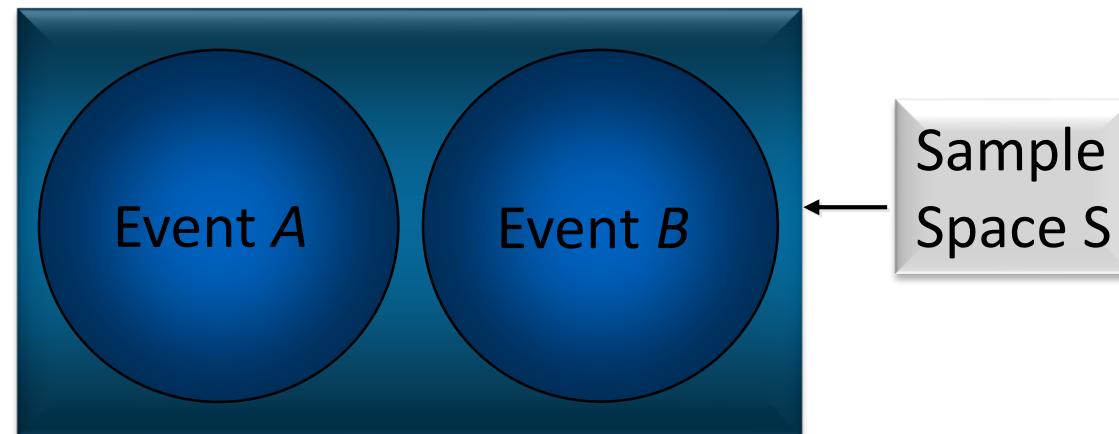
(This result is the same as that obtained earlier using the definition of the probability of an event.)

# Mutually Exclusive Events



Two events are said to be mutually exclusive if the events have no sample points in common.

Two events are mutually exclusive if, when one event occurs, the other cannot occur.



# Mutually Exclusive Events



If events  $A$  and  $B$  are mutually exclusive,  $P(A \cap B) = 0$ .

The addition law for mutually exclusive events is:

$$P(A \cup B) = P(A) + P(B)$$

There is no need to include “ $- P(A \cap B)$ ”

# Conditional Probability



The probability of an event given that another event has occurred is called a conditional probability.

The conditional probability of  $A$  given  $B$  is denoted by  $P(A | B)$ .

A conditional probability is computed as follows :

$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$

# Conditional Probability



Example: Bradley Investments

Event  $M$  = Markley Oil Profitable

Event  $C$  = Collins Mining Profitable

$P(C | M)$  = Collins Mining Profitable  
given Markley Oil Profitable

We know:  $P(M \cap C) = .36$ ,  $P(M) = .70$

Thus: 
$$P(C | M) = \frac{P(C \cap M)}{P(M)} = \frac{.36}{.70} = .5143$$

# Multiplication Law



The multiplication law provides a way to compute the probability of the intersection of two events.

The law is written as:

$$P(A \cap B) = P(B)P(A | B)$$

# Multiplication Law



Example: Bradley Investments

Event  $M$  = Markley Oil Profitable

Event  $C$  = Collins Mining Profitable

$M \cap C$  = Markley Oil Profitable

and Collins Mining Profitable

We know:  $P(M) = .70$ ,  $P(C|M) = .5143$

Thus:  $P(M \cap C) = P(M)P(C|M)$

$$= (.70)(.5143)$$

$$= .36$$

(This result is the same as that obtained earlier using the definition of the probability of an event.)

# Joint Probability Table



		Collins Mining		Total
		Profitable (C)	Not Profitable ( $C^c$ )	
<u>Markley Oil</u>	Profitable (M)	.36	.34	.70
	Not Profitable ( $M^c$ )	.12	.18	.30
Total		.48	.52	1.00

Joint Probabilities  
(appear in the body  
of the table)

Marginal Probabilities  
(appear in the margins  
of the table)

# Independent Events



If the probability of event  $A$  is not changed by the existence of event  $B$ , we would say that events  $A$  and  $B$  are independent.

Two events  $A$  and  $B$  are independent if:

$$P(A | B) = P(A)$$

or

$$P(B | A) = P(B)$$

# Multiplication Law for Independent Events



The multiplication law also can be used as a test to see if two events are independent.

The law is written as:

$$P(A \cap B) = P(A)P(B)$$

# Multiplication Law for Independent Events



Example: Bradley Investments

Event  $M$  = Markley Oil Profitable

Event  $C$  = Collins Mining Profitable

Are events  $M$  and  $C$  independent?

Does  $P(M \cap C) = P(M)P(C)$  ?

We know:  $P(M \cap C) = .36$ ,  $P(M) = .70$ ,  $P(C) = .48$

But:  $P(M)P(C) = (.70)(.48) = .34$ , not .36

Hence:  $M$  and  $C$  are not independent.



# Bayes Theorem

# Bayesian probability



- “**Probability**”: often used to refer to frequency
- ... but
- **Bayesian Probability**: a measure of a state of knowledge.
  
- It quantifies uncertainty. Allows us to reason using uncertain statements.
  
- A Bayesian model is continually updated as more data is acquired.

# How did this come about?



Billiard Table:

- A white billiard ball is rolled along a line and we look at where it stops.
- We suppose that it has a uniform probability of falling anywhere on the line. It stops at a point  $p$ .
- A red billiard ball is then rolled  $n$  times under the same uniform assumption.
- How many times does the red ball roll further than the white ball?

# Bayes' Theorem

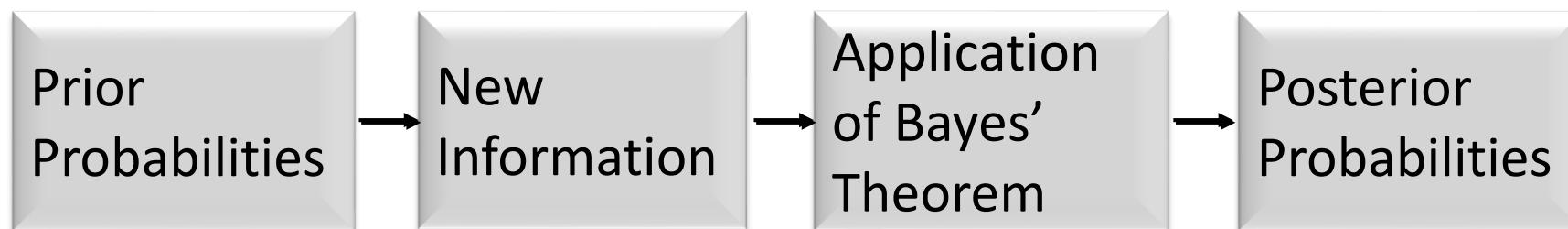


Often we begin probability analysis with initial or prior probabilities.

Then, from a sample, special report, or a product test we obtain some additional information.

Given this information, we calculate revised or posterior probabilities.

Bayes' theorem provides the means for revising the posterior probabilities.



# Bayes Theorem



- Bayes' Theorem shows the relationship between a conditional probability and its inverse.
  - i.e. it allows us to make an inference from
    - the probability of a hypothesis given the evidence to the probability of that evidence given the hypothesis and vice versa

# Bayes Theorem



$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

- $P(A)$  – the PRIOR PROBABILITY – represents your knowledge about A before you have gathered data.
- e.g. probability of a person drawn at random would have schizophrenia (disease) is 0.01

# Bayes Theorem



$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

- $P(B|A)$  – the CONDITIONAL PROBABILITY – the probability of B, given A.
- e.g. you are trying to roll a total of 8 on two dice. What is the probability that you achieve this, given that the first die rolled a 6?

# Bayes Theorem



$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

- So the theorem says:
- The probability of A given B is equal to the probability of B given A, times the prior probability of A, divided by the prior probability of B.

# Understand Bayes Theorem



Understand Bayes Theorem (prior/likelihood/posterior/evidence)

Formula

$$P(X|Y) = ( P(Y|X) * P(X) ) / P(Y)$$

Posterior = (Likelihood\*Prior)/Evidence

# Example



Suppose we have 100 movies and 50 books.

There are 3 different movie types: Action, Sci-fi, Romance,  
2 different book types: Sci-fi, Romance

- 20 of those 100 movies are Action.
- 30 are Sci-fi 50 are Romance.
- 15 of those 50 books are Sci-fi 35 are Romance

# Example



If we want to know that given an object which has type Sci-fi, what the probability is if it's a movie?

Using Bayes theorem, we know that the formula is:

$$P(\text{movie} | \text{Sci-fi}) = P(\text{Sci-fi} | \text{Movie}) * P(\text{Movie}) / P(\text{Sci-fi})$$

Here,  $P(\text{ movie } | \text{ Sci-fi })$  is called **Posterior**,

$P(\text{Sci-fi} | \text{ Movie })$  is **Likelihood**,

$P(\text{movie})$  is **Prior**,

$P(\text{Sci-fi})$  is **Evidence**.

Now let's see why they are

# Example



**Prior:** Before we **observe** it's a Sci-fi type, the object is completely unknown to us.

Our goal is to find out the possibility that it's a movie, we actually have the data **prior(or before)** our **observation**, which is the possibility that it's a movie if it's a completely unknown object:  $P(\text{movie})$ .

**Posterior:** After we **observed** it's a Sci-fi type, we know something about the object. Because it's **post(or after)** the **observation**, we call it **posterior**:  $P(\text{movie} \mid \text{Sci-fi})$ .

**Evidence:** Because we've already known it's a Sci-fi type, what has happened is happened. We **witness** its appearance, so to us, it's an **evidence**, and the chance we get this evidence is  $P(\text{Sci-fi})$ .

**Likelihood:** The dictionary meaning of this word is chance or probability that one thing will happen. Here it means when it's a movie, what the chance will be if it is also a Sci-fi type. This term is very important in Machine Learning.

# A Simple Example



<u>Mode of transport:</u>	<u>Probability he is late:</u>
Car	50%
Bus	20%
Train	1%

Suppose that Bob is late one day.

His boss wishes to estimate the probability that he traveled to work that day by car.

He does not know which mode of transportation Bob usually uses, so he gives a prior probability of 1 in 3 to each of the three possibilities.

# A Simple Example



$$P(A|B) = P(B|A) P(A) / P(B)$$

$$P(\text{car}|\text{late}) = P(\text{late}|\text{car}) \times P(\text{car}) / P(\text{late})$$

$P(\text{late}|\text{car}) = 0.5$  (he will be late half the time he drives)

$P(\text{car}) = 0.33$  (this is the boss' assumption)

$$P(\text{late}) = 0.5 \times 0.33 + 0.2 \times 0.33 + 0.01 \times 0.33$$

(all the probabilities that he will be late added together)

$$P(\text{car}|\text{late}) = 0.5 \times 0.33 / 0.5 \times 0.33 + 0.2 \times 0.33 + 0.01 \times 0.33$$

$$= 0.165 / 0.71 \times 0.33$$

$$= 0.7042$$

# Example



Disease present in 0.5% population (i.e. 0.005)

Blood test is 99% accurate (i.e. 0.99)

False positive 5% (i.e. 0.05)

If someone tests positive, what is the probability that they have the disease?

# Example



Disease present in 0.5% population (i.e. 0.005)

Blood test is 99% accurate (i.e. 0.99)

False positive 5% (i.e. 0.05)

If someone tests positive, what is the probability that they have the disease?

$$P(A|B) = P(B|A) P(A) / P(B)$$

$$P(\text{disease}| \text{pos}) = P(\text{pos} | \text{disease}) \times P(\text{disease}) / P(\text{pos})$$

$$= 0.99 \times 0.005 / (0.99 \times 0.005) + (0.05 \times 0.995)$$

$$= 0.00495 / 0.00495 + 0.04975$$

$$= 0.00495 / 0.0547$$

$$= 0.0905$$

# What does this mean?



- If someone tests positive for the disease, they have a 0.0905 chance of having the disease.
- i.e. there is just a 9% chance that they have it.
- Even though the test is very accurate, because the condition is so rare the test may not be useful.

# So why is Bayesian probability useful?



- It allows us to put probability values on unknowns. We can make logical inferences even regarding uncertain statements.
- This can show counterintuitive results – e.g. that the disease test may not be useful.



# Probability Distribution

# Random Variables



A random variable is a numerical description of the outcome of an experiment.

A discrete random variable may assume either a finite number of values or an infinite sequence of values.

A continuous random variable may assume any numerical value in an interval or collection of intervals.

# Discrete Random Variable with a Finite Number of Values



- Example: JSL Appliances

Let  $x$  = number of TVs sold at the store in one day,  
where  $x$  can take on 5 values (0, 1, 2, 3, 4)

We can count the TVs sold, and there is a finite upper limit on the number that might be sold (which is the number of TVs in stock).

# Discrete Random Variable with an Infinite Sequence of Values



Example: JSL Appliances

Let  $x$  = number of customers arriving in one day,  
where  $x$  can take on the values 0, 1, 2, . . .

We can count the customers arriving, but there is no finite upper limit on the number that might arrive.

# Random Variables



Question	Random Variable $x$	Type
Family size	$x = \text{Number of dependents reported on tax return}$	Discrete
Distance from home to store	$x = \text{Distance in miles from home to the store site}$	Continuous
Own dog or cat	$x = 1 \text{ if own no pet;} \\ = 2 \text{ if own dog(s) only;} \\ = 3 \text{ if own cat(s) only;} \\ = 4 \text{ if own dog(s) and cat(s)}$	Discrete

# Probability Distributions



The probability distribution for a random variable describes how probabilities are distributed over the values of the random variable.

We can describe a discrete probability distribution with a table, graph, or formula.

# Discrete Probability Distributions



**Two types of discrete probability distributions:**

**First type:** uses the rules of assigning probabilities to experimental outcomes to determine probabilities for each value of the random variable.

**Second type:** uses a special mathematical formula to compute the probabilities for each value of the random variable.

# Discrete Probability Distributions



The probability distribution is defined by a probability function, denoted by  $f(x)$ , that provides the probability for each value of the random variable.

**The required conditions for a discrete probability function are:**

$$f(x) \geq 0$$

$$\sum f(x) = 1$$

# Discrete Probability Distributions



**There are three methods for assign probabilities to random variables: the classical method, the subjective method, and the relative frequency method.**

**The use of the relative frequency method to develop discrete probability distributions leads to what is called an empirical discrete distribution.**

example  
on next  
slide

# Discrete Probability Distributions



Example: JSL Appliances

Using past data on TV sales, ...

a tabular representation of the probability distribution for TV sales was developed.

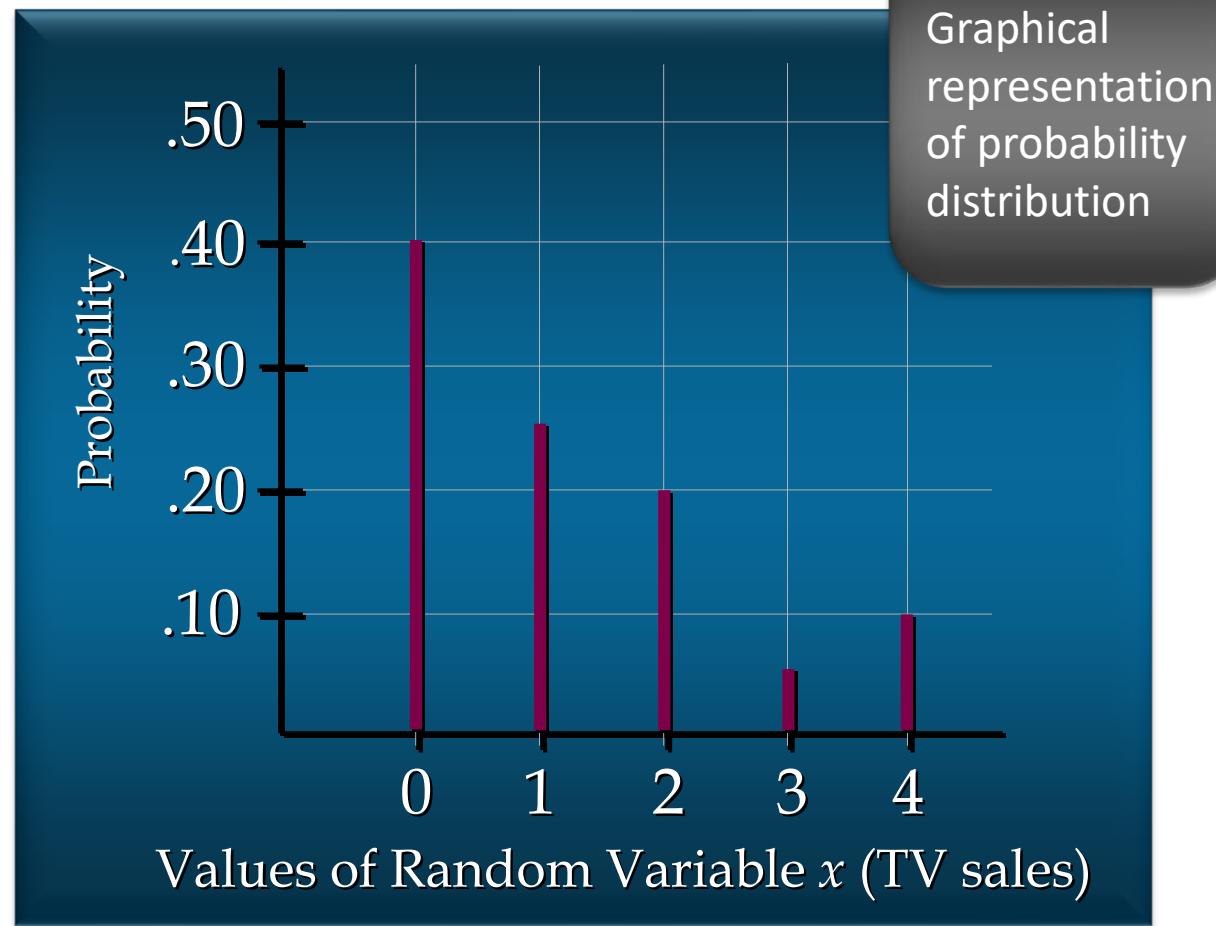
<u>Units Sold</u>	Number <u>of Days</u>	<u>x</u>	<u>f(x)</u>
0	80	0	.40
1	50	1	.25
2	40	2	.20
3	10	3	.05
4	<u>20</u>	4	<u>.10</u>
	200		1.00

80/200

# Discrete Probability Distributions



Example: JSL Appliances



# Discrete Probability Distributions



In addition to tables and graphs, a formula that gives the probability function,  $f(x)$ , for every value of  $x$  is often used to describe the probability distributions.

Several discrete probability distributions specified by formulas are the discrete -uniform, binomial, Poisson.

# Discrete Uniform Probability Distribution



The discrete uniform probability distribution is the simplest example of a discrete probability distribution given by a formula.

The discrete uniform probability function is

$$f(x) = 1/n$$

the values of the random variable  
are equally likely

where:

$n$  = the number of values the random variable may assume

Expected Mean  $E(x) = (n+1)/2$

Variance =  $(n^2-1)/12$

# Bivariate Distributions

A probability distribution involving two random variables is called a bivariate probability distribution.

Each outcome of a bivariate experiment consists of two values, one for each random variable.

Example: rolling a pair of dice

When dealing with bivariate probability distributions, we are often interested in the relationship between the random variables.

# Expected Value



The expected value, or mean, of a random variable is a measure of its central location.

$$E(x) = \mu = \Sigma xf(x)$$

The expected value is a weighted average of the values the random variable may assume. The weights are the probabilities.

The expected value does not have to be a value the random variable can assume.

# Variance and Standard Deviation



The variance summarizes the variability in the values of a random variable.

$$Var(x) = \sigma^2 = \sum(x - \mu)^2 f(x)$$

The variance is a weighted average of the squared deviations of a random variable from its mean.  
The weights are the probabilities.

The standard deviation,  $\sigma$ , is defined as the positive square root of the variance.

# A Bivariate Discrete Probability Distribution

A company asked 200 of its employees how they rated their benefit package and job satisfaction. The cross tabulation below shows the ratings data.

Benefits Package ( $x$ )	Job Satisfaction ( $y$ )			Total
	1	2	3	
1	28	26	4	58
2	22	42	34	98
3	2	10	32	44
Total	52	78	70	200

# A Bivariate Discrete Probability Distribution

The bivariate empirical discrete probabilities for benefits rating and job satisfaction are shown below.

Benefits Package ( $x$ )	<u>Job Satisfaction (<math>y</math>)</u>			Total
	1	2	3	
1	.14	.13	.02	.29
2	.11	.21	.17	.49
3	.01	.05	.16	.22
Total	.26	.39	.35	1.00

# A Bivariate Discrete Probability Distribution

Expected Value and Variance for Benefits Package,  $x$

<u><math>x</math></u>	<u><math>f(x)</math></u>	<u><math>xf(x)</math></u>	<u><math>x - E(x)</math></u>	<u><math>(x - E(x))^2</math></u>	<u><math>(x - E(x))^2f(x)</math></u>
1	0.29	0.29	-0.93	0.8649	0.250821
2	0.49	0.98	0.07	0.0049	0.002401
3	0.22	<u>0.66</u>	1.07	1.1449	<u>0.251878</u>
$E(x) =$		1.93	$Var(x) =$		0.505100

# A Bivariate Discrete Probability Distribution



Expected Value and Variance for Job Satisfaction,  $y$

$y$	$f(y)$	$yf(y)$	$y - E(y)$	$(y - E(y))^2$	$(y - E(y))^2 f(y)$
1	0.26	0.26	-1.09	1.1881	0.308906
2	0.39	0.78	-0.09	0.0081	0.003159
3	0.35	<u>1.05</u>	0.91	0.8281	<u>0.289835</u>
$E(y) =$		2.09	$Var(y) =$		0.601900

# A Bivariate Discrete Probability Distribution

Expected Value and Variance for Bivariate Distrib.

<u>s</u>	<u>f(s)</u>	<u>sf(s)</u>	<u>s - E(s)</u>	<u>(s - E(s))<sup>2</sup></u>	<u>(s - E(s))<sup>2</sup>f(s)</u>
2	0.14	0.28	-2.02	4.0804	0.571256
3	0.24	0.72	-1.02	1.0404	0.249696
4	0.24	0.96	-0.02	0.0004	0.000960
5	0.22	1.10	0.98	0.9604	0.211376
6	0.16	<u>0.96</u>	1.98	3.9204	<u>0.627264</u>
	$E(s) =$	4.02		$Var(s) =$	1.660552

# A Bivariate Discrete Probability Distribution

Covariance for Random Variables  $x$  and  $y$

$$Var_{xy} = [Var(x + y) - Var(x) - Var(y)]/2$$

$$Var_{xy} = [1.660552 - 0.5051 - 0.6019]/2 = 0.276776$$

# A Bivariate Discrete Probability Distribution

Correlation Between Variables  $x$  and  $y$

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$$\sigma_x = \sqrt{0.5051} = 0.7107038$$

$$\sigma_y = \sqrt{0.6019} = 0.7758221$$

$$\sigma_{xy} = \sqrt{0.276776} = 0.526095$$

$$\rho_{xy} = \frac{0.526095}{0.7107038(0.7758221)} = 0.954$$

# Binomial Probability Distribution



## Four Properties of a Binomial Experiment

1. The experiment consists of a sequence of  $n$  identical trials.
2. Two outcomes, success and failure, are possible on each trial.
3. The probability of a success, denoted by  $p$ , does not change from trial to trial.
4. The trials are independent.

stationarity  
assumption

# Binomial Probability Distribution

Our interest is in the number of successes occurring in the  $n$  trials.

Let  $x$  denote the number of successes occurring in the  $n$  trials.

# Binomial Probability Distribution



## Binomial Probability Function

$$f(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{(n-x)}$$

**where:**

**$x$  = the number of successes**

**$p$  = the probability of a success on one trial**

**$n$  = the number of trials**

**$f(x)$  = the probability of  $x$  successes in  $n$  trials**

**$n! = n(n - 1)(n - 2) \dots (2)(1)$**

# Binomial Probability Distribution



## Binomial Probability Function

$$f(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{(n-x)}$$

Number of experimental outcomes providing exactly  $x$  successes in  $n$  trials

Probability of a particular sequence of trial outcomes with  $x$  successes in  $n$  trials

# Binomial Probability Distribution



## Example: Evans Electronics

Evans Electronics is concerned about a low retention rate for its employees. In recent years, management has seen a turnover of 10% of the employees annually.

Thus, for any employee chosen at random, management estimates a probability of 0.1 that the person will not be with the company next year.

Choosing 3 employees at random, what is the probability that 1 of them will leave the company this year?

# Binomial Probability Distribution



Example: Evans Electronics

The probability of the first employee leaving and the second and third employees staying, denoted  $(S, F, F)$ , is given by

$$p(1 - p)(1 - p)$$

With a .10 probability of an employee leaving on any one trial, the probability of an employee leaving on the first trial and not on the second and third trials is given by

$$(.10)(.90)(.90) = (.10)(.90)^2 = .081$$

# Binomial Probability Distribution



## Example: Evans Electronics

Two other experimental outcomes also result in one success and two failures. The probabilities for all three experimental outcomes involving one success follow.

<u>Experimental Outcome</u>	<u>Probability of Experimental Outcome</u>
(S, F, F)	$p(1 - p)(1 - p) = (.1)(.9)(.9) = .081$
(F, S, F)	$(1 - p)p(1 - p) = (.9)(.1)(.9) = .081$
(F, F, S)	$(1 - p)(1 - p)p = (.9)(.9)(.1) = \underline{.081}$
	Total = <u>.243</u>

# Binomial Probability Distribution



Example: Evans Electronics

Using the  
probability  
function

Let:  $p = .10$ ,  $n = 3$ ,  $x = 1$

$$f(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{(n-x)}$$

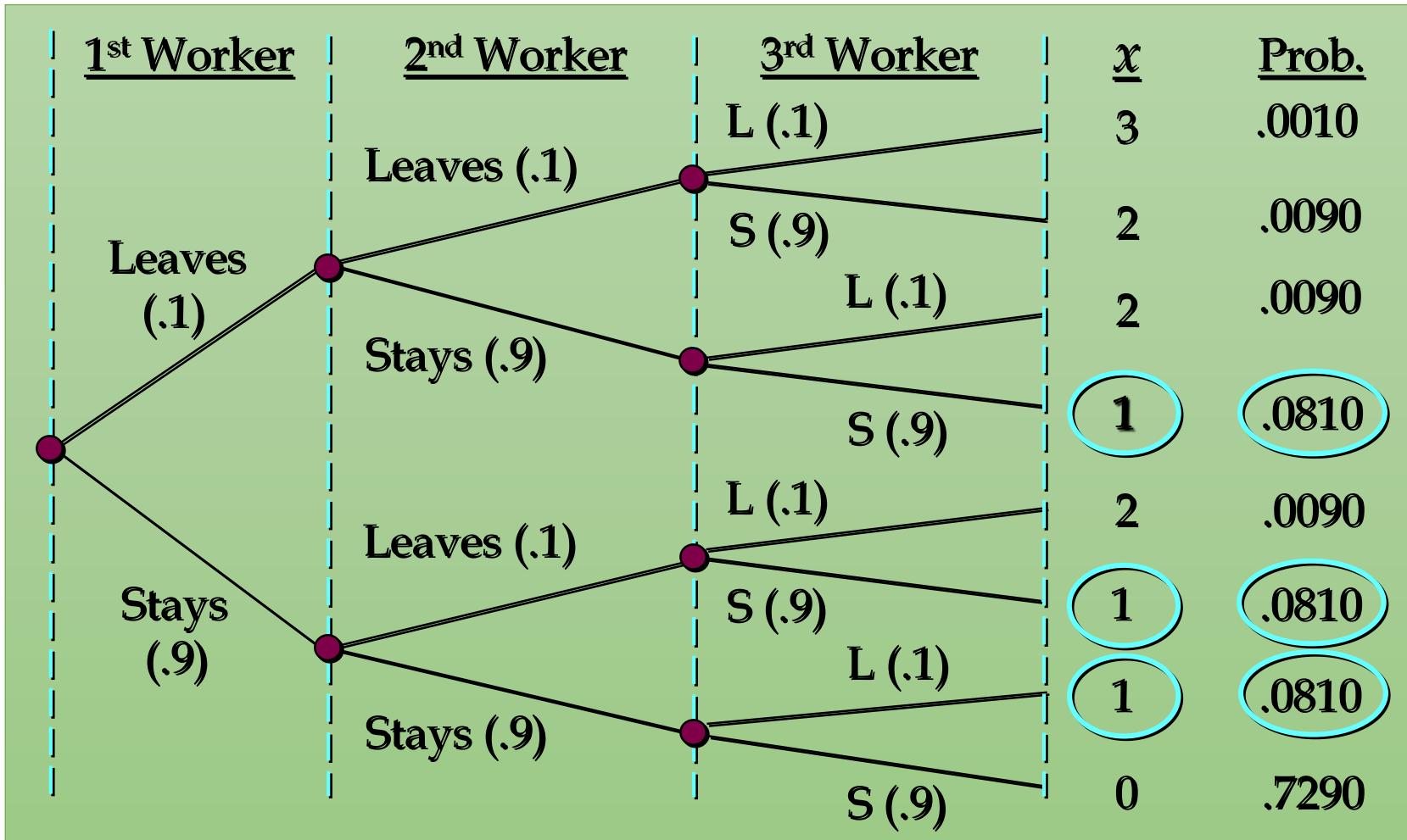
$$f(1) = \frac{3!}{1!(3-1)!} (0.1)^1 (0.9)^2 = 3(.1)(.81) = \textcircled{.243}$$

# Binomial Probability Distribution



Example: Evans Electronics

Using a tree diagram



# Binomial Probabilities and Cumulative Probabilities



Statisticians have developed tables that give probabilities and cumulative probabilities for a binomial random variable.

These tables can be found in some statistics textbooks.

With modern calculators and the capability of statistical software packages, such tables are almost unnecessary.

# Binomial Probability Distribution



## Using Tables of Binomial Probabilities

n	x	p									
		.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
3	0	.8574	.7290	.6141	.5120	.4219	.3430	.2746	.2160	.1664	.1250
	1	.1354	.2430	.3251	.3840	.4219	.4410	.4436	.4320	.4084	.3750
	2	.0071	.0270	.0574	.0960	.1406	.1890	.2389	.2880	.3341	.3750
	3	.0001	.0010	.0034	.0080	.0156	.0270	.0429	.0640	.0911	.1250

# Binomial Probability Distribution



## Expected Value

$$E(x) = \mu = np$$

## Variance

$$Var(x) = \sigma^2 = np(1 - p)$$

## Standard Deviation

$$\sigma = \sqrt{np(1 - p)}$$

# Binomial Probability Distribution



**Example:** Evans Electronics

**Expected Value**

$$E(x) = np = 3(.1) = .3 \text{ employees out of 3}$$

**Variance**

$$Var(x) = np(1 - p) = 3(.1)(.9) = .27$$

**Standard Deviation**

$$\sigma = \sqrt{3(.1)(.9)} = .52 \text{ employees}$$

# Poisson Probability Distribution



A Poisson distributed random variable is often useful in estimating the number of occurrences over a specified interval of time or space

It is a discrete random variable that may assume an infinite sequence of values ( $x = 0, 1, 2, \dots$ ).

# Poisson Probability Distribution



Examples of Poisson distributed random variables:

**the number of customers withdrawing  
money from ATM in odd hours**

**the number of vehicles arriving at a toll  
booth in one hour**

Bell Labs used the Poisson distribution to model  
the arrival of phone calls.

# Poisson Probability Distribution



## Two Properties of a Poisson Experiment

1. The probability of an occurrence is the same for any two intervals of equal length.
2. The occurrence or nonoccurrence in any interval is independent of the occurrence or nonoccurrence in any other interval.

# Poisson Probability Distribution



## Poisson Probability Function

$$f(x) = \frac{\mu^x e^{-\mu}}{x!}$$

where:

$x$  = the number of occurrences in an interval

$f(x)$  = the probability of  $x$  occurrences in an interval

$\mu$  = mean number of occurrences in an interval

$e = 2.71828$

$x! = x(x - 1)(x - 2) \dots (2)(1)$

# Poisson Probability Distribution



## Poisson Probability Function

Since there is no stated upper limit for the number of occurrences, the probability function  $f(x)$  is applicable for values  $x = 0, 1, 2, \dots$  without limit.

In practical applications,  $x$  will eventually become large enough so that  $f(x)$  is approximately zero and the probability of any larger values of  $x$  becomes negligible.

# Poisson Probability Distribution



Example: Mercy Hospital

Patients arrive at the emergency room of Mercy Hospital at the average rate of 6 per hour on weekend evenings.

What is the probability of 4 arrivals in 30 minutes on a weekend evening?

# Poisson Probability Distribution

Example: Mercy Hospital

$$\mu = 6/\text{hour} = 3/\text{half-hour}, \quad x = 4$$

$$f(4) = \frac{3^4 (2.71828)^{-3}}{4!} = .1680$$

Using the  
probability  
function

# Poisson Probability Distribution



Example: Mercy Hospital

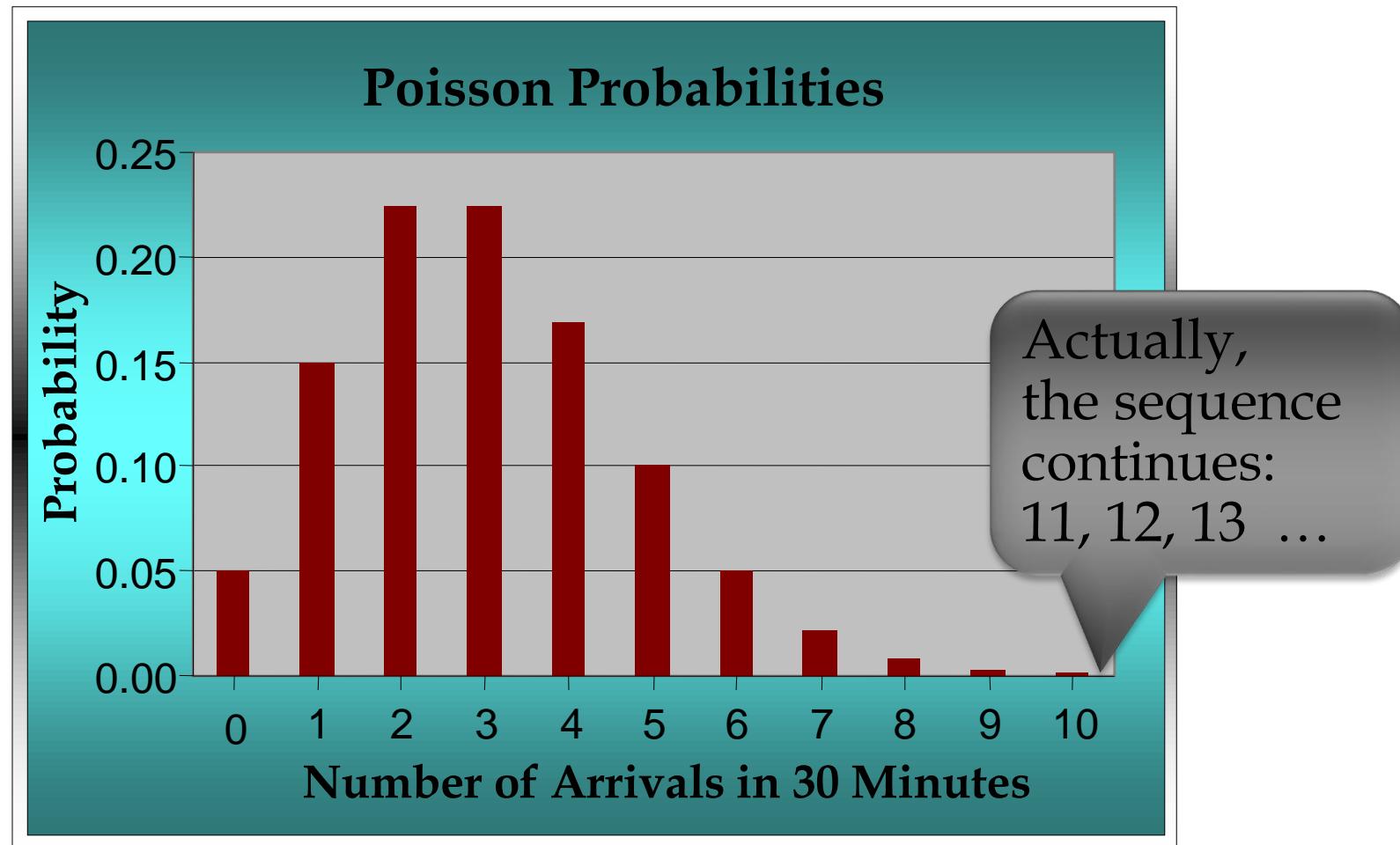
Variance for Number of Arrivals  
During 30-Minute Periods

$$\mu = \sigma^2 = 3$$

# Poisson Probability Distribution



Example: Mercy Hospital



# Poisson Probability Distribution



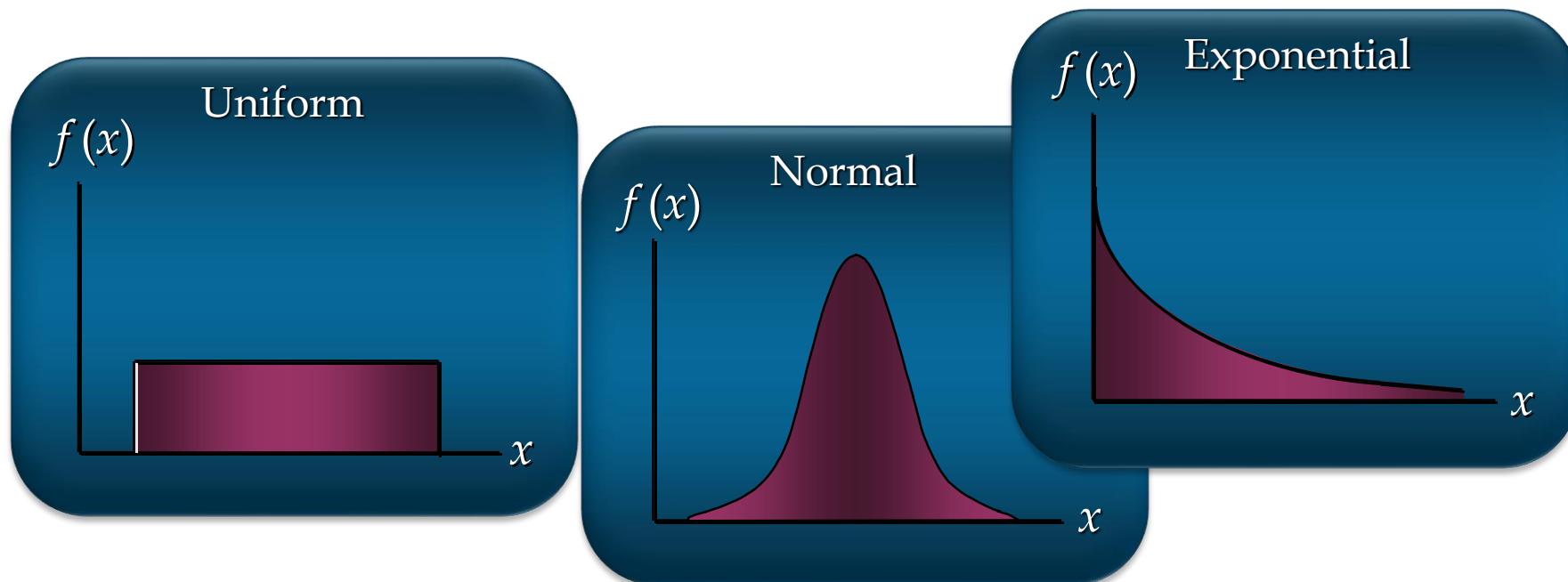
A property of the Poisson distribution is that the mean and variance are equal.

$$\mu = \sigma^2$$

# Continuous Probability Distributions



- Uniform Probability Distribution
- Normal Probability Distribution
- Exponential Probability Distribution



# Continuous Probability Distributions

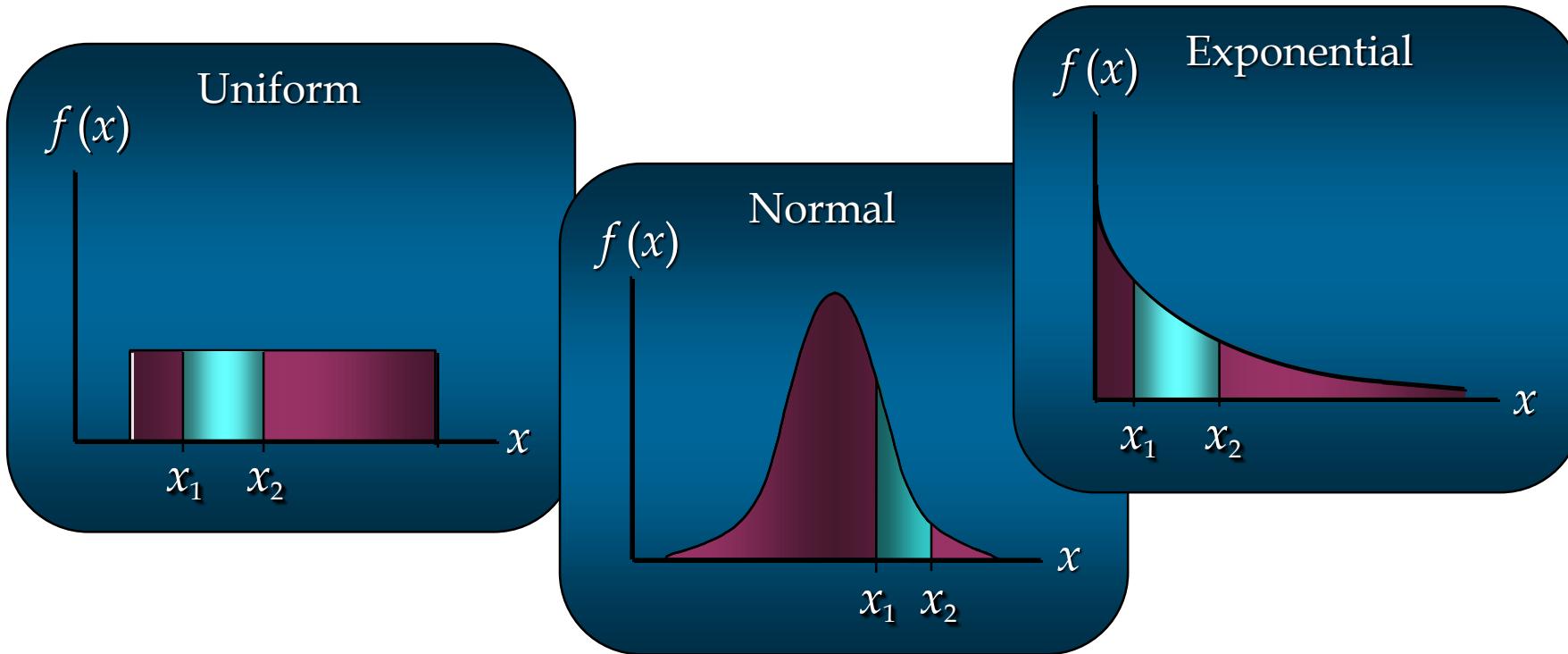


- A continuous random variable can assume any value in an interval on the real line or in a collection of intervals.
- It is not possible to talk about the probability of the random variable assuming a particular value.
- Instead, we talk about the probability of the random variable assuming a value within a given interval.

# Continuous Probability Distributions



The probability of the random variable assuming a value within some given interval from  $x_1$  to  $x_2$  is defined to be the area under the graph of the probability density function between  $x_1$  and  $x_2$ .



# Uniform Probability Distribution



- A random variable is uniformly distributed whenever the probability is proportional to the interval's length.
- The uniform probability density function is:

$$\begin{aligned} f(x) &= 1/(b - a) && \text{for } a < x < b \\ &= 0 && \text{elsewhere} \end{aligned}$$

where:  $a$  = smallest value the variable can assume

$b$  = largest value the variable can assume

# Uniform Probability Distribution



Expected Value of  $x$

$$E(x) = (a + b)/2$$

Variance of  $x$

$$\text{Var}(x) = (b - a)^2/12$$

# Uniform Probability Distribution



## Example: Slater's Buffet

Slater customers are charged for the amount of salad they take. Sampling suggests that the amount of salad taken is uniformly distributed between 5 ounces and 15 ounces.

# Uniform Probability Distribution



## Uniform Probability Density Function

$$\begin{aligned} f(x) &= 1/10 \quad \text{for } 5 < x < 15 \\ &= 0 \quad \quad \quad \text{elsewhere} \end{aligned}$$

where:

$x$  = salad plate filling weight

# Uniform Probability Distribution



Expected Value of  $x$

$$\begin{aligned} E(x) &= (a + b)/2 \\ &= (5 + 15)/2 \\ &= 10 \end{aligned}$$

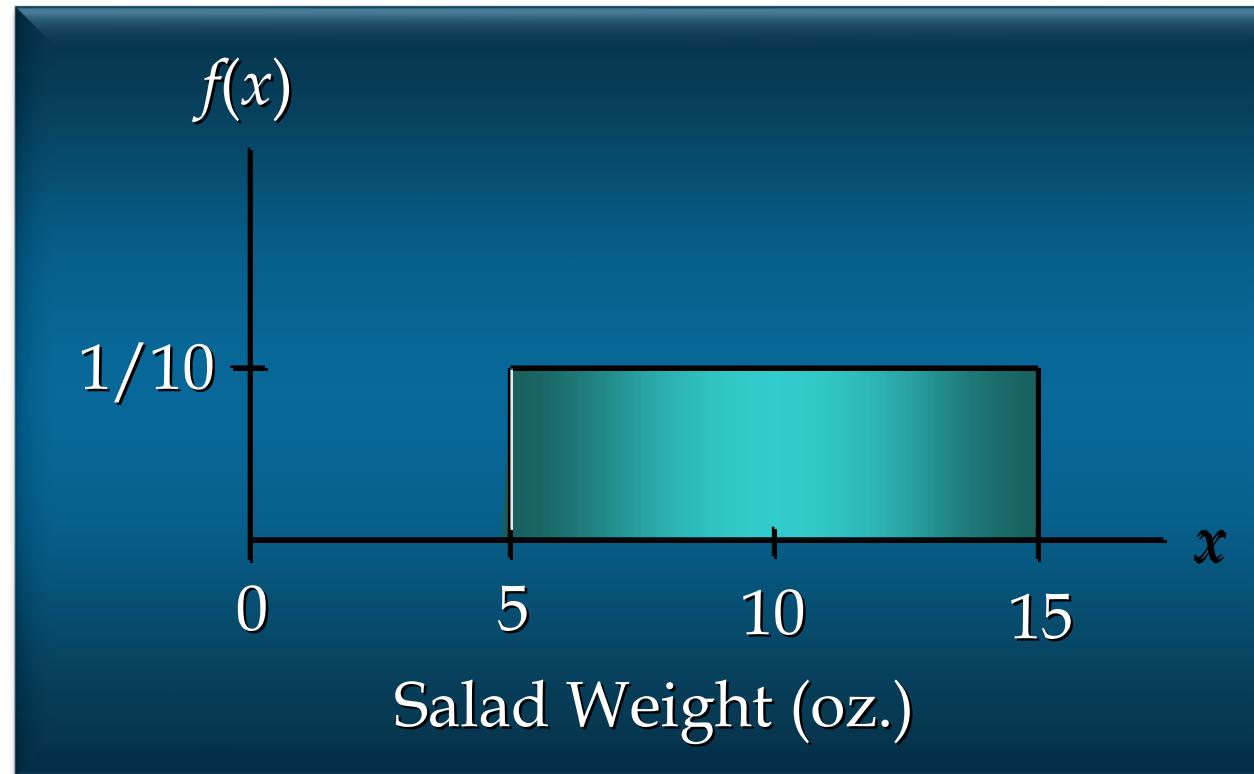
Variance of  $x$

$$\begin{aligned} \text{Var}(x) &= (b - a)^2/12 \\ &= (15 - 5)^2/12 \\ &= 8.33 \end{aligned}$$

# Uniform Probability Distribution



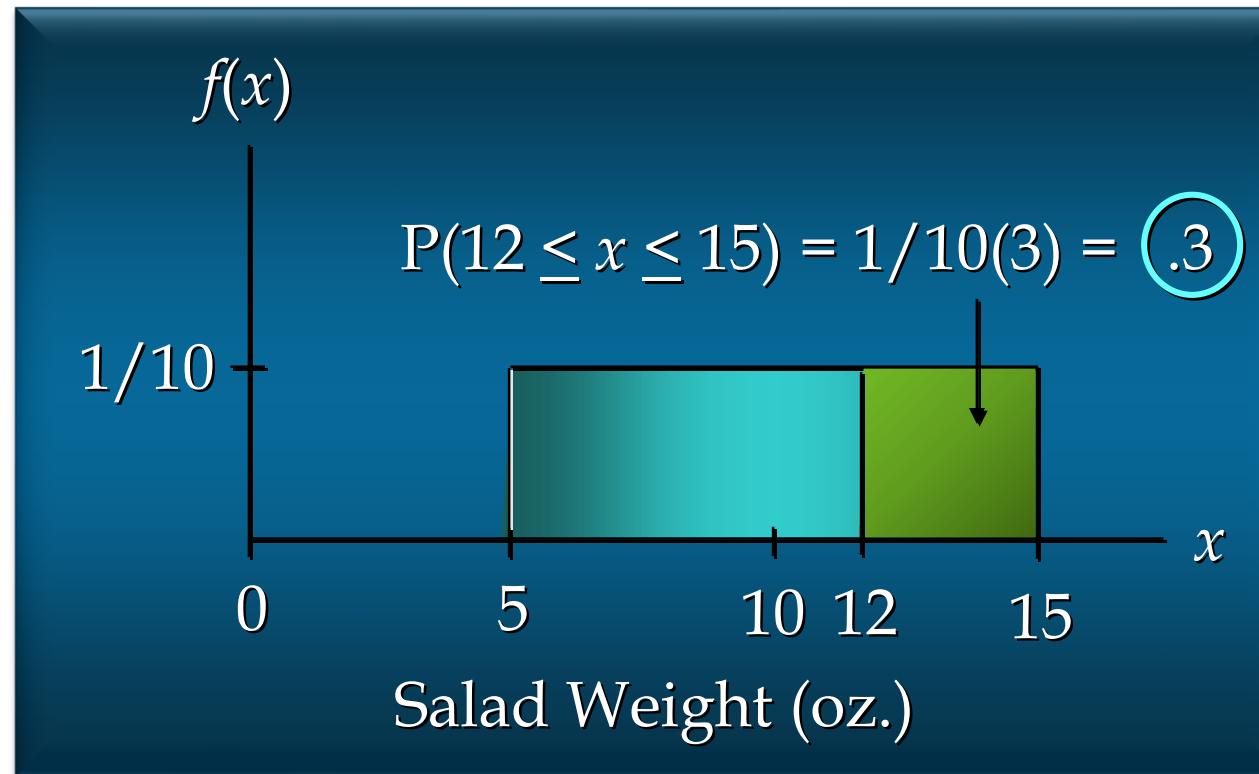
- Uniform Probability Distribution for Salad Plate Filling Weight



# Uniform Probability Distribution



What is the probability that a customer will take between 12 and 15 ounces of salad?



# Area as a Measure of Probability



- The area under the graph of  $f(x)$  and probability are identical.
- This is valid for all continuous random variables.
- The probability that  $x$  takes on a value between some lower value  $x_1$  and some higher value  $x_2$  can be found by computing the area under the graph of  $f(x)$  over the interval from  $x_1$  to  $x_2$ .

# Normal Probability Distribution



- The normal probability distribution is the most important distribution for describing a continuous random variable.
- It is widely used in statistical inference.
- It has been used in a wide variety of applications including:
  - Heights of people
  - Rainfall amounts
  - Test scores
  - Scientific measurements
- Abraham de Moivre, a French mathematician, published *The Doctrine of Chances* in 1733.
- He derived the normal distribution.

# Normal Probability Distribution



- Normal Probability Density Function

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

where:

$\mu$  = mean

$\sigma$  = standard deviation

$\pi$  = 3.14159

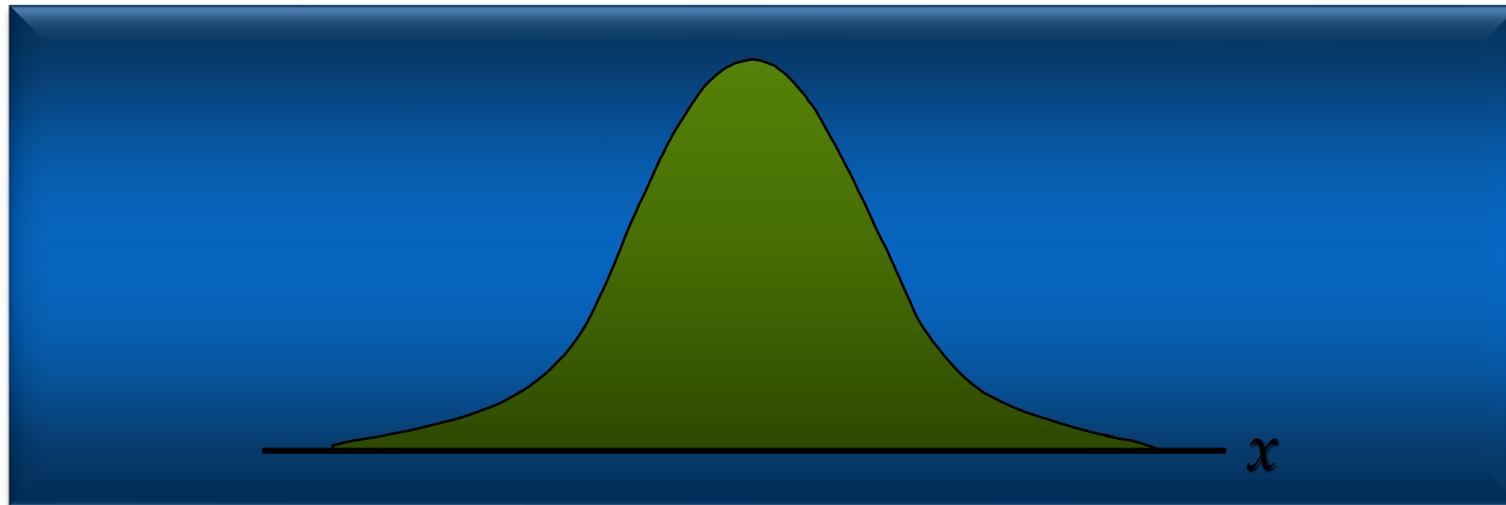
$e$  = 2.71828

# Normal Probability Distribution



## Characteristics

The distribution is symmetric; its skewness measure is zero.

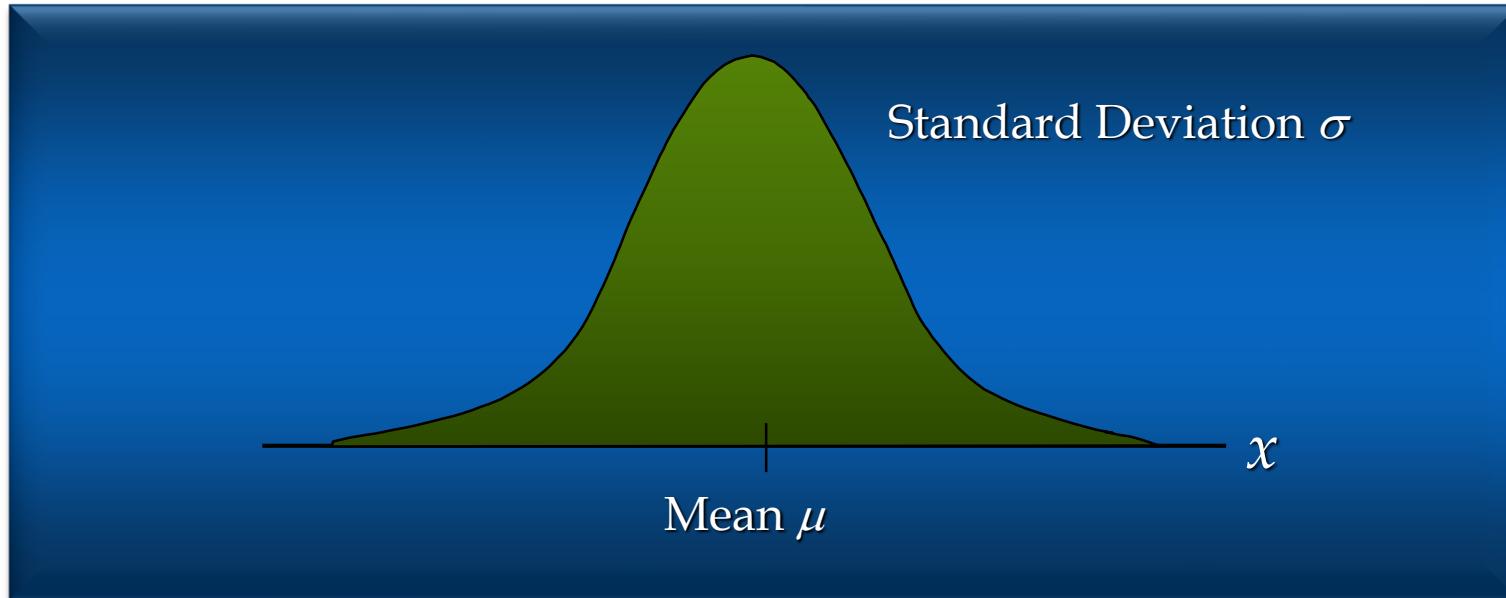


# Normal Probability Distribution



## □ Characteristics

The entire family of normal probability distributions is defined by its mean  $\mu$  and its standard deviation  $\sigma$ .

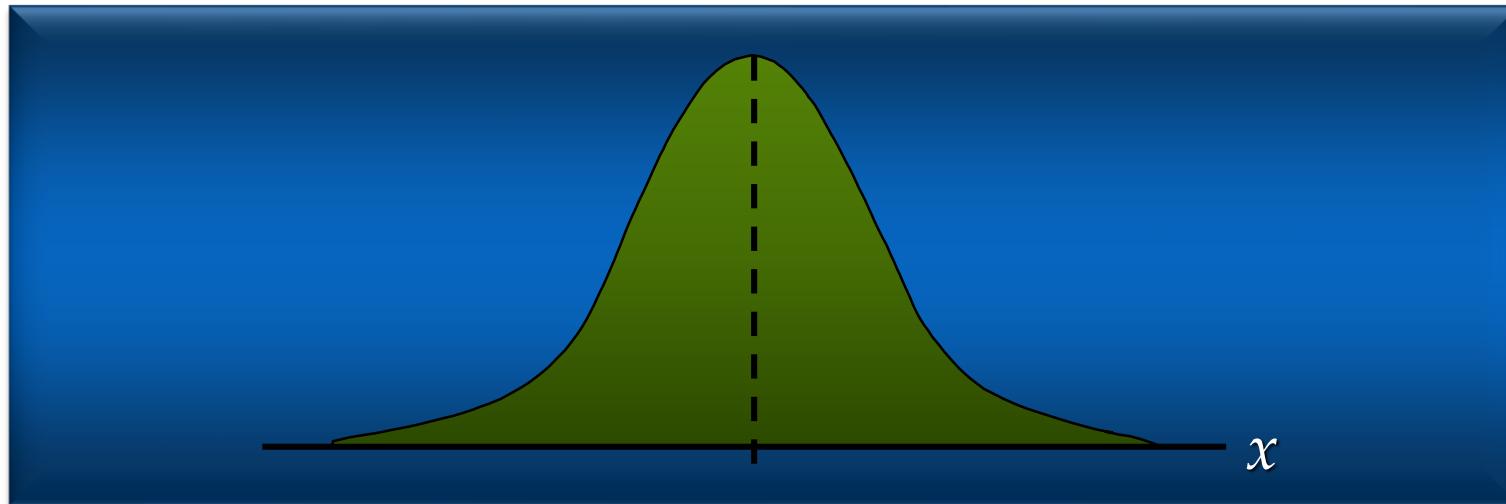


# Normal Probability Distribution



## Characteristics

The highest point on the normal curve is at the mean, which is also the median and mode.

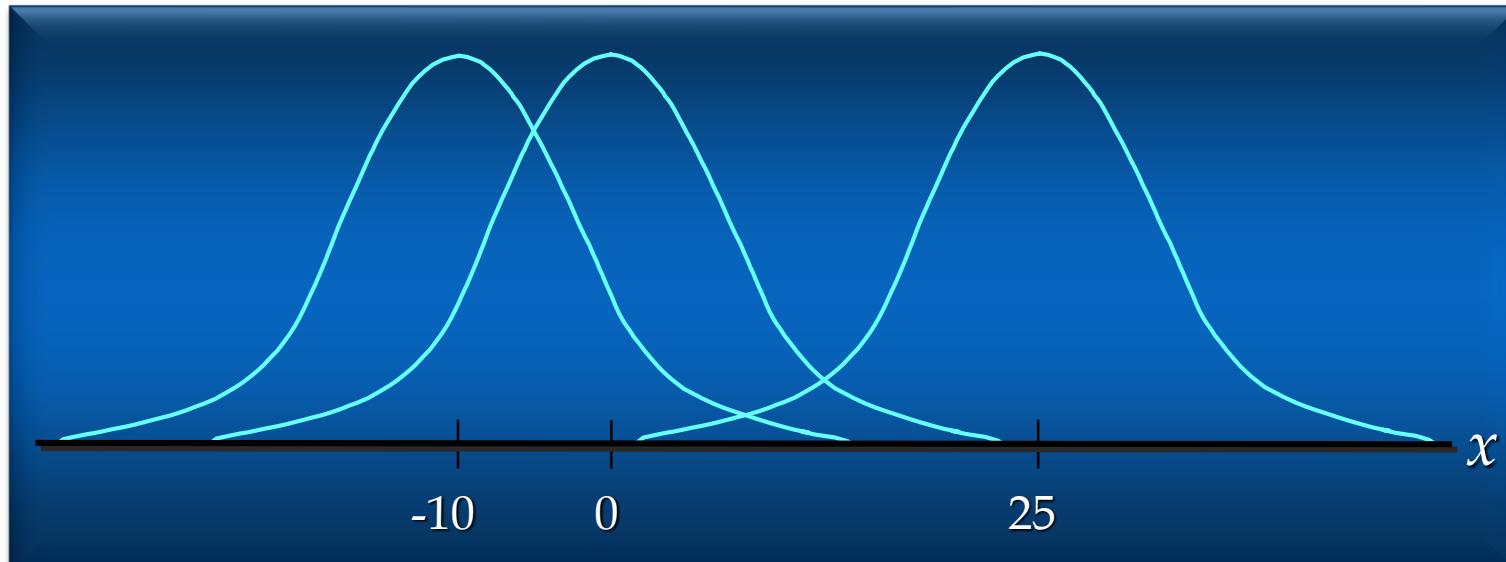


# Normal Probability Distribution



## Characteristics

The mean can be any numerical value: negative, zero, or positive.

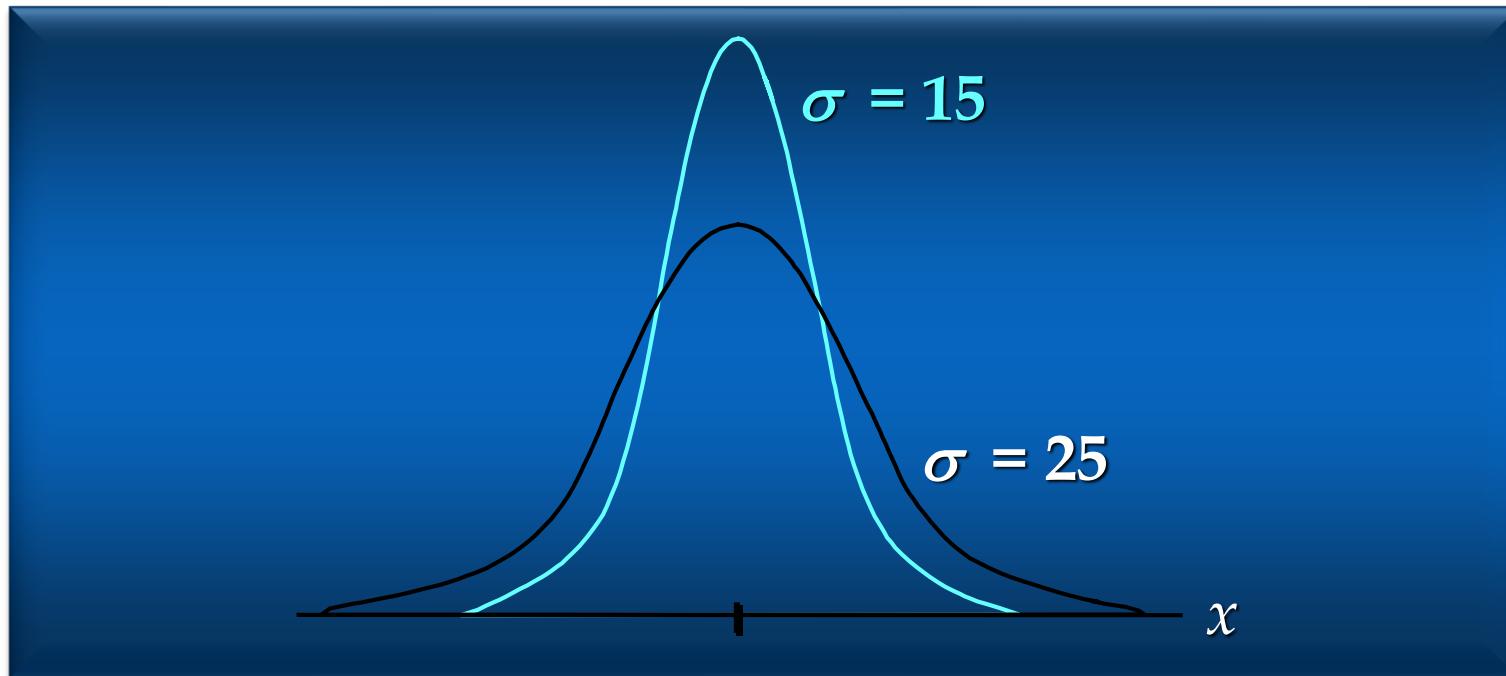


# Normal Probability Distribution



## Characteristics

The standard deviation determines the width of the curve: larger values result in wider, flatter curves.

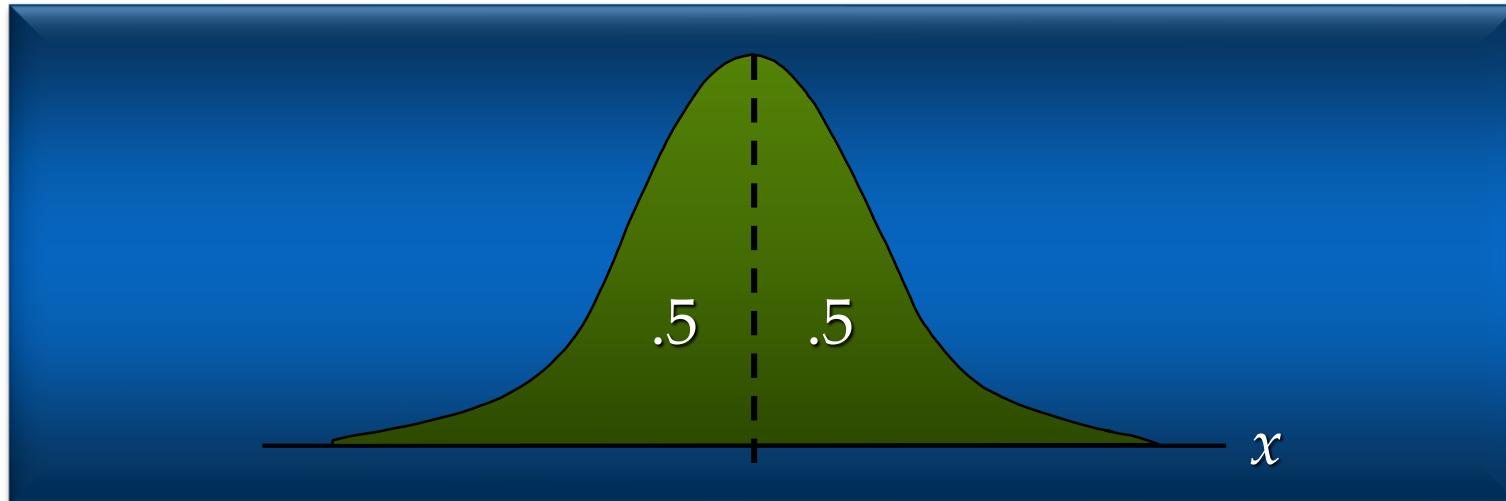


# Normal Probability Distribution



## Characteristics

Probabilities for the normal random variable are given by areas under the curve. The total area under the curve is 1 (.5 to the left of the mean and .5 to the right).



# Normal Probability Distribution



## Characteristics (basis for the empirical rule)

68.26% of values of a normal random variable are within  $+/- 1$  standard deviation of its mean.

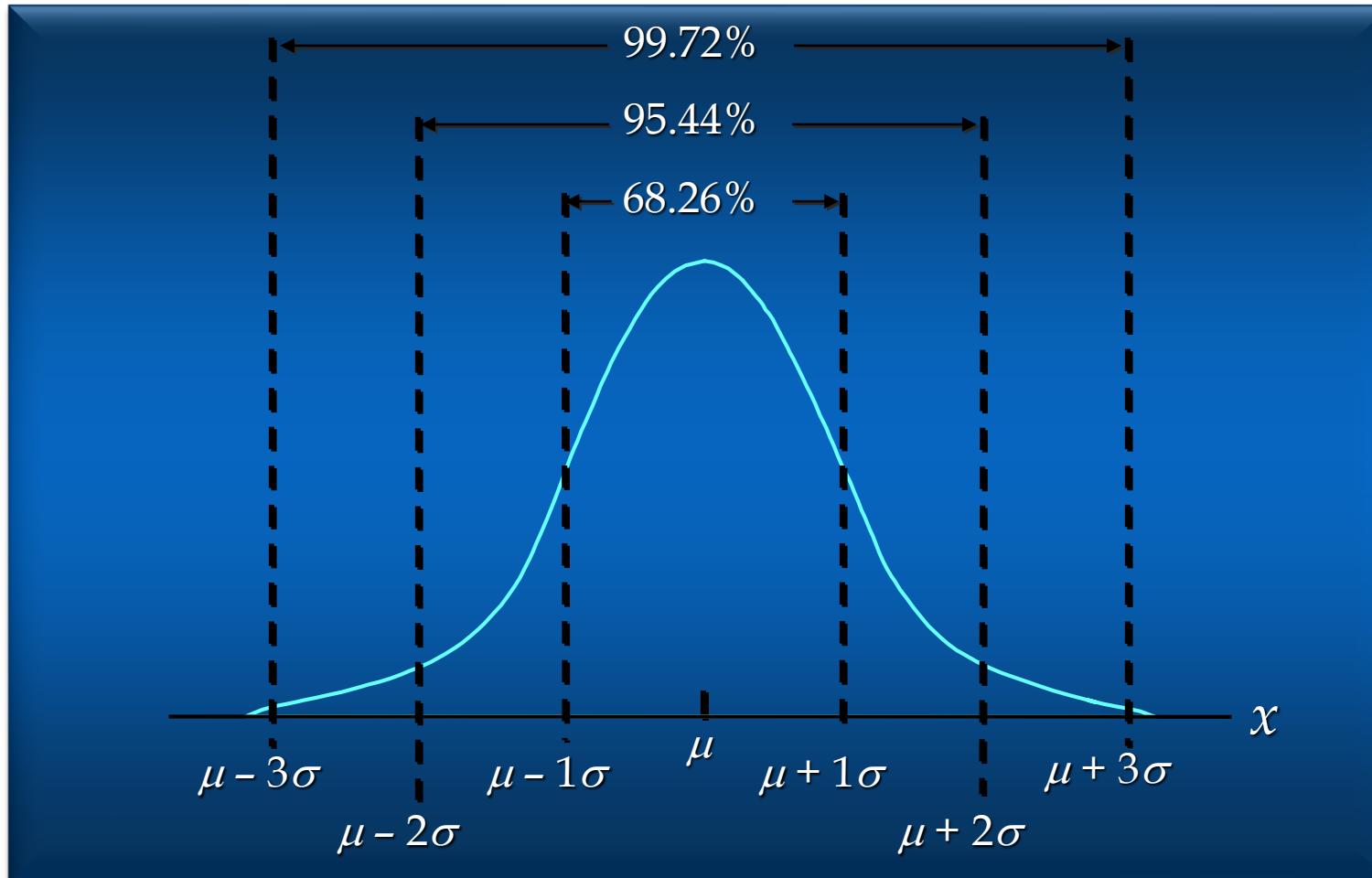
95.44% of values of a normal random variable are within  $+/- 2$  standard deviations of its mean.

99.72% of values of a normal random variable are within  $+/- 3$  standard deviations of its mean.

# Normal Probability Distribution

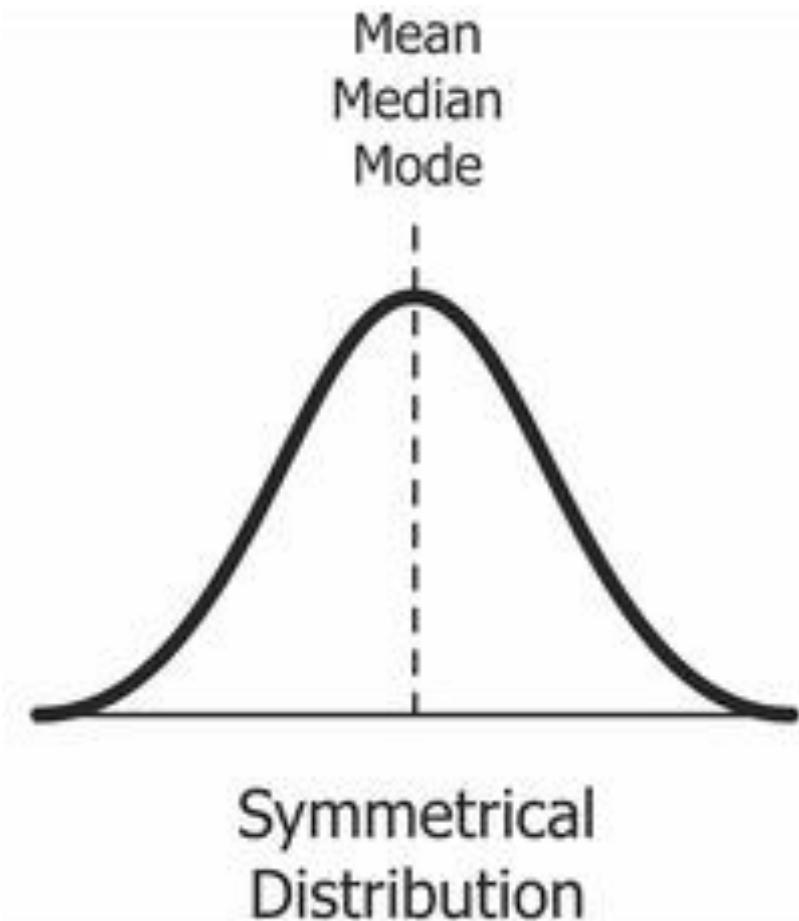


Characteristics (basis for the empirical rule)



# SKEWNESS

- Measure of the symmetry in a distribution.
- A symmetrical dataset will have a skewness equal to 0.
- A normal distribution will have a skewness of 0.
- Skewness essentially measures the relative size of the two tails.



# SKEWNESS

- The skewness is defined as

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1$$

where

$$g_1 = m_3 / m_2^{3/2}$$

with

$$m_3 = \sum(x - \bar{x})^3 / n,$$

$$m_2 = \sum(x - \bar{x})^2 / n,$$

$\bar{x}$  is the mean,

n is the sample size.

$m_3$  is called the third moment of the data set.

$m_2$  is the variance (square of the standard deviation).

# EXAMPLE

- Here, sample size is 100.
- By computation, mean is 67.45. ( $\sum xf$ )

Class Mark, $x$	Frequency, $f$	$xf$	$(x-\bar{x})$	$(x-\bar{x})^2f$	$(x-\bar{x})^3f$
61	5	305	-6.45	208.01	-1341.68
64	18	1152	-3.45	214.25	-739.15
67	42	2814	-0.45	8.51	-3.83
70	27	1890	2.55	175.57	447.70
73	8	584	5.55	246.42	1367.63
$\Sigma$		6745	n/a	852.75	-269.33
$\bar{x}, m_2, m_3$		67.45	n/a	8.5275	-2.6933

Example:

Height (inches)	Class Mark, $x$	Freq- uency, $f$
59.5-62.5	61	5
62.5-65.5	64	18
65.5-68.5	67	42
68.5-71.5	70	27
71.5-74.5	73	8

Skewness:

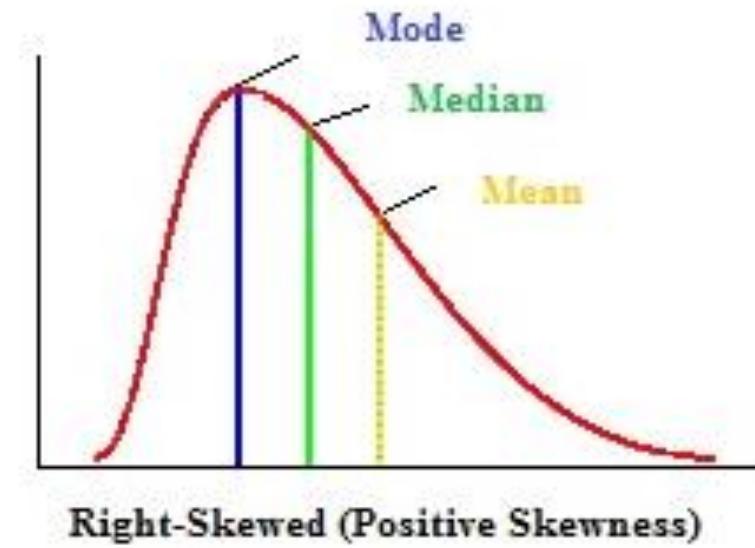
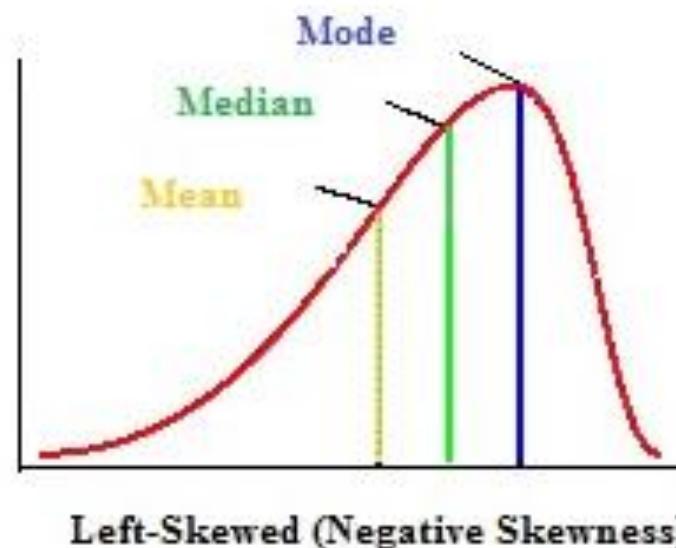
$$g_1 = m_3 / m_2^{3/2} = -2.6933 / 8.52753/2 = -0.1082$$

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1 = [\sqrt{(100 \times 99) / 98}] [-2.6933 / 8.5275^{3/2}] = -0.1098$$

With a skewness of -0.1098, the sample data for student heights are approximately symmetric.

# POSITIVE & NEGATIVE SKEWNESS

- If skewness is **positive**, the data are positively skewed or skewed right, meaning that the right tail of the distribution is longer than the left.
- If skewness is **negative**, the data are negatively skewed or skewed left, meaning that the left tail is longer.



# Skewness RULE of Thumb

The rule of thumb for Skewness:

- If skewness is less than  $-1$  or greater than  $+1$ , the distribution is highly skewed.
- If skewness is between  $-1$  and  $-\frac{1}{2}$  or between  $\frac{1}{2}$  and  $+1$ , the distribution is moderately skewed.
- If skewness is between  $-\frac{1}{2}$  and  $+\frac{1}{2}$ , the distribution is approximately symmetric.

# Kurtosis



- Kurtosis measures how peaked the histogram is

$$kurtosis = \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{ns^4} - 3$$

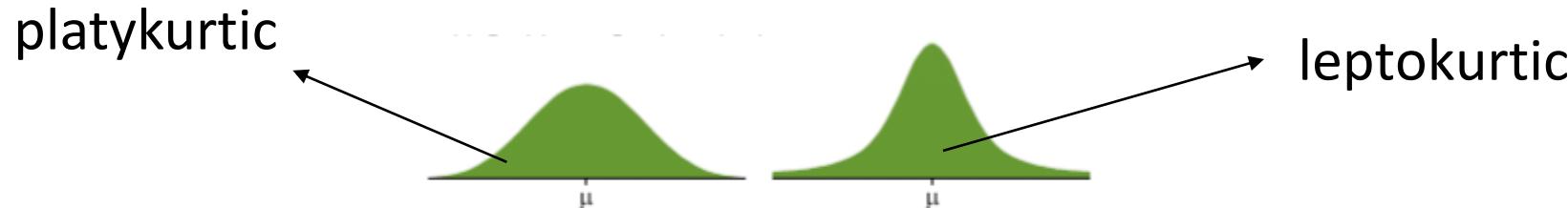
- The kurtosis of a normal distribution is 0
- Kurtosis characterizes the relative peakedness or flatness of a distribution compared to the normal distribution

# Kurtosis



- **Platykurtic**— When the kurtosis  $< 0$ , the frequencies throughout the curve are closer to be equal (i.e., the curve is more flat and wide)
- Thus, **negative kurtosis** indicates a relatively **flat** distribution
- **Leptokurtic**— When the kurtosis  $> 0$ , there are high frequencies in only a small part of the curve (i.e, the curve is more peaked)
- Thus, **positive kurtosis** indicates a relatively **peaked** distribution

# Kurtosis

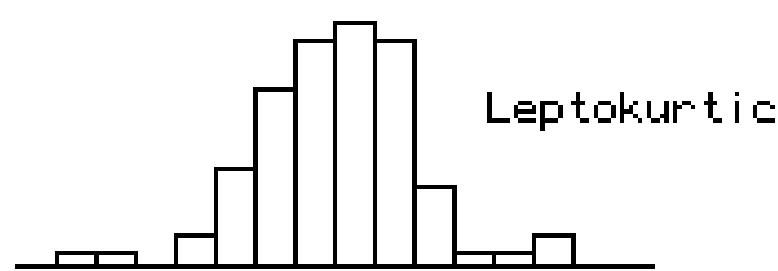


These graphs illustrate the notion of kurtosis. The PDF on the right has higher kurtosis than the PDF on the left. It is more peaked at the center, and it has fatter tails.

- **Kurtosis** is based on the size of a distribution's tails.
- **Negative kurtosis (platykurtic)** – distributions with short tails
- **Positive kurtosis (leptokurtic)** – distributions with relatively long tails

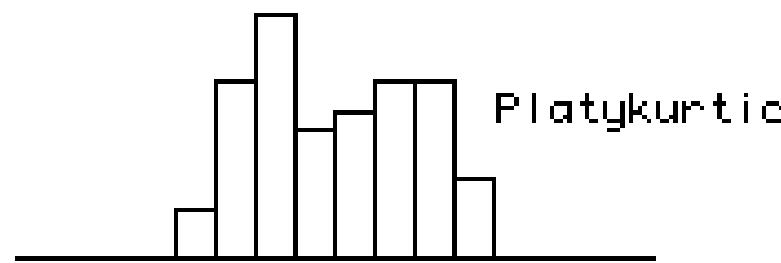
# Why we need Kurtosis

Kurtosis = 1.25



Leptokurtic

Kurtosis = -1.23



Platykurtic

- These two distributions have the same variance, approximately the same skew, but differ markedly in kurtosis.

# Standard Normal Probability Distribution



## Characteristics

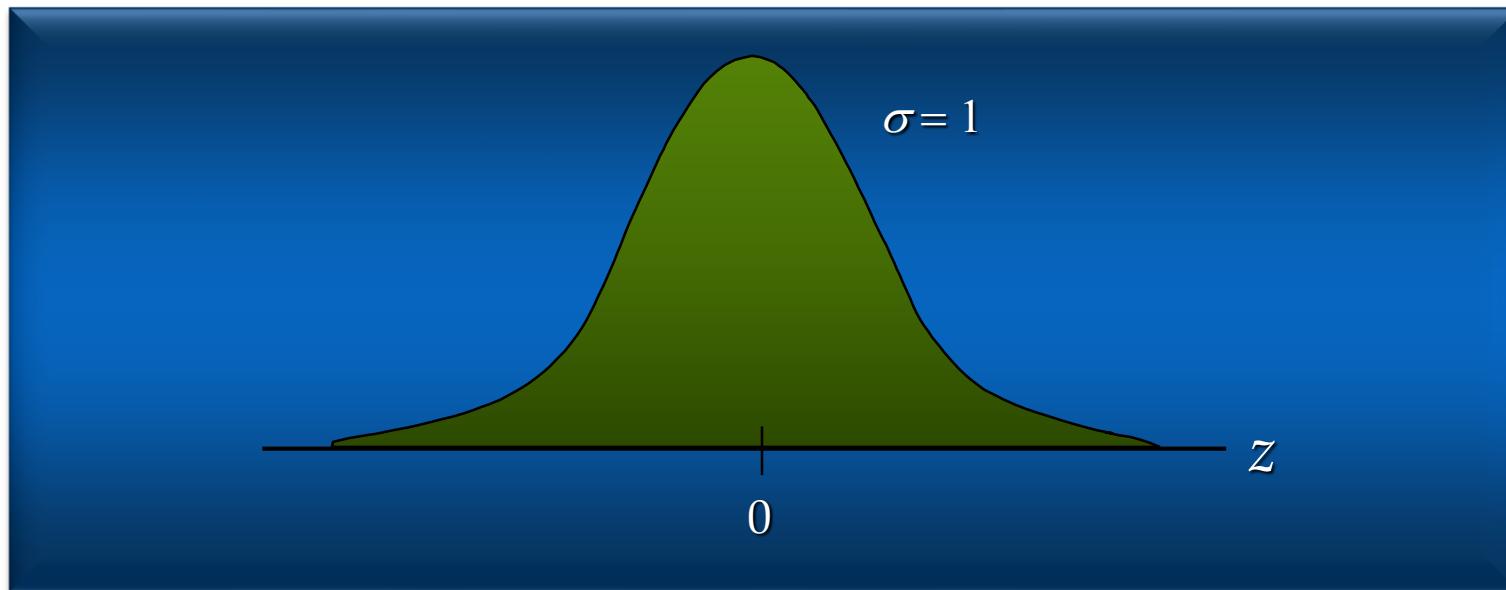
A random variable having a normal distribution with a mean of 0 and a standard deviation of 1 is said to have a standard normal probability distribution.

# Standard Normal Probability Distribution



## Characteristics

The letter  $z$  is used to designate the standard normal random variable.



# Standard Normal Probability Distribution



Converting to the Standard Normal Distribution

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

We can think of  $z$  as a measure of the number of standard deviations  $x$  is from  $\mu$ .

# Standard Normal Probability Distribution



## Example: Pep Zone

Pep Zone sells auto parts and supplies including a popular multi-grade motor oil. When the stock of this oil drops to 20 gallons, a replenishment order is placed.

The store manager is concerned that sales are being lost due to stock outs while waiting for a replenishment order.

# Standard Normal Probability Distribution



## Example: Pep Zone

It has been determined that demand during replenishment lead-time is normally distributed with a mean of 15 gallons and a standard deviation of 6 gallons.

The manager would like to know the probability of a stock out during replenishment lead-time. In other words, what is the probability that demand during lead-time will exceed 20 gallons?

$$P(x > 20) = ?$$

# Standard Normal Probability Distribution



Solving for the Stock out Probability

**Step 1: Convert  $x$  to the standard normal distribution.**

$$\begin{aligned} z &= (x - \mu) / \sigma \\ &= (20 - 15) / 6 \\ &= .83 \end{aligned}$$

**Step 2: Find the area under the standard normal curve to the left of  $z = .83$ .**

# Standard Normal Probability Distribution



Cumulative Probability Table for  
the Standard Normal Distribution

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.	.	.	.	.	.	.	.	.	.	.
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
.	.	.	.	.	.	.	.	.	.	.

$$P(z \leq .83)$$

# Standard Normal Probability Distribution



## Solving for the Stock out Probability

Step 3: Compute the area under the standard normal curve to the right of  $z = .83$ .

$$\begin{aligned}P(z > .83) &= 1 - P(z \leq .83) \\&= 1 - .7967 \\&= .2033\end{aligned}$$

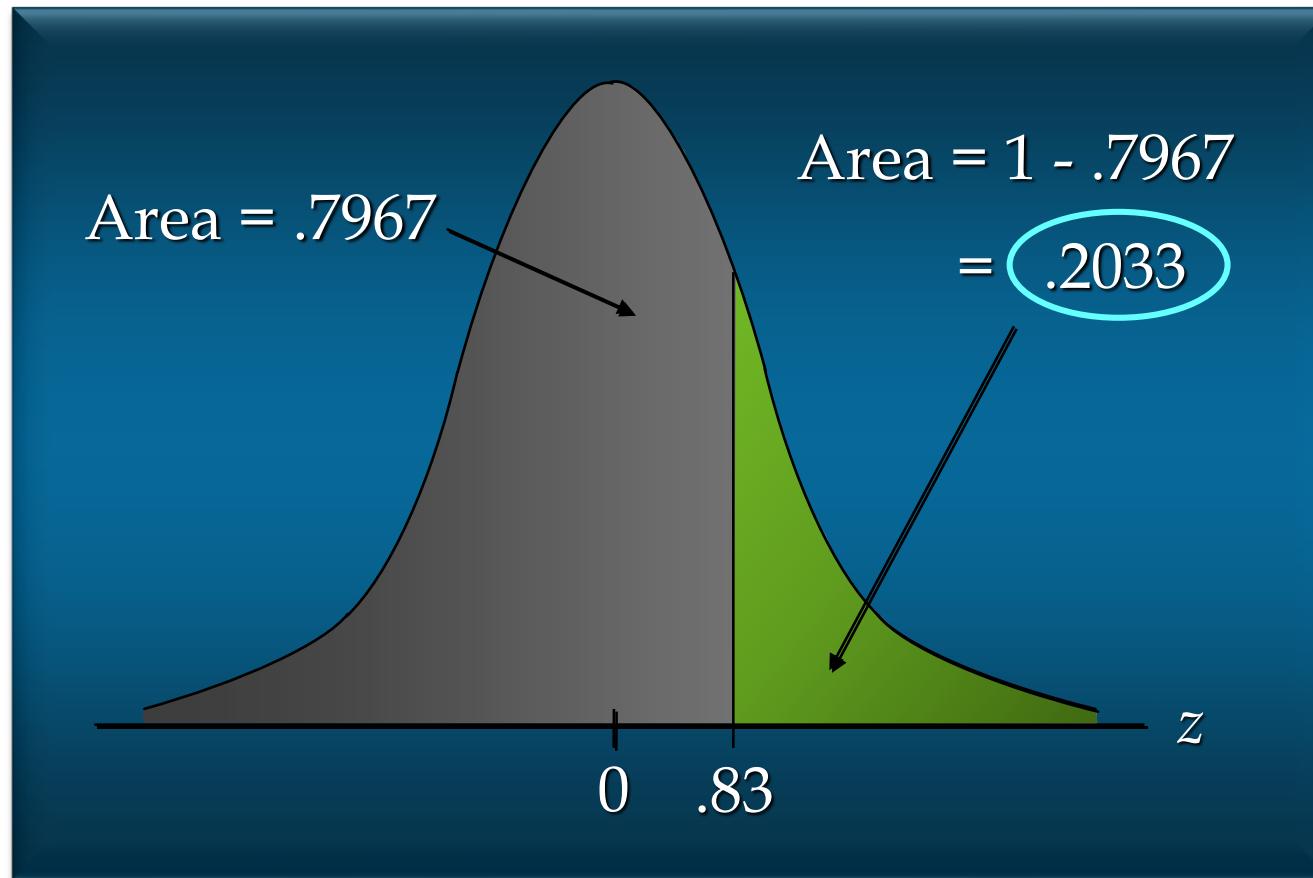
Probability  
of a stockout

$P(x > 20)$

# Standard Normal Probability Distribution



Solving for the Stock out Probability



# Standard Normal Probability Distribution



- Standard Normal Probability Distribution

If the manager of Pep Zone wants the probability of a stock out during replenishment lead-time to be no more than .05, what should the reorder point be?

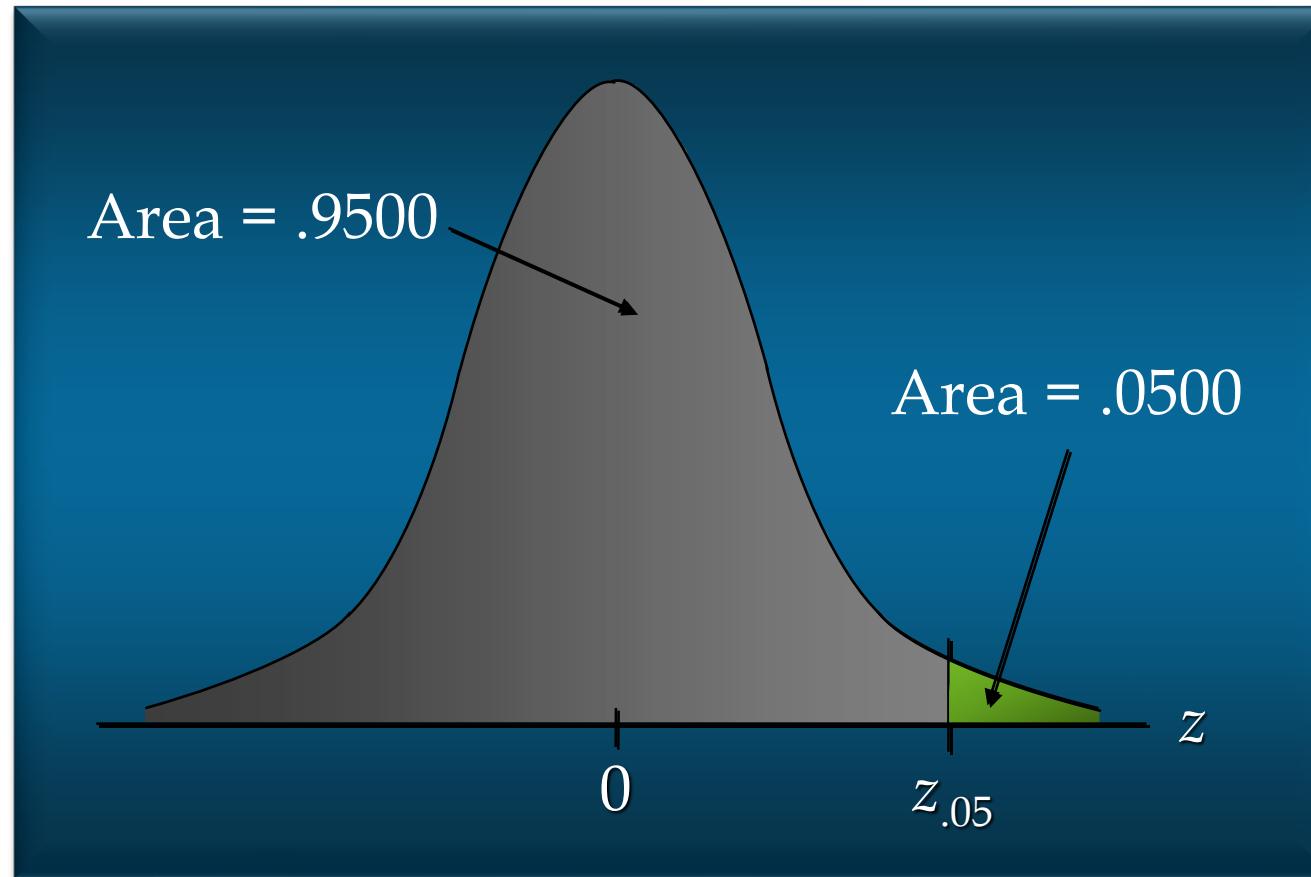
---

(Hint: Given a probability, we can use the standard normal table in an inverse fashion to find the corresponding z value.)

# Standard Normal Probability Distribution



- Solving for the Reorder Point



# Standard Normal Probability Distribution



## Solving for the Reorder Point

Step 1: Find the z-value that cuts off an area of .05 in the right tail of the standard normal distribution.

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.	.	.	.	.	.	.	.	.	.	.
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9685	.9692	.9699	.9705
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9757	.9763	.9767
.	.	.	.	.	.	.	.	.	.	.

We look up  
the complement  
of the tail area  
 $(1 - .05 = .95)$

# Standard Normal Probability Distribution



## Solving for the Reorder Point

Step 2: Convert  $z_{.05}$  to the corresponding value of  $x$ .

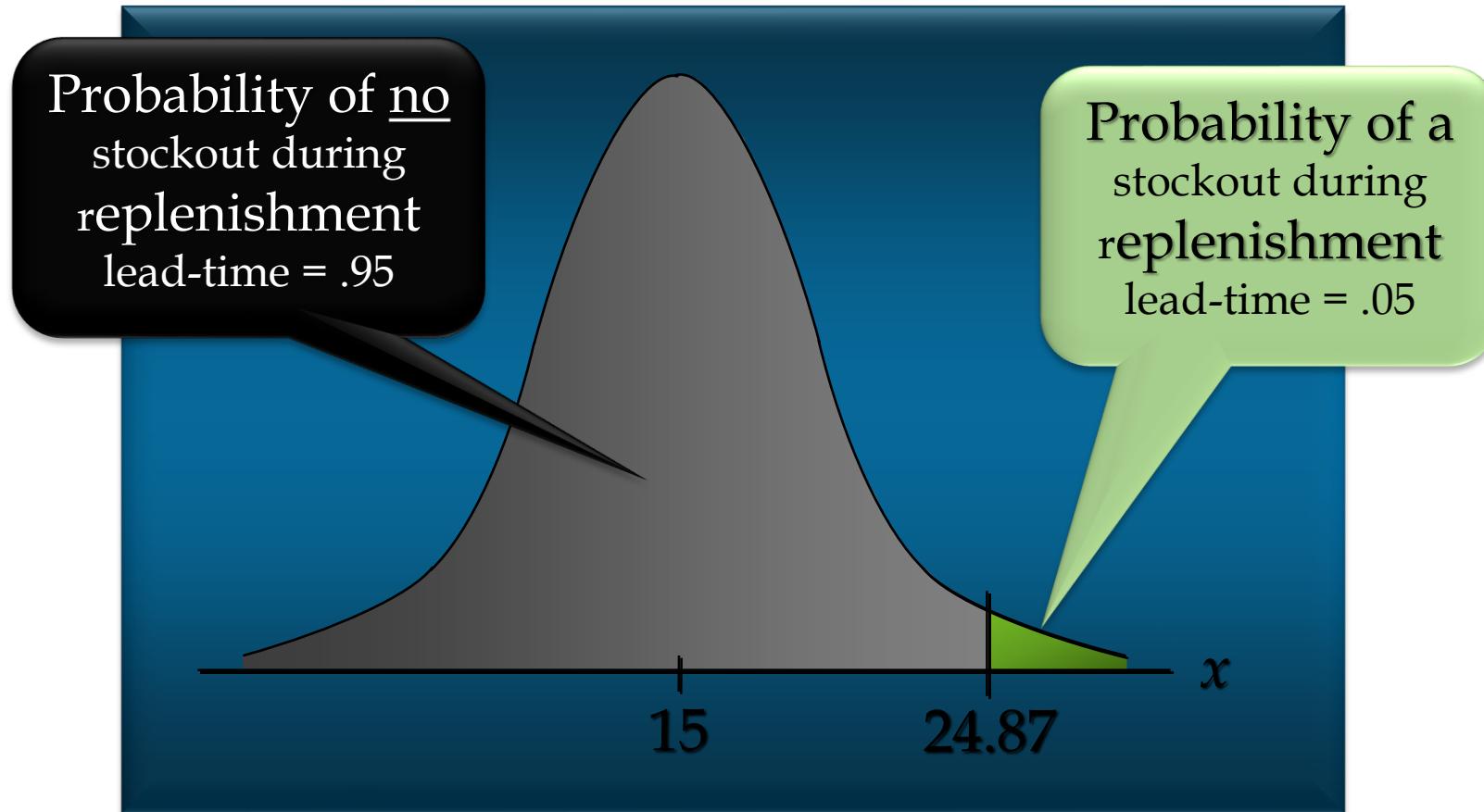
$$\begin{aligned}x &= \mu + z_{.05}\sigma \\&= 15 + 1.645(6) \\&= 24.87 \text{ or } 25\end{aligned}$$

A reorder point of 25 gallons will place the probability of a stockout during leadtime at (slightly less than) .05.

# Normal Probability Distribution



## □ Solving for the Reorder Point



# Standard Normal Probability Distribution



## Solving for the Reorder Point

By raising the reorder point from 20 gallons to 25 gallons on hand, the probability of a stockout decreases from about .20 to .05.

This is a significant decrease in the chance that Pep Zone will be out of stock and unable to meet a customer's desire to make a purchase.

# Exponential Probability Distribution



The exponential probability distribution is useful in describing the time it takes to complete a task.

The exponential random variables can be used to describe:

- Time between vehicle arrivals at a toll booth
- Time required to complete a questionnaire
- Distance between major defects in a highway

In waiting line applications, the exponential distribution is often used for service times.

# Exponential Probability Distribution



A property of the exponential distribution is that the mean and standard deviation are equal.

The exponential distribution is skewed to the right. Its skewness measure is 2.

# Exponential Probability Distribution



- Density Function

$$f(x) = \frac{1}{\mu} e^{-x/\mu} \quad \text{for } x \geq 0$$

where:     $\mu$  = expected or mean  
               $e = 2.71828$

# Exponential Probability Distribution



- Cumulative Probabilities

$$P(x \leq x_0) = 1 - e^{-x_0/\mu}$$

where:

$x_0$  = some specific value of  $x$

# Exponential Probability Distribution



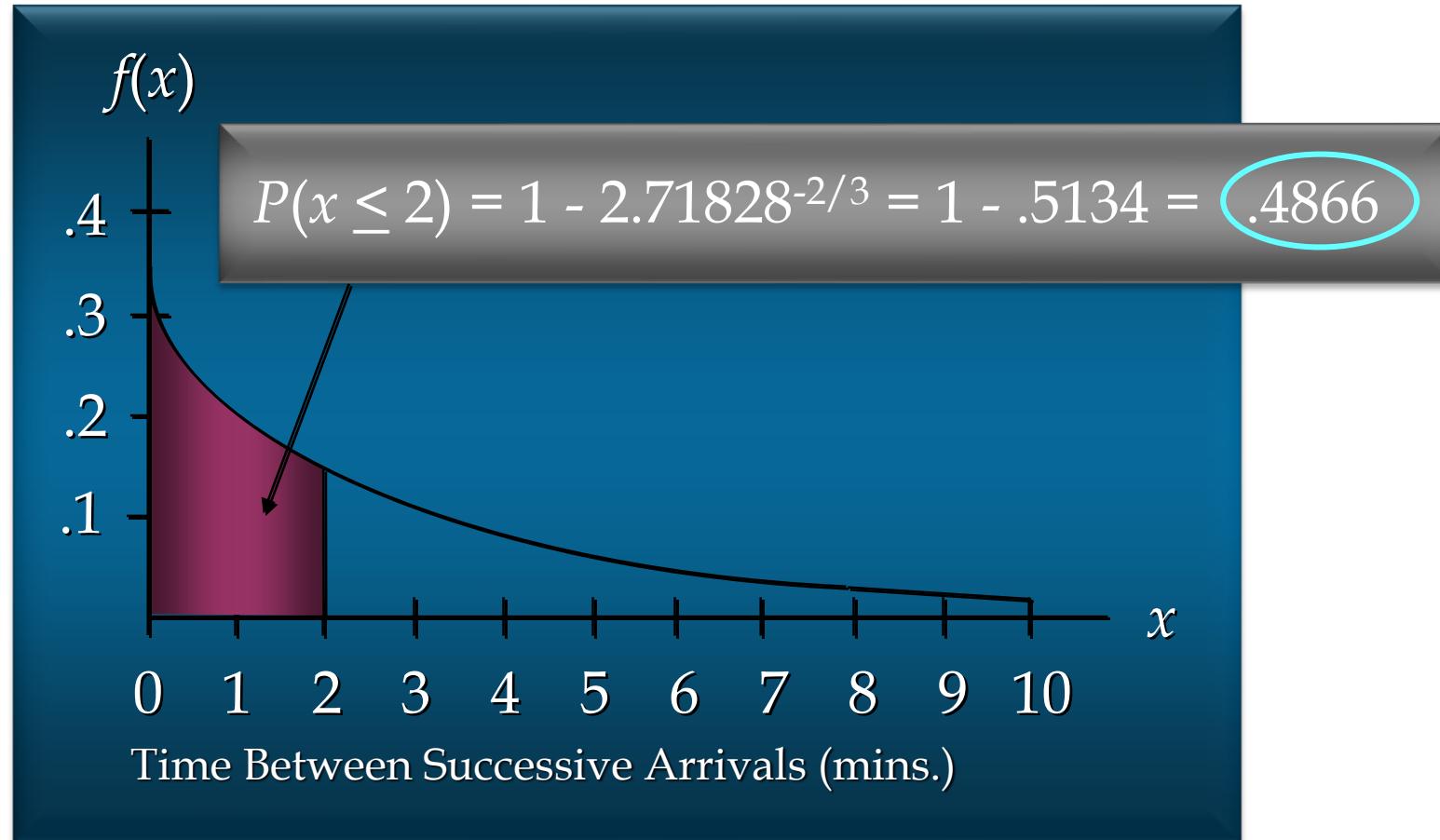
## Example: Al's Full-Service Pump

The time between arrivals of cars at Al's full-service gas pump follows an exponential probability distribution with a mean time between arrivals of 3 minutes. Al would like to know the probability that the time between two successive arrivals will be 2 minutes or less.

# Exponential Probability Distribution



Example: Al's Full-Service Pump



# Relationship between the Poisson and Exponential Distributions



The Poisson distribution provides an appropriate description of the number of occurrences per interval



The exponential distribution provides an appropriate description of the length of the interval between occurrences



# Central limit theorem

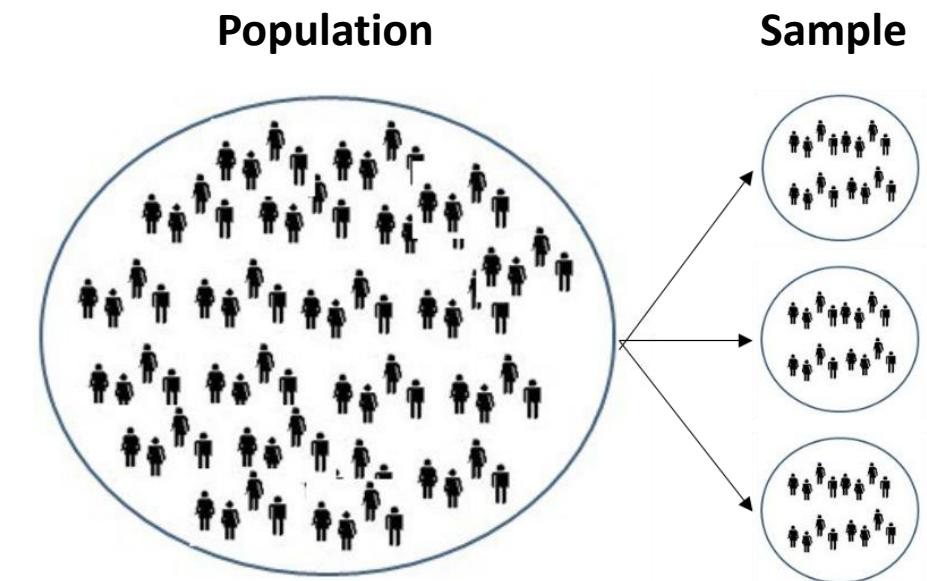
# What is Central limit theorem?



- Central limit theorem is the relationship between the population distribution and sampling distribution.

- Central limit theorem states:**
- Mean of sample is same as mean of the population.
- Standard deviation of the sample is equal to standard deviation of the population divided by square root of sample size.

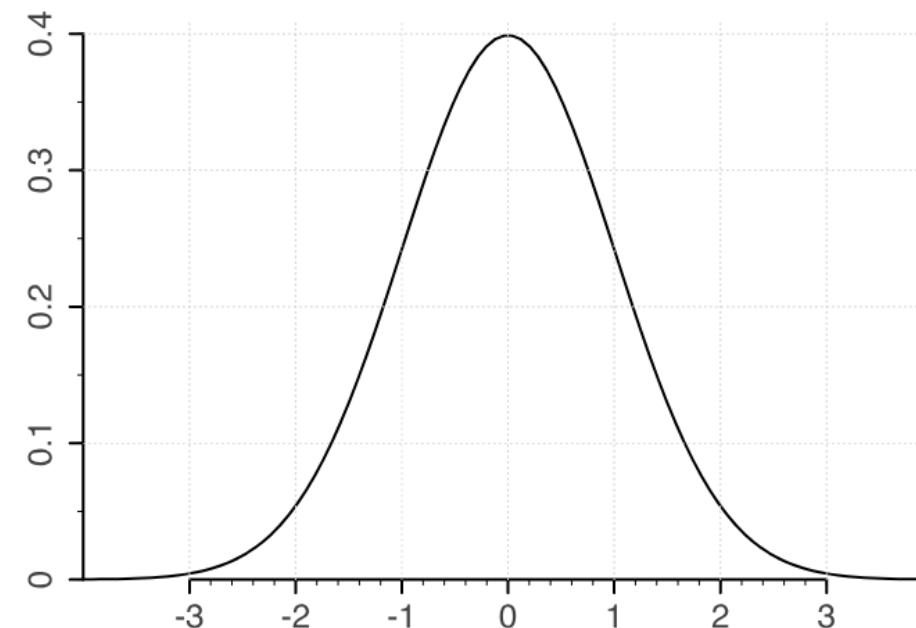
Sample Mean	$\mu_{\bar{x}} = \mu$	Population Mean
Sample Standard deviation	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$	Standard deviation / square root of sample size



# Why central limit theorem?



- Theorem states that if a population with mean  $\mu$  and standard deviation  $\sigma$  take sufficiently large random samples from the population then the distribution of the sample means should be approximately normally distributed.
- To check whether data is distributed normally or not (skewness).
- This fact holds especially true for sample sizes over 30.



# Central limit theorem



- **Problem Statement**
- At a coastal area, the number of crabs caught per day are recorded.
- The average of which is 10 and standard deviation is 3.
- If the record of 60 days is chosen randomly, estimate the mean and standard deviation of the chosen sample.

- Mean of the population  
 $\mu = 10$
- Standard deviation of the population  
 $\sigma = 3$
- Sample size  
 $n = 60$
- Mean of the sample is given by:  
 $\mu_{\bar{x}} = \mu$   
 $\mu_{\bar{x}} = 10$
- Standard deviation of the sample is given by:  
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = 3/\sqrt{60} = 0.387$$

$$\sigma_{\bar{x}} = 0.387$$

# Central limit theorem



- **Problem Statement**
- In a survey of a company, mean salary of employees is 29321 dollars with SD of 2120 dollars.
- Consider the sample of 100 employees and find the probability their mean salary will be less than 29000 dollars?

- Solution:
- Total number of employees (n) = 100
- Mean  $(\mu)$  = 29321
- standard deviation  $(\sigma)$  = 2120
- z-formula  $Z = (x - \mu) / (\sigma / \sqrt{n})$ 
$$z = (29,000 - 29,321) / (2,120 / \sqrt{100})$$
$$= -321 / 212$$
$$= -1.51$$
- Using z-table, we found -1.51 has an area of 93.45%.
- Since we have to find result for "less than", so minus 93.45 from 100 to get required result.
- $= 100 - 93.45 = 0.07$
- Hence the probability of employees having mean salary less than 29000 dollars is 0.07%.

# Z - Table



<i>z</i>	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995

# Central limit theorem



- **Problem Statement**
- The record of weights of male population follows normal distribution.
- Its mean and standard deviation are 70 kg and 15 kg respectively.
- If a researcher considers the records of 50 males, then what would be the mean and standard deviation of the chosen sample?

Solution:

Mean of the population  $\mu = 70 \text{ kg}$

Standard deviation of the population = 15 kg  
sample size  $n = 50$

Mean of the sample is given by:

$$\begin{aligned}\mu_{\bar{x}} &= \mu \\ \mu_{\bar{x}} &= 70 \text{ kg}\end{aligned}$$

Standard deviation of the sample is given by:

$$\begin{aligned}\sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} \\ \sigma_{\bar{x}} &= \frac{15}{\sqrt{50}} \\ \sigma_{\bar{x}} &= 2.121 = 2.1 \text{ kg (approx)}\end{aligned}$$



# Confidence Interval

# Confidence Interval Formula

- Assumptions
  - Population standard deviation  $\sigma$  is known
  - Population is normally distributed
  - If sample is not normal, use large sample
- Confidence interval estimate:

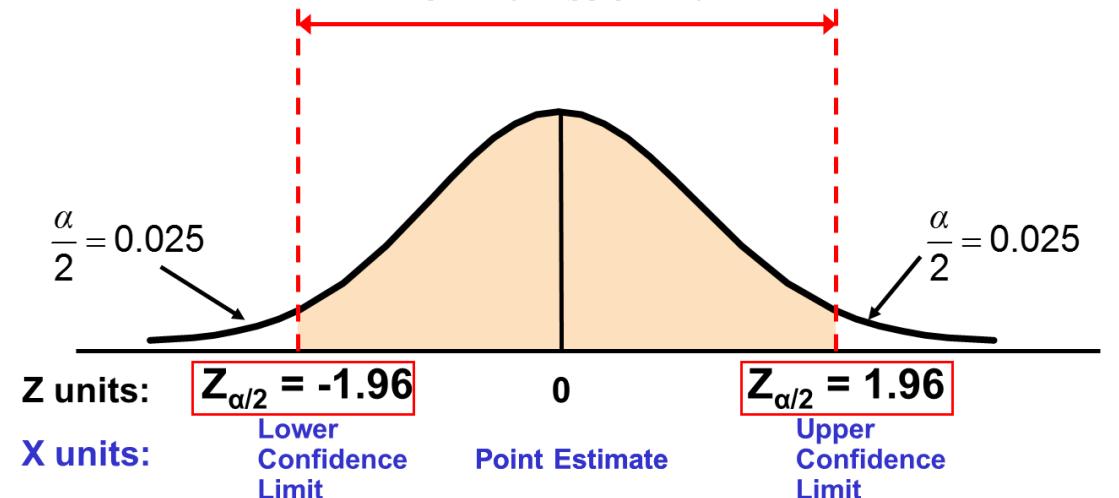
$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

where  $\bar{X}$  is the point estimate

$Z_{\alpha/2}$  is the normal distribution critical value for a probability of  $\alpha/2$  in each tail  
 $\sigma/\sqrt{n}$  is the standard error

- Consider a 95% confidence interval:

$$1 - \alpha = 0.95 \text{ so } \alpha = 0.05$$



$$Z_{\alpha/2} = \pm 1.96$$

Confidence Interval	$z$
80%	1.282
85%	1.440
90%	1.645
95%	1.960
99%	2.576
99.5%	2.807
99.9%	3.291

# Example



Suppose that in a sample of 50 college students in Illinois, the mean credit card debt was \$346. Suppose that we also have reason to believe (from previous studies) that the population standard deviation of credit card debts for this group is \$108. Use this information to calculate a 95% confidence interval for the mean credit card debt of all college students in Illinois.

## Solution

From reading the problem, we also have:

Mean is \$346:  $\bar{x} = 346$

Population standard deviation is 108:  $s=108$

Applying the formula:

$$\bar{x} \pm z_c \left( \frac{\sigma}{\sqrt{n}} \right)$$
$$346 \pm 1.96 \left( \frac{108}{\sqrt{50}} \right)$$

Left hand endpoint:  $346 - 1.96(108/\sqrt{50}) = 316.1$

Right hand endpoint:  $346 + 1.96(108/\sqrt{50}) = 375.9$

This gives our 95% confidence interval for  $\mu$ , the population mean, as (316.1, 375.9).

## Interpretation

We are 95% confident that the mean amount of credit card debt for all college students in Illinois is between \$316.10 and \$375.90.

# Example

## Example: Average Height

We measure the heights of **40** randomly chosen men, and get a mean height of **175cm**,

We also know the standard deviation of men's heights is **20cm**.



The **95% Confidence Interval** (we show how to calculate it later) is:

$$175\text{cm} \pm 6.2\text{cm}$$



This says the **true mean** of ALL men (if we could measure all their heights) is likely to be between 168.8cm and 181.2cm.



# Statistical Tests

# What is a Hypothesis?



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

- population mean

**Example: The mean monthly cell phone bill in this city is  $\mu = \$42$**

- population proportion

**Example: The proportion of adults in this city with cell phones is  $\pi = 0.68$**

# The Null Hypothesis, H<sub>0</sub>



- States the claim or assertion to be tested

**Example:** The average number of TV sets in U.S. Homes is equal to three

$$( H_0 : \mu = 3 )$$

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

$$H_0 : \mu = 3$$

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

# The Null Hypothesis, H<sub>0</sub>



- Begin with the assumption that the null hypothesis is true
  - Similar to the notion of innocent until proven guilty
- Refers to the status quo or historical value
- Always contains “=” , “≤” or “≥” sign
- May or may not be rejected

# The Alternative Hypothesis, H<sub>1</sub>

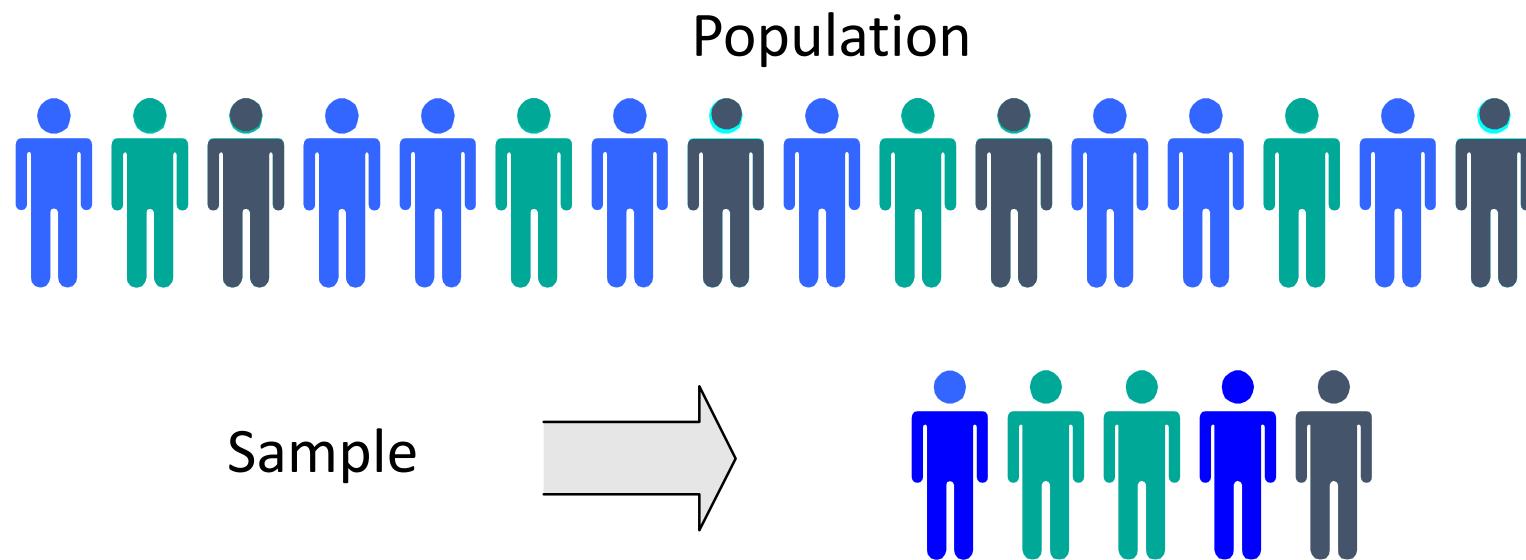


- Is the opposite of the null hypothesis
  - e.g., The average number of TV sets in U.S. homes is not equal to 3 (  $H_1: \mu \neq 3$  )
- Challenges the status quo
- Never contains the “=” , “≤” or “≥” sign
- May or may not be proven
- Is generally the hypothesis that the researcher is trying to prove

# The Hypothesis Testing Process



- Claim: The population mean age is 50.
  - $H_0: \mu = 50$ ,       $H_1: \mu \neq 50$
- Sample the population and find sample mean.



# The Hypothesis Testing Process



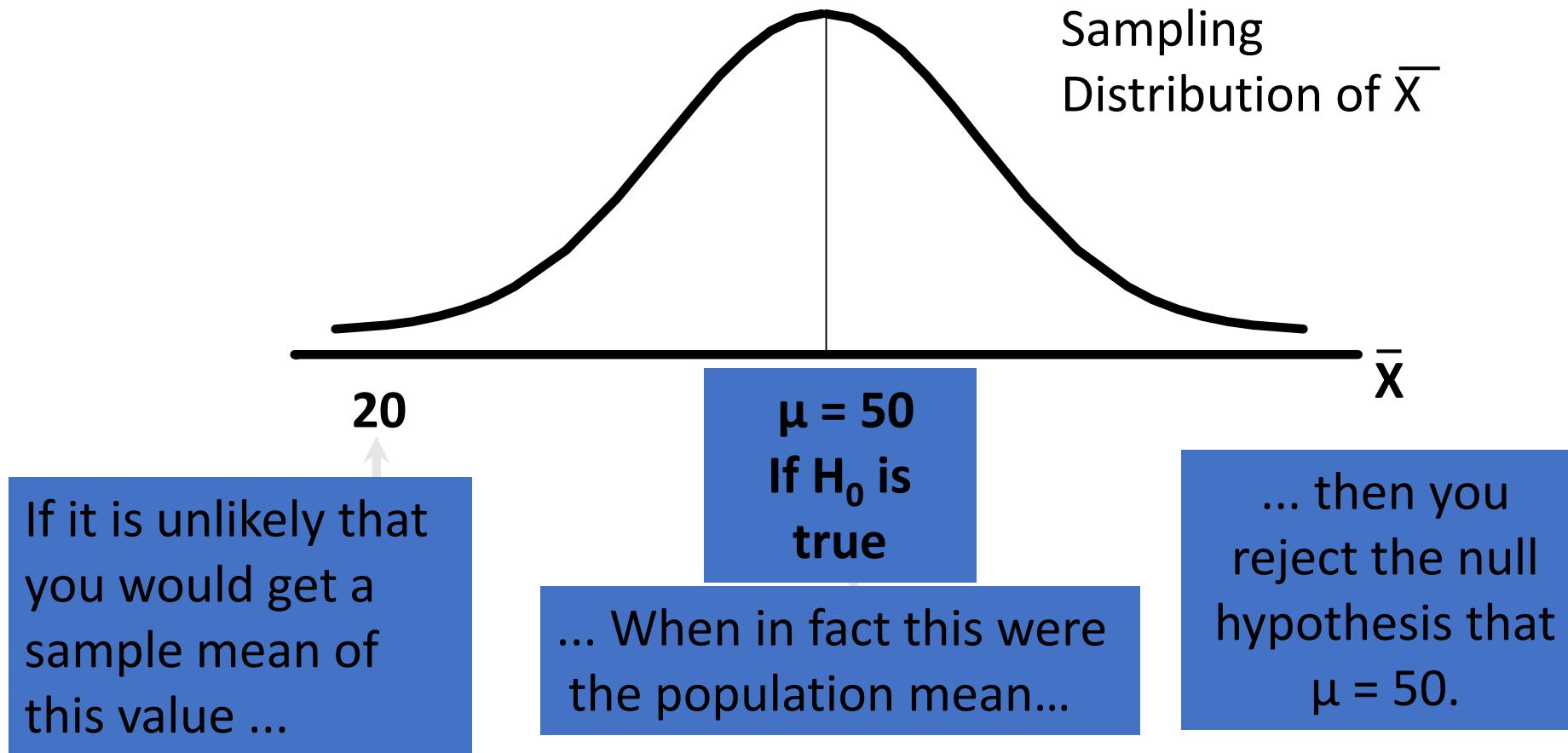
*(continued)*

- Suppose the sample mean age was  $\bar{X} = 20$ .
- This is significantly lower than the claimed mean population age of 50.
- If the null hypothesis were true, the probability of getting such a different sample mean would be very small, so you reject the null hypothesis .
- In other words, getting a sample mean of 20 is so unlikely if the population mean was 50, you conclude that the population mean must not be 50.

# The Hypothesis Testing Process



(continued)



# The Test Statistic and Critical Values

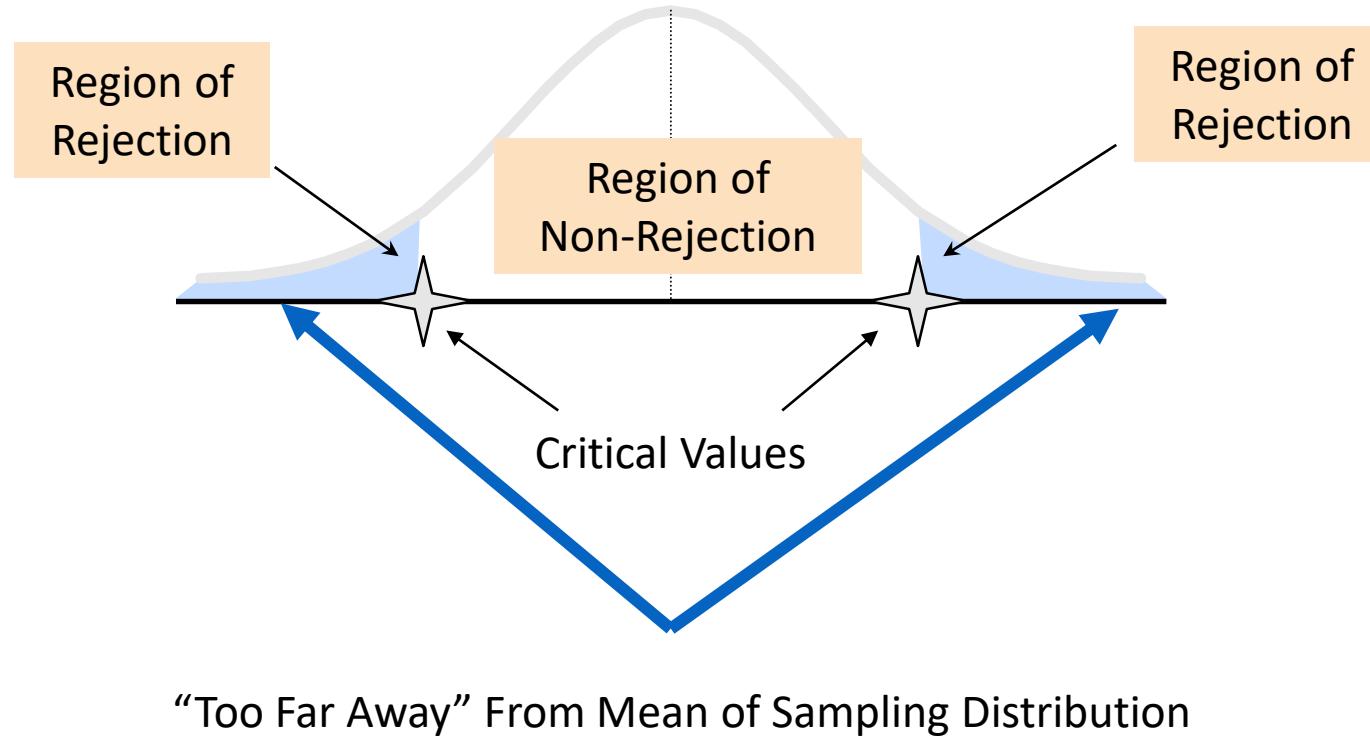


- If the sample mean is close to the assumed population mean, the null hypothesis is not rejected.
- If the sample mean is far from the assumed population mean, the null hypothesis is rejected.
- How far is “far enough” to reject  $H_0$ ?
- The critical value of a test statistic creates a “line in the sand” for decision making -- it answers the question of how far is far enough.

# The Test Statistic and Critical Values



Sampling Distribution of the test statistic



# Possible Errors in Hypothesis Test Decision Making



- **Type I Error**
  - Reject a true null hypothesis
  - Considered a serious type of error
  - The probability of a Type I Error is  $\alpha$ 
    - Called level of significance of the test
    - Set by researcher in advance
- **Type II Error**
  - Failure to reject false null hypothesis
  - The probability of a Type II Error is  $\beta$

# Possible Errors in Hypothesis Test Decision Making



(continued)

Possible Hypothesis Test Outcomes		
	Actual Situation	
Decision	$H_0$ True	$H_0$ False
Do Not Reject $H_0$	No Error Probability $1 - \alpha$	Type II Error Probability $\beta$
Reject $H_0$	Type I Error Probability $\alpha$	No Error Probability $1 - \beta$

# Possible Errors in Hypothesis Test Decision Making



*(continued)*

- The **confidence coefficient** ( $1-\alpha$ ) is the probability of not rejecting  $H_0$  when it is true.
- The **confidence level** of a hypothesis test is  $(1-\alpha)*100\%$ .
- The **power of a statistical test** ( $1-\beta$ ) is the probability of rejecting  $H_0$  when it is false.

# Type I & II Error Relationship



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

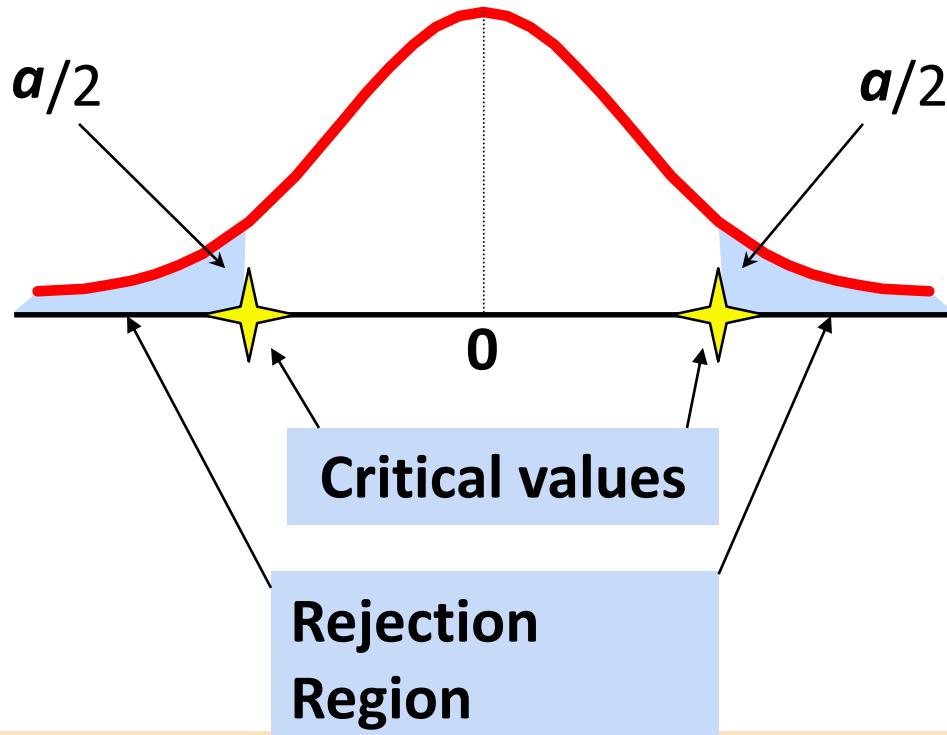
If Type I error probability (  $\alpha$  ) ↑ , then  
Type II error probability (  $\beta$  ) ↓

# Factors Level of Significance and the Rejection Region



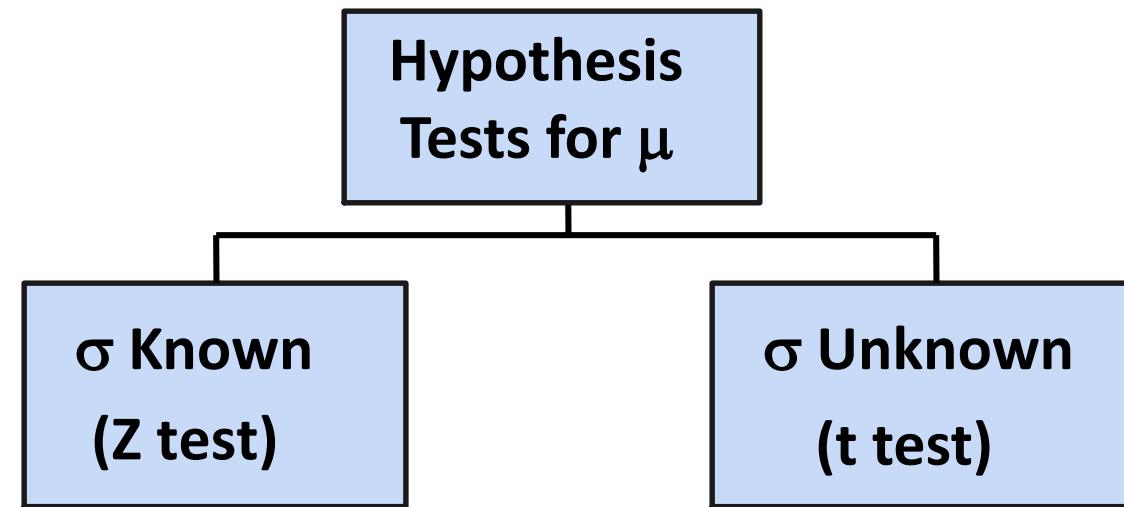
$H_0: \mu = 3$   
 $H_1: \mu \neq 3$

Level of significance =  $\alpha$



This is a **two-tail test** because there is a rejection region in both tails

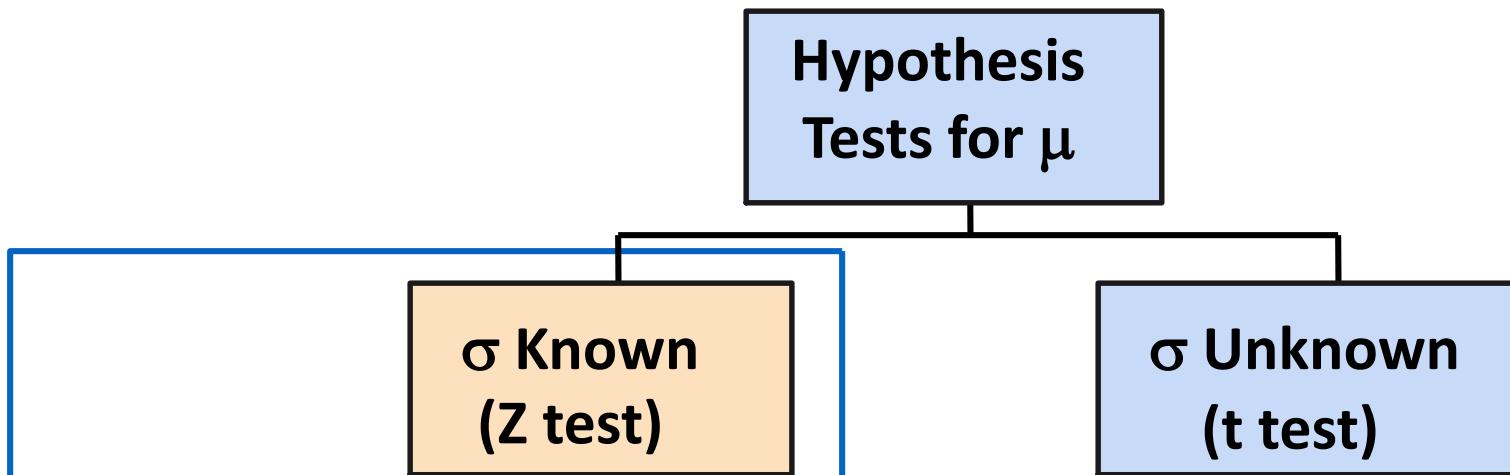
# Hypothesis Tests for the Mean



# Z Test of Hypothesis for the Mean ( $\sigma$ Known)



- Convert sample statistic ( $\bar{X}$ ) to a  $Z_{\text{STAT}}$  test statistic



The test statistic is:

$$Z_{\text{STAT}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

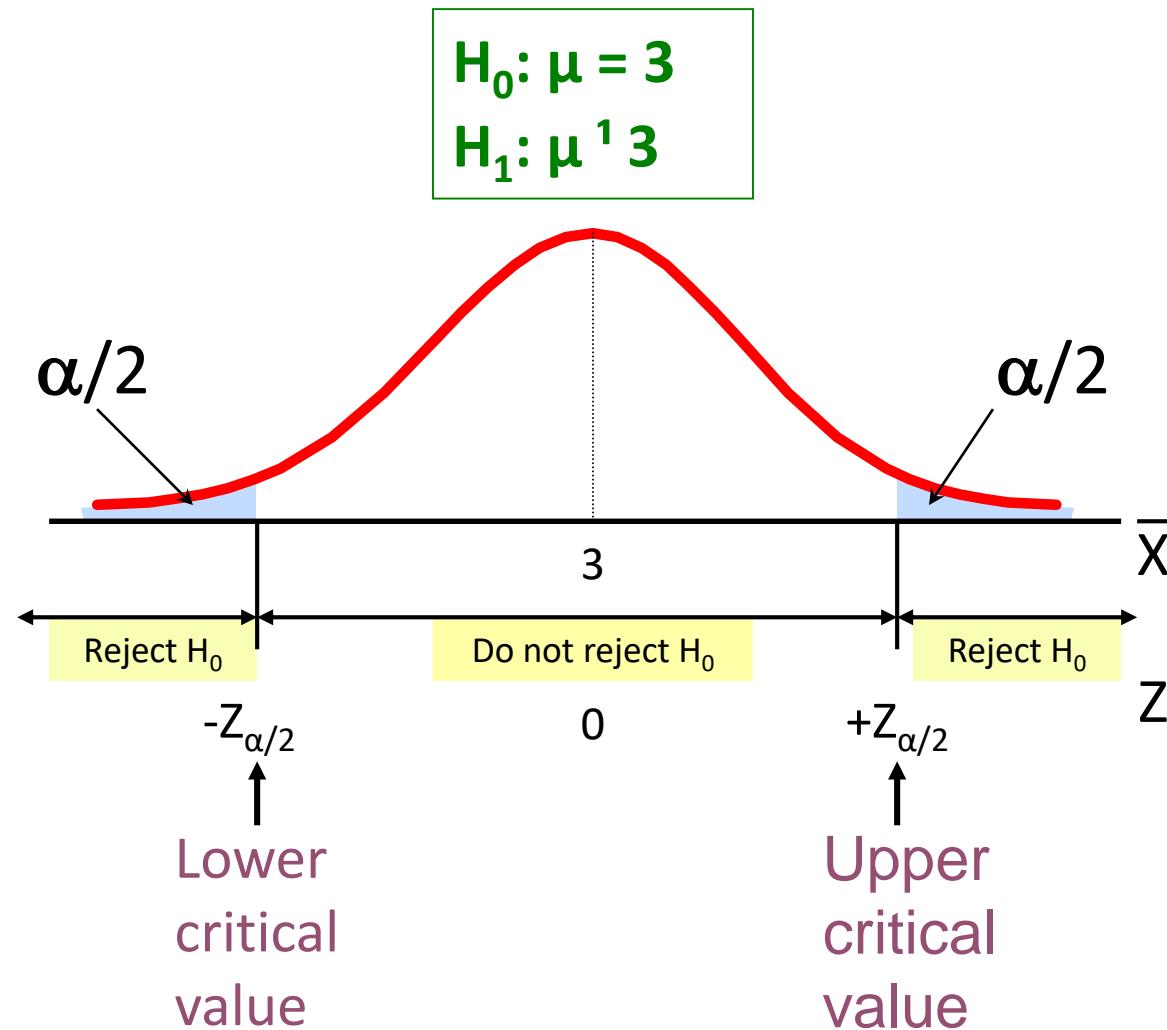
# Z Critical Value Approach to Testing



- For a two-tail test for the mean,  $\sigma$  known:
- Convert sample statistic ( $\bar{X}$ ) to test statistic ( $Z_{STAT}$ )
- Determine the critical Z values for a specified level of significance  $\alpha$  from a table or computer
- **Decision Rule:** If the test statistic falls in the rejection region, reject  $H_0$  ; otherwise do not reject  $H_0$

# Two-Tail Tests

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



# 6 Steps in Hypothesis Testing



1. State the null hypothesis,  $H_0$  and the alternative hypothesis,  $H_1$
2. Choose the level of significance,  $\alpha$ , and the sample size,  $n$
3. Determine the appropriate test statistic and sampling distribution
4. Determine the critical values that divide the rejection and non rejection regions
5. Collect data and compute the value of the test statistic
6. Make the statistical decision and state the managerial conclusion. If the test statistic falls into the non rejection region, do not reject the null hypothesis  $H_0$ . If the test statistic falls into the rejection region, reject the null hypothesis. Express the managerial conclusion in the context of the problem

# Hypothesis Testing Example



**Test the claim that the true mean # of TV sets in US homes is equal to 3.**

**(Assume  $\sigma = 0.8$ )**

1. State the appropriate null and alternative hypotheses
  - $H_0: \mu = 3$     $H_1: \mu \neq 3$  (This is a two-tail test)
2. Specify the desired level of significance and the sample size
  - Suppose that  $\alpha = 0.05$  and  $n = 100$  are chosen for this test

# Hypothesis Testing Example

(continued)



3. Determine the appropriate technique
  - $\sigma$  is assumed known so this is a Z test.
4. Determine the critical values
  - For  $\alpha = 0.05$  the critical Z values are  $\pm 1.96$
5. Collect the data and compute the test statistic
  - Suppose the sample results are  
 $n = 100, \bar{X} = 2.84$  ( $\sigma = 0.8$  is assumed known)

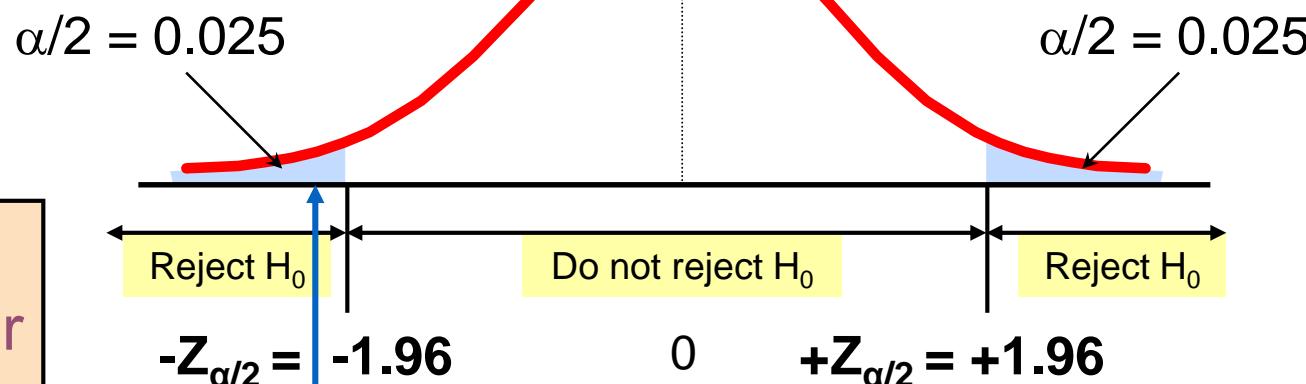
So the test statistic is:

$$Z_{\text{STAT}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{2.84 - 3}{\frac{0.8}{\sqrt{100}}} = \frac{-0.16}{0.08} = -2.0$$

# Hypothesis Testing Example

(continued)

- 6. Is the test statistic in the rejection region?



Reject H<sub>0</sub> if  
 $Z_{STAT} < -1.96$  or  
 $Z_{STAT} > 1.96$ ;  
otherwise do  
not reject H<sub>0</sub>

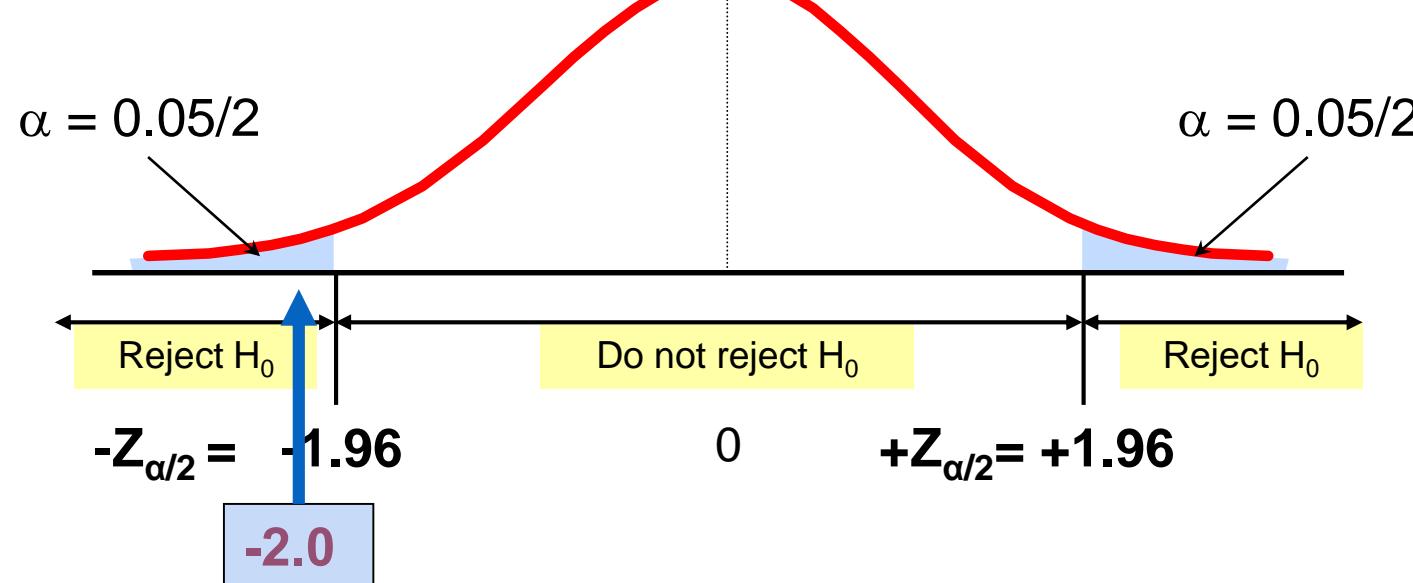
Here,  $Z_{STAT} = -2.0 < -1.96$ , so the  
test statistic is in the rejection  
region

# Hypothesis Testing Example



(continued)

6 (continued). Reach a decision and interpret the result



Since  $Z_{\text{STAT}} = -2.0 < -1.96$ , reject the null hypothesis and conclude there is sufficient evidence that the mean number of TVs in US homes is not equal to 3

# p-Value Approach to Testing



- p-value: Probability of obtaining a test statistic equal to or more extreme than the observed sample value **given  $H_0$  is true**
  - The p-value is also called the observed level of significance
  - It is the smallest value of  $\alpha$  for which  $H_0$  can be rejected

# p-Value Approach to Testing : Interpreting the p-value



- Compare the **p-value** with  $\alpha$

- If  $p\text{-value} < \alpha$  , reject  $H_0$
- If  $p\text{-value} \geq \alpha$  , do not reject  $H_0$

# The 5 Step p-value approach to Hypothesis Testing



1. State the null hypothesis,  $H_0$  and the alternative hypothesis,  $H_1$
2. Choose the level of significance,  $\alpha$ , and the sample size,  $n$
3. Determine the appropriate test statistic and sampling distribution
4. Collect data and compute the value of the test statistic and the p-value
5. Make the statistical decision and state the managerial conclusion. If the p-value is  $< \alpha$  then reject  $H_0$ , otherwise do not reject  $H_0$ . State the managerial conclusion in the context of the problem

# p-value Hypothesis Testing Example



**Test the claim that the true mean # of TV sets in US homes is equal to 3.**

**(Assume  $\sigma = 0.8$ )**

1. State the appropriate null and alternative hypotheses
  - $H_0: \mu = 3$     $H_1: \mu \neq 3$  (This is a two-tail test)
2. Specify the desired level of significance and the sample size
  - Suppose that  $\alpha = 0.05$  and  $n = 100$  are chosen for this test

# p-value Hypothesis Testing Example



(continued)

3. Determine the appropriate technique
  - $\sigma$  is assumed known so this is a Z test.
4. Collect the data, compute the test statistic and the p-value
  - Suppose the sample results are  
 $n = 100, \bar{X} = 2.84$  ( $\sigma = 0.8$  is assumed known)

So the test statistic is:

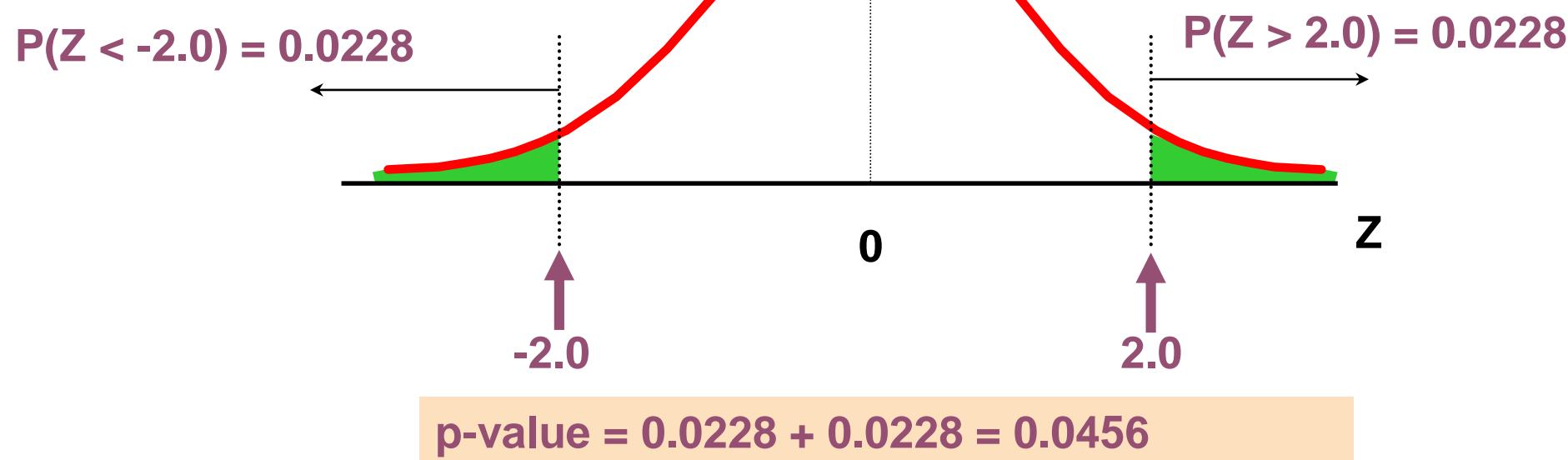
$$Z_{\text{STAT}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{2.84 - 3}{\frac{0.8}{\sqrt{100}}} = \frac{-0.16}{0.08} = -2.0$$

# p-value Hypothesis Testing Example: Calculating the p-value



4. (continued) Calculate the p-value.

- How likely is it to get a  $Z_{\text{STAT}}$  of -2 (or something further from the mean (0), in either direction) if  $H_0$  is true?



# p-value Hypothesis Testing Example



*(continued)*

- 5. Is the p-value  $< \alpha$ ?
  - Since  $p\text{-value} = 0.0456 < \alpha = 0.05$  Reject  $H_0$
- 5. (continued) State the managerial conclusion in the context of the situation.
  - There is sufficient evidence to conclude the average number of TVs in US homes is not equal to 3.

# Connection Between Two Tail Tests and Confidence Intervals



- For  $\bar{X} = 2.84$ ,  $\sigma = 0.8$  and  $n = 100$ , the 95% confidence interval is:

$$2.84 - (1.96) \frac{0.8}{\sqrt{100}} \text{ to } 2.84 + (1.96) \frac{0.8}{\sqrt{100}}$$

$$2.6832 \leq \mu \leq 2.9968$$

- Since this interval does not contain the hypothesized mean (3.0), we reject the null hypothesis at  $\alpha = 0.05$

# Do You Ever Truly Know $\sigma$ ?



- Probably not!
- In virtually all real world business situations,  $\sigma$  is not known.
- If there is a situation where  $\sigma$  is known then  $\mu$  is also known (since to calculate  $\sigma$  you need to know  $\mu$ .)
- If you truly know  $\mu$  there would be no need to gather a sample to estimate it.

# Hypothesis Testing: $\sigma$ Unknown

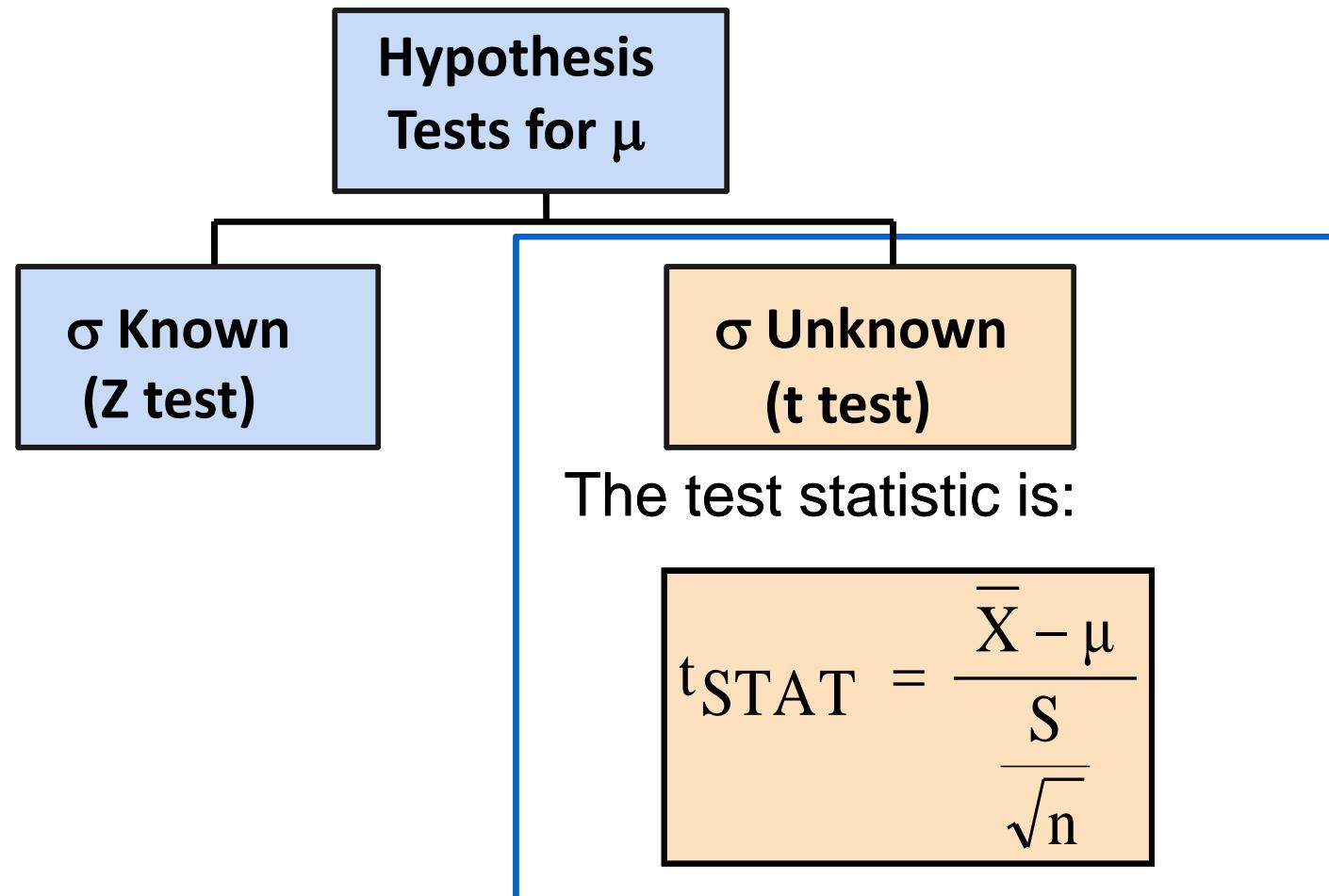


- If the population standard deviation is unknown, you instead use the sample standard deviation S.
- Because of this change, you use the t distribution instead of the Z distribution to test the null hypothesis about the mean.
- When using the t distribution you must assume the population you are sampling from follows a normal distribution.
- All other steps, concepts, and conclusions are the same.

# t Test of Hypothesis for the Mean ( $\sigma$ Unknown)



- Convert sample statistic ( $\bar{X}$ ) to a  $t_{STAT}$  test statistic



# Example: Two-Tail Test( $\sigma$ Unknown)

The average cost of a hotel room in New York is said to be \$168 per night. To determine if this is true, a random sample of 25 hotels is taken and resulted in an  $X$  of \$172.50 and an  $S$  of \$15.40. Test the appropriate hypotheses at  $\alpha = 0.05$ .

(Assume the population distribution is normal)

$$\begin{aligned}H_0: \mu &= 168 \\H_1: \mu &\neq 168\end{aligned}$$

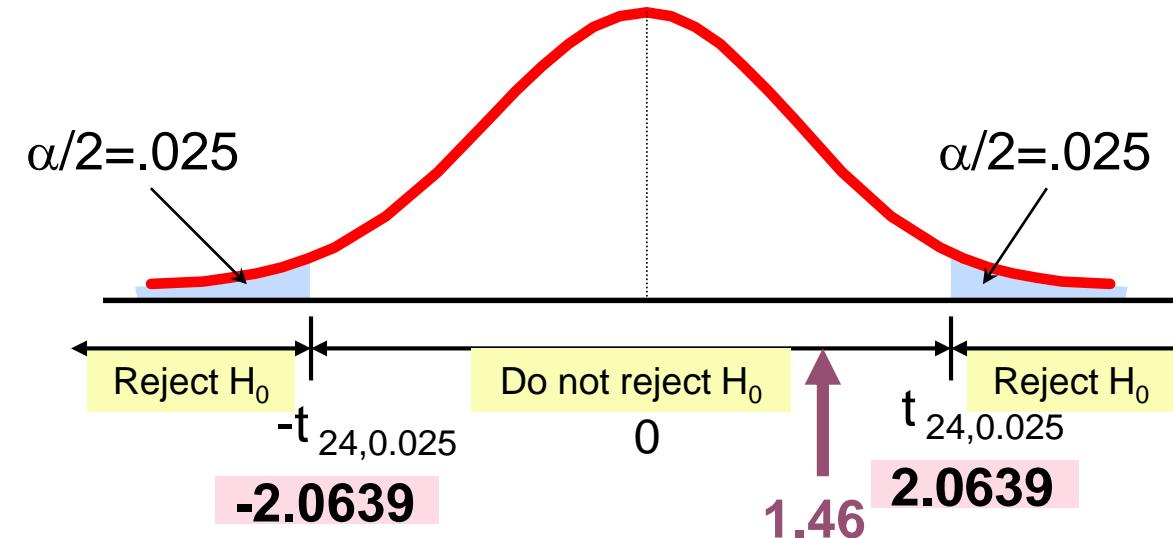
# Example: Example Solution: Two-Tail t Test



$$\begin{aligned} H_0: \mu &= 168 \\ H_1: \mu &\neq 168 \end{aligned}$$

- $\alpha = 0.05$
- $n = 25, df = 25-1=24$
- $\sigma$  is unknown, so use a **t statistic**
- Critical Value:

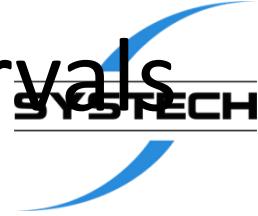
$$\pm t_{24,0.025} = \pm 2.0639$$



$$\rightarrow t_{\text{STAT}} = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{172.50 - 168}{\frac{15.40}{\sqrt{25}}} = 1.46$$

**Do not reject H<sub>0</sub>:** insufficient evidence that true mean cost is different than \$168

# Connection of Two Tail Tests to Confidence Intervals



- For  $\bar{X} = 172.5$ ,  $S = 15.40$  and  $n = 25$ , the 95% confidence interval for  $\mu$  is:

$$172.5 - (2.0639) \frac{15.4}{\sqrt{25}} \text{ to } 172.5 + (2.0639) \frac{15.4}{\sqrt{25}}$$

$$166.14 \leq \mu \leq 178.86$$

- Since this interval contains the Hypothesized mean (168), we do not reject the null hypothesis at  $\alpha = 0.05$

# One-Tail Tests



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

$$H_0: \mu \geq 3$$

$$H_1: \mu < 3$$



This is a **lower**-tail test since the alternative hypothesis is focused on the lower tail below the mean of 3

$$H_0: \mu \leq 3$$

$$H_1: \mu > 3$$



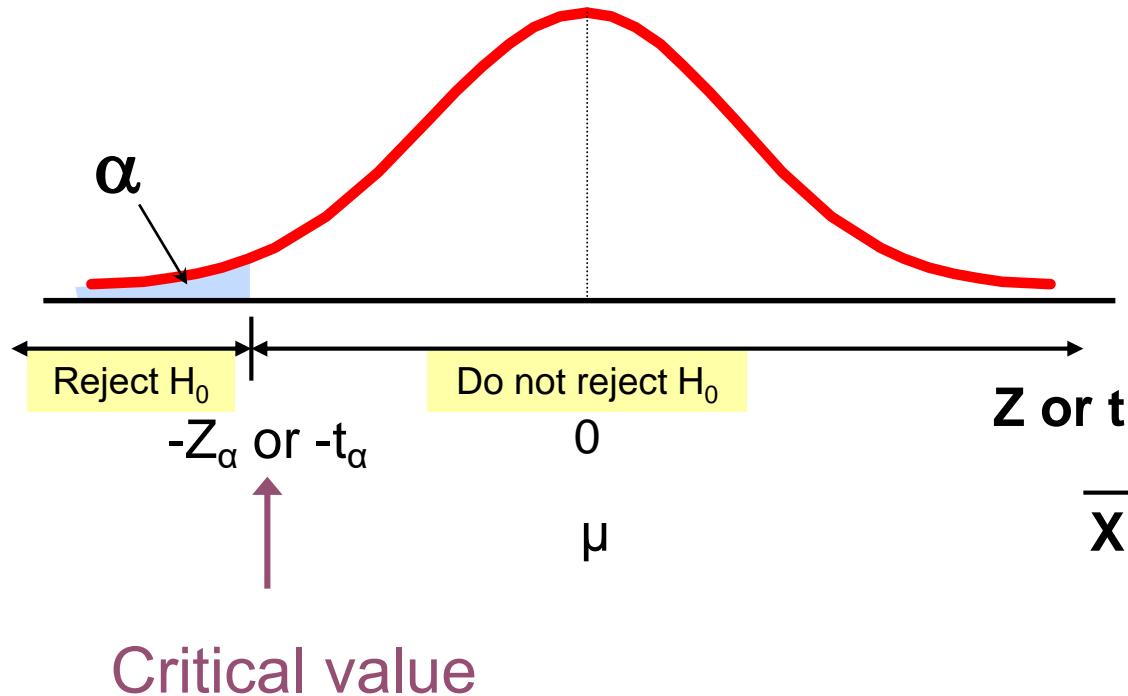
This is an **upper**-tail test since the alternative hypothesis is focused on the upper tail above the mean of 3

# Lower-Tail Tests



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

$$\begin{aligned} H_0: \mu &\geq 3 \\ H_1: \mu &< 3 \end{aligned}$$

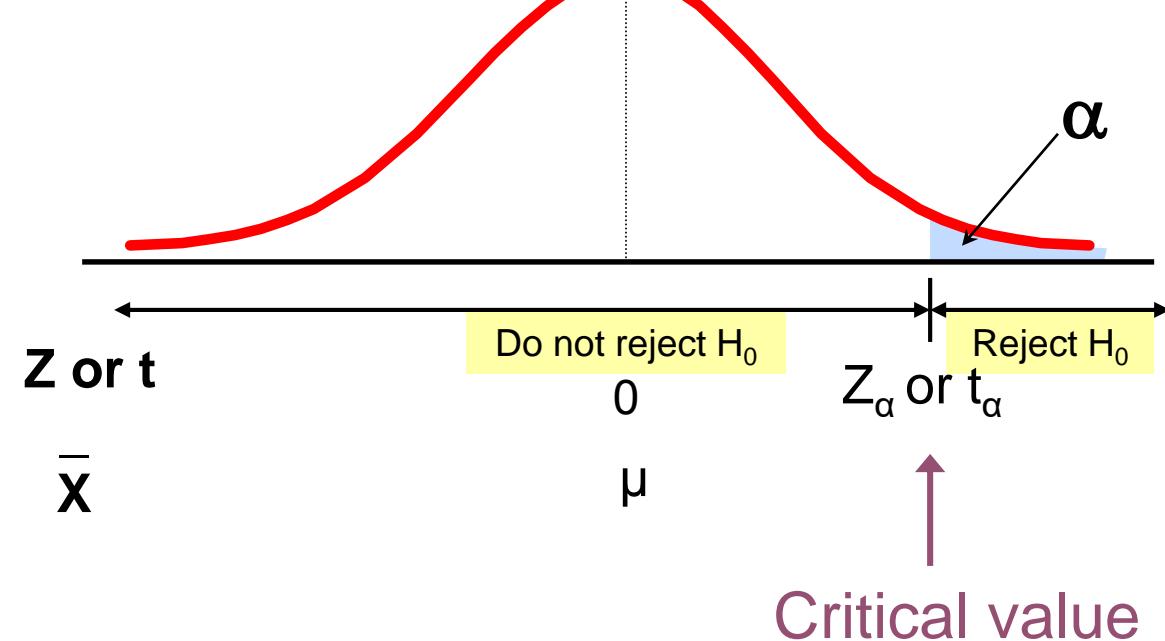


# Upper-Tail Tests



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

$$\begin{aligned} H_0: \mu &\leq 3 \\ H_1: \mu &> 3 \end{aligned}$$



# Example: Upper-Tail t Test for Mean ( $\sigma$ unknown)



A phone industry manager thinks that customer monthly cell phone bills have increased, and now average over \$52 per month. The company wishes to test this claim. (Assume a normal population)

Form hypothesis test:

$H_0: \mu \leq 52$  the average is not over \$52 per month

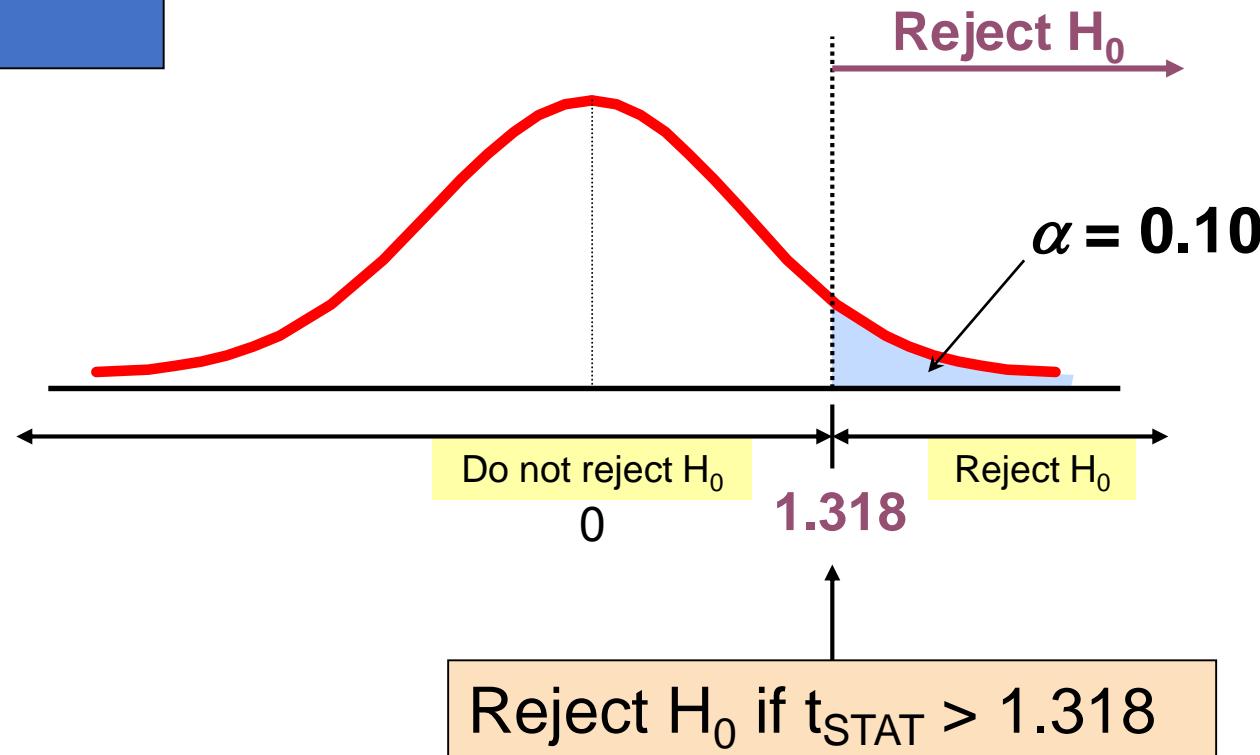
$H_1: \mu > 52$  the average **is** greater than \$52 per month  
(i.e., sufficient evidence exists to support the manager's claim)

# Example: Find Rejection Region

(continued)

- Suppose that  $\alpha = 0.10$  is chosen for this test and  $n = 25$

Find the rejection region:



# Example: Test Statistic



(continued)

Obtain sample and compute the test statistic

Suppose a sample is taken with the following results:  $n = \overline{25}$ ,  $X = 53.1$ , and  $S = 10$

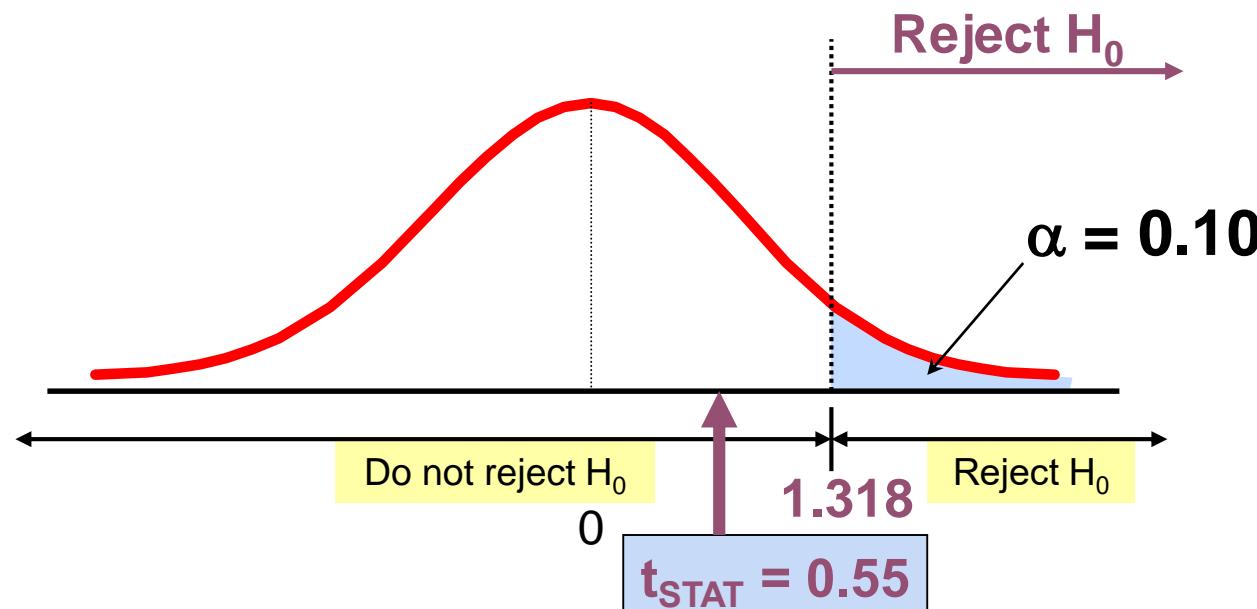
- Then the test statistic is:

$$t_{\text{STAT}} = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{53.1 - 52}{\frac{10}{\sqrt{25}}} = 0.55$$

# Example: Decision

(continued)

Reach a decision and interpret the result:



**Do not reject  $H_0$  since  $t_{STAT} = 0.55 \leq 1.318$**

there is not sufficient evidence that the mean bill is over \$52

# Chi Square Test

# Example – Goodness of fit

The American Pet Products Association conducted a survey in 2011 and determined that

- 60% of dog owners have only one dog,
- 28% have two dogs, and
- 12% have three or more

Supposing that you have decided to conduct your

own survey and have collected the data below, determine whether your data supports the results of the APPA study.

Use a significance level of 0.05.

Data: Out of 129 dog owners, 73 had one dog and 38 had two dogs

# Solution

**Step 1:** Clearly state the null and alternative hypotheses.

$H_0$ : The proportion of dog owners with one, two or three dogs is 0.60, 0.28 and 0.12 respectively.

$H_1$ : The proportion of dog owners with one, two or three dogs does not match the proposed model.

**Step 2:** Identify an appropriate test and significance level. In the absence of a stated significance level in the problem, we assume the default 0.05.

# Solution

**Step 3:** Analyze sample data. Create a table to organize data and compare the observed data to the expected data

	1 dog	2 dogs	3+ dogs	Total
Observed	73	38	18	129
Expected				

To identify the expected values, multiply the expected % by the total number observed

	1 dog	2 dogs	3+ dogs	Total
Observed	73	38	18	129
Expected	.60*129=77.4	0.28*129=36.1	0.12*129=15.5	129

# Solution

To calculate our chi-square statistic, we need to sum the squared difference between each observed and expected value divided by the expected value:

$$\chi^2 = \sum (\text{observed} - \text{expected})^2 / \text{expected}$$

$$\chi^2 = (73 - 77.4)^2 / 77.4 + (38 - 36.1)^2 / 36.1 + (18 - 15.5)^2 / 15.5$$

$$\chi^2 = (-4.4)^2 / 77.4 + (1.9)^2 / 36.1 + (2.5)^2 / 15.5$$

$$\chi^2 = 19.36 / 77.4 + 3.61 / 36.1 + 6.25 / 15.5$$

$$\chi^2 = 0.2501 + 0.1000 + 0.4032$$

$$\chi^2 = 0.7533$$

# Solution

Chi-square Distribution Table

d.f.	.995	.99	.975	.95	.9	.1	.05	.025	.01
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58

# Solution

Using the table, we find that the critical value for a 0.05 significance level with df = 2 is 5.99.

If our chi-square value is greater than 5.995 or the null hypothesis is incorrect. Our chi-square statistic is only 0.7533 so we will not reject the null hypothesis. we can say that our survey data does support the distribution that is in American Pet Products Association

# Example- Test of Independence

Rachel claims that girls take more black and white and color photographs than boys, but Jack (who is a photographer) is skeptical.

If Jack collects the following data, would it be correct to say that he should reject Rachel's claim that gender affects tendency to take photographs?

	Black/White	Color	Total
Female	72	489	561
Male	48	530	578
Total	120	1019	1139

# Solution

1. The upper-left cell, female X black/white:  
expected cell value =  $C \times R/n$

$= (\text{column total}) \times (\text{row total}) /$   
total number of observations

$$= 120 \times 561 / 1139$$

expected cell value = 59.1

2. The cell below that, male X black/white:  
expected cell value =  $C \times R/n$

$= (\text{column total}) \times (\text{row total}) /$   
total number of observations

$$= 120 \times 578 / 1139$$

expected cell value = 60.9

# Solution

3. Top-right cell, female X color:  
expected cell value =  $C \times R/n$

$= (\text{column total}) \times (\text{row total}) /$   
total number of observations

$$= 1019 \times 561 / 1139$$

expected cell value = 501.9

4. Bottom-right cell, male X color:  
expected cell value =  $C \times R/n$

$= (\text{column total}) \times (\text{row total}) /$   
total number of observations

$$= 1019 \times 578 / 1139$$

expected cell value = 517.1

# Solution

	Black/White	Color	Total
Female	72(59.1)	489(501.9)	561
Male	48(60.9)	530(517.1)	578
Total	120	1019	1139

$$\chi^2 = \sum (\text{observed} - \text{expected})^2 / \text{expected}$$

$$\chi^2 = (72-59.1)^2 / 59.10 + (48-60.1)^2 / 60.89 + (489-501.9)^2 / 501.89 + (530-517.1)^2 / 517.10$$

$$\chi^2 = (12.9)^2 / 59.10 + (-12.9)^2 / 60.89 + (-12.9)^2 / 501.89 + (12.9)^2 / 517.10$$

$$\chi^2 = 166.41 / 59.10 + 166.41 / 60.89 + 166.41 / 501.89 + 166.41 / 517.10$$

$$\chi^2 = 2.82 + 2.73 + .33 + .32 \quad \chi^2 = 6.20$$

# Solution

$$d.f = (\text{rows}-1)(\text{columns}-1) \quad d.f = (2-1)(2-1) \quad d.f = 1$$

we find that the critical value for 0.05 with  $d.f = 1$  is 3.8414.

we compare our calculated chi-squared value of 6.2 to the critical value of 3.8414 and determine that since  $6.2 > 3.8414$

we can reject  $H_0$ , in other words, we reject the independence of the variables .

The observed data indicates that there is a gender bias on picture-taking tendency.



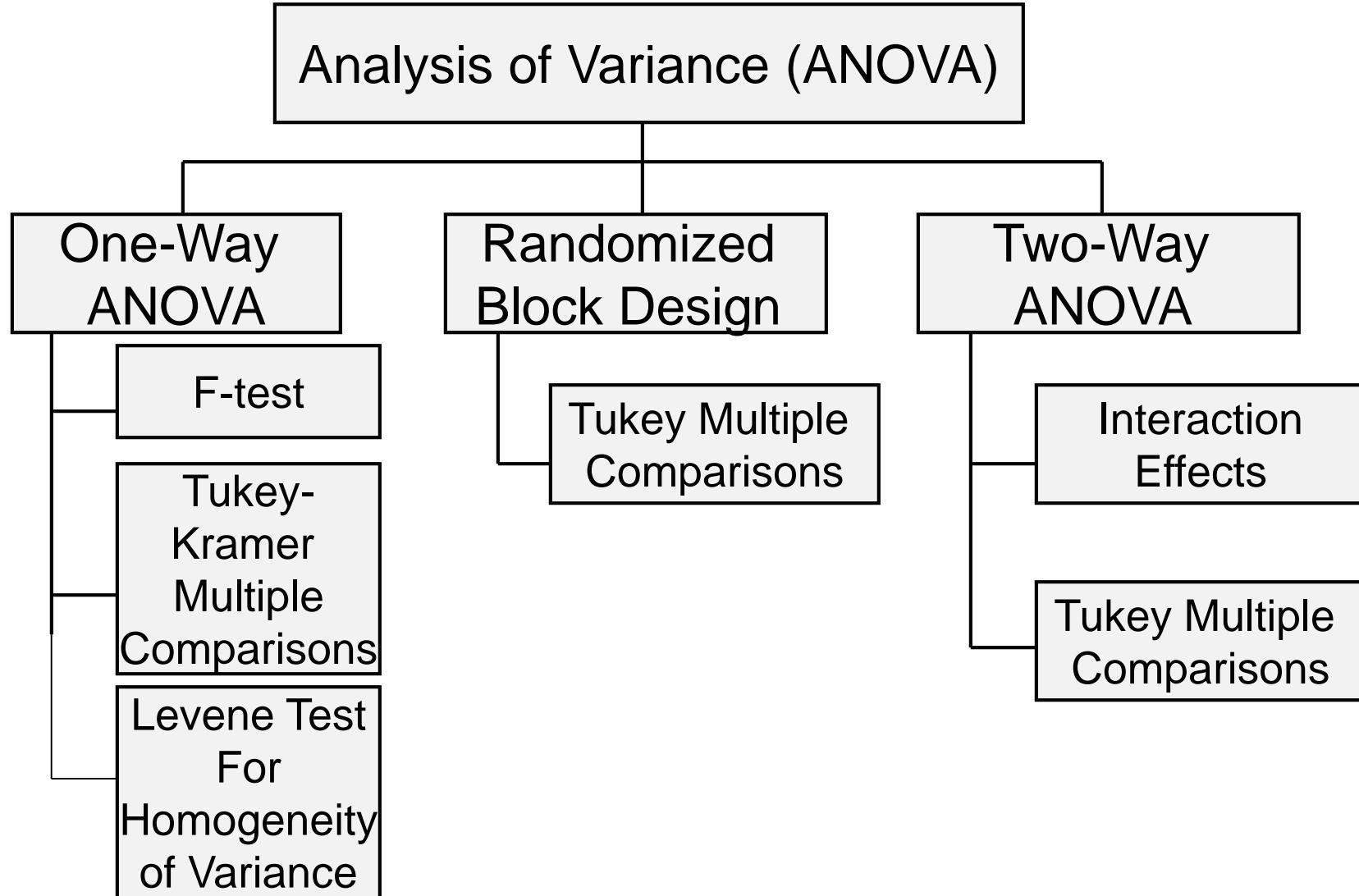
# ANOVA - Analysis of Variance

# Introduction



- An ANOVA test enables us to decide whether to reject or accept the hypothesis.
- Basically, we will be testing samples from different population.
- Examples of when you might want to test different groups:
  - A group of psychiatric patients are trying three different therapies: counseling, medication and biofeedback. You want to see if one therapy is better than the others.
  - A manufacturer has two different processes to make light bulbs. They want to know if one process is better than the other.

# Overview



# One-Way Analysis of Variance



- Evaluate the difference among means of three or more groups

## Examples:

- Accident rates for 1<sup>st</sup>, 2<sup>nd</sup>, and 3<sup>rd</sup> shift.
- Expected mileage for five brands of tires

## • Assumptions

- Populations are normally distributed
- Populations have equal variances
- Samples are randomly and independently drawn

# Hypotheses of One-Way ANOVA



- $H_0 : \mu_1 = \mu_2 = \mu_3 = \dots = \mu_c$ 
  - All population means are equal
  - i.e., no factor effect (no variation in means among groups)
- $H_1 : \text{Not all of the population means are the same}$ 
  - At least one population mean is different
  - i.e., there is a factor effect
  - Does not mean that all population means are different  
(some pairs may be the same)

# One-Way ANOVA

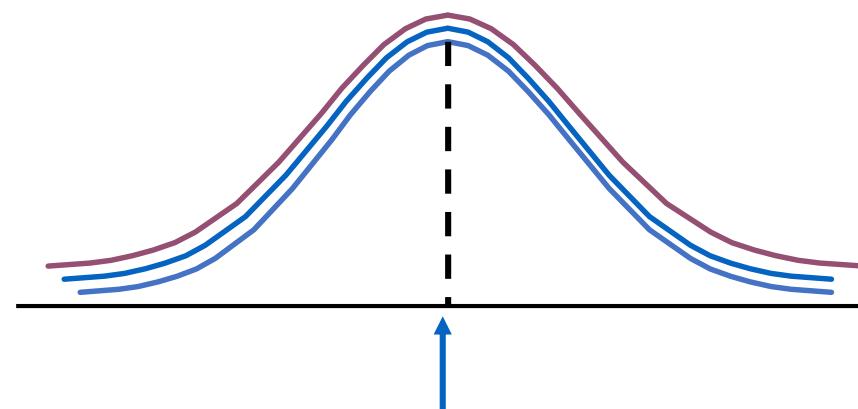


$$H_0 : \mu_1 = \mu_2 = \mu_3 = \dots = \mu_c$$

$H_1$  : Not all  $\mu_j$  are the same

The Null Hypothesis is True

All Means are same:  
(No Factor Effect)



$$\mu_1 = \mu_2 = \mu_3$$

# One-Way ANOVA

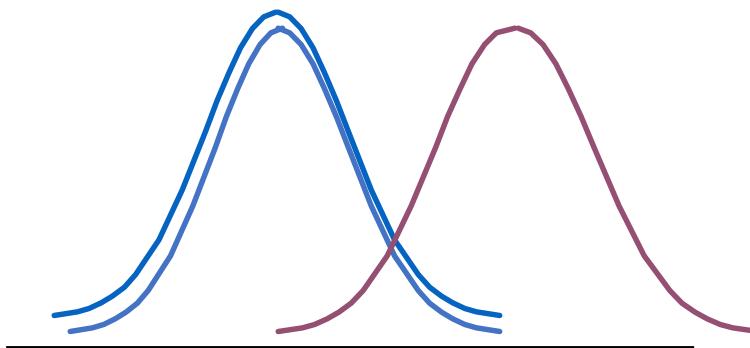


$$H_0 : \mu_1 = \mu_2 = \mu_3 = \dots = \mu_c$$

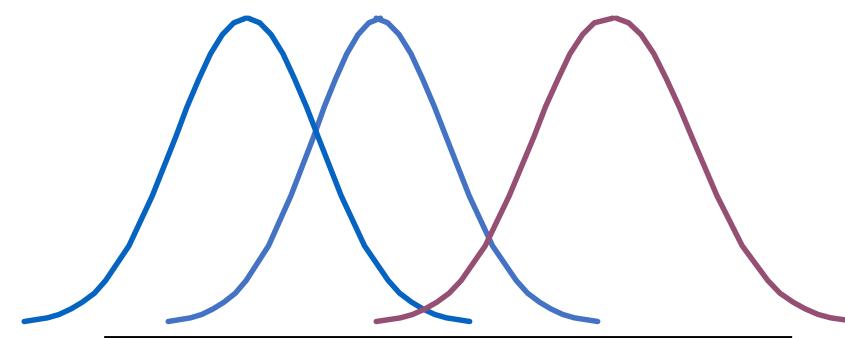
$H_1$  : Not all  $\mu_j$  are the same

The Null Hypothesis is NOT true

At least one of the means is different  
(Factor Effect is present)



or



$$\mu_1 = \mu_2 \neq \mu_3$$

$$\mu_1 \neq \mu_2 \neq \mu_3$$

# Partitioning the Variation



- Total variation can be split into two parts:

$$\text{SST} = \text{SSA} + \text{SSW}$$

SST = Total Sum of Squares  
*(Total variation)*

SSA = Sum of Squares Among Groups  
*(Among-group variation)*

SSW = Sum of Squares Within  
Groups  
*(Within-group variation)*

# Partitioning the Variation



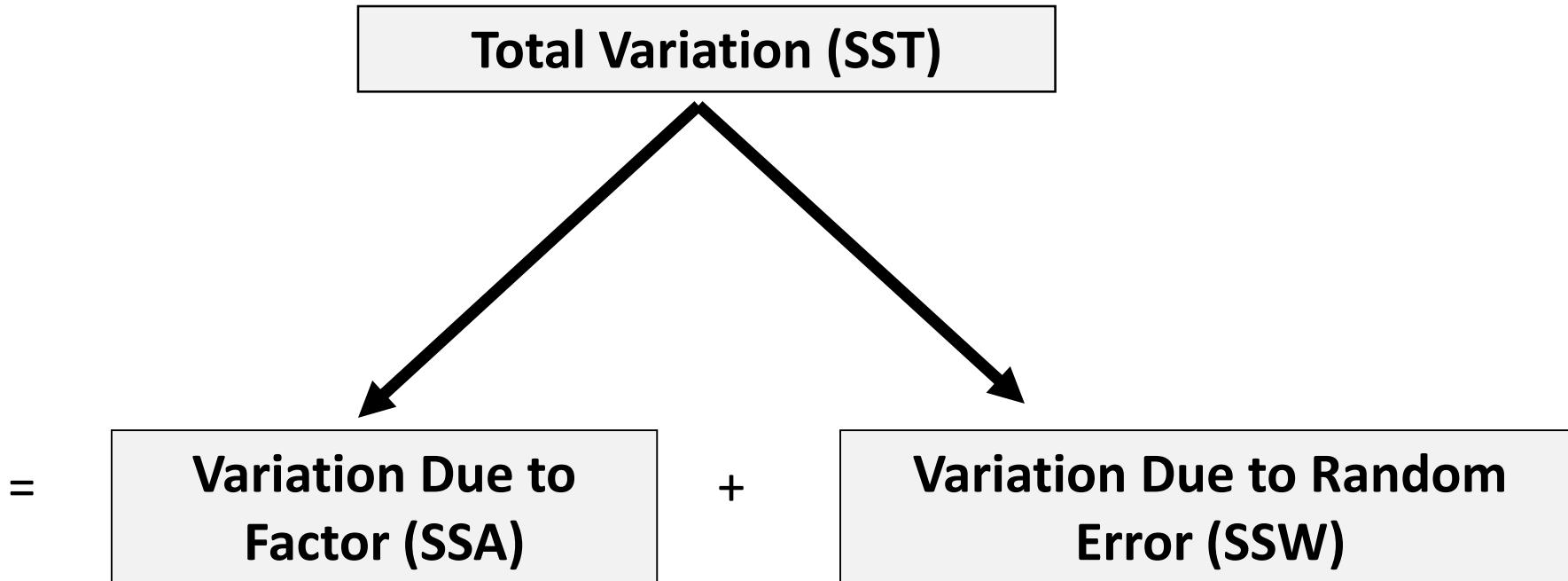
$$SST = SSA + SSW$$

**Total Variation** = the aggregate variation of the individual data values across the various factor levels (SST)

**Among-Group Variation** = variation among the factor sample means (SSA)

**Within-Group Variation** = variation that exists among the data values within a particular factor level (SSW)

# Partition of Total Variation



# Total Sum of Squares

$$\boxed{\text{SST} = \text{SSA} + \text{SSW}}$$

$$\text{SST} = \sum_{j=1}^c \sum_{i=1}^{n_j} (X_{ij} - \bar{X})^2$$

Where:

SST = Total sum of squares

c = number of groups or levels

$n_j$  = number of observations in group j

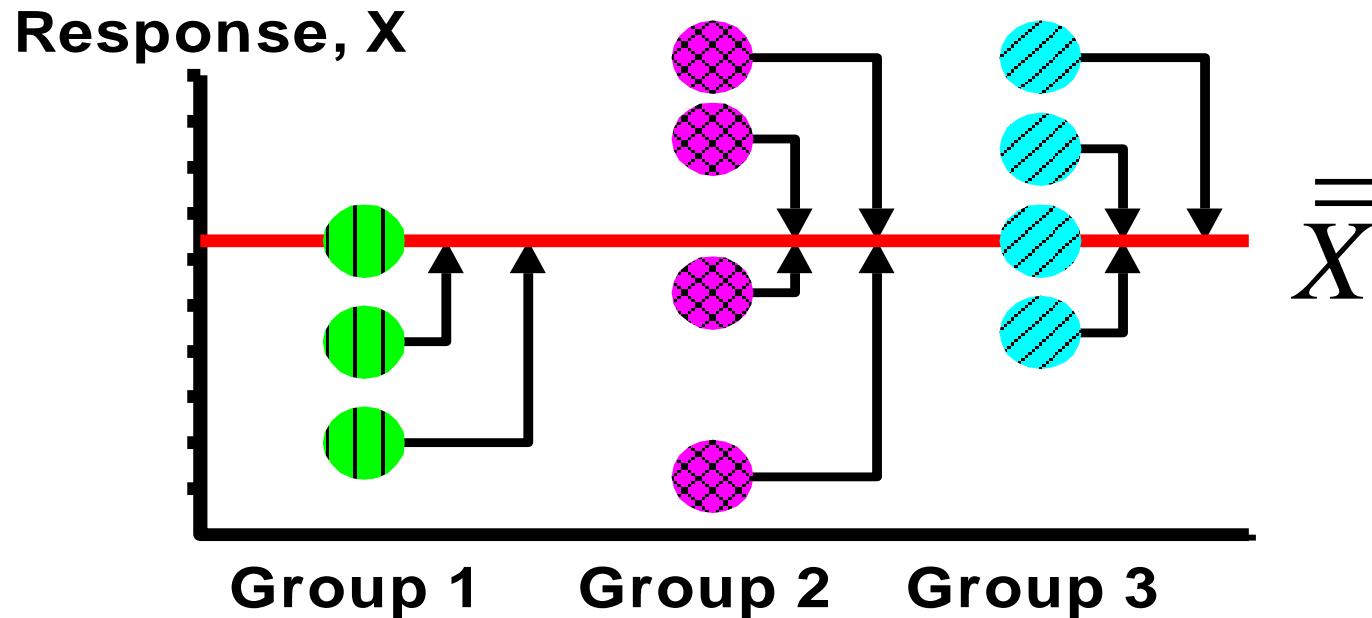
$X_{ij}$  = i<sup>th</sup> observation from group j

X = grand mean (mean of all data values)

# Total Variation



$$SST = (X_{11} - \bar{\bar{X}})^2 + (X_{12} - \bar{\bar{X}})^2 + \dots + (X_{cn_c} - \bar{\bar{X}})^2$$



# Among-Group Variation



$$SST = SSA + SSW$$

$$SSA = \sum_{j=1}^c n_j (\bar{X}_j - \bar{\bar{X}})^2$$

Where:

SSA = Sum of squares among groups

c = number of groups

$n_j$  = sample size from group j

$\bar{X}_j$  = sample mean from group j

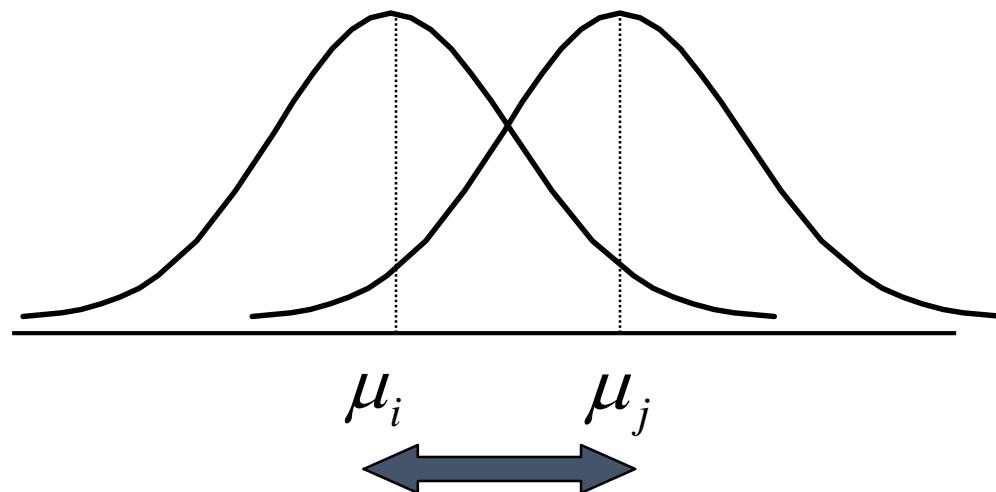
X = grand mean (mean of all data values)

# Among-Group Variation



$$SSA = \sum_{j=1}^c n_j (\bar{X}_j - \bar{\bar{X}})^2$$

Variation Due to  
Differences Among Groups



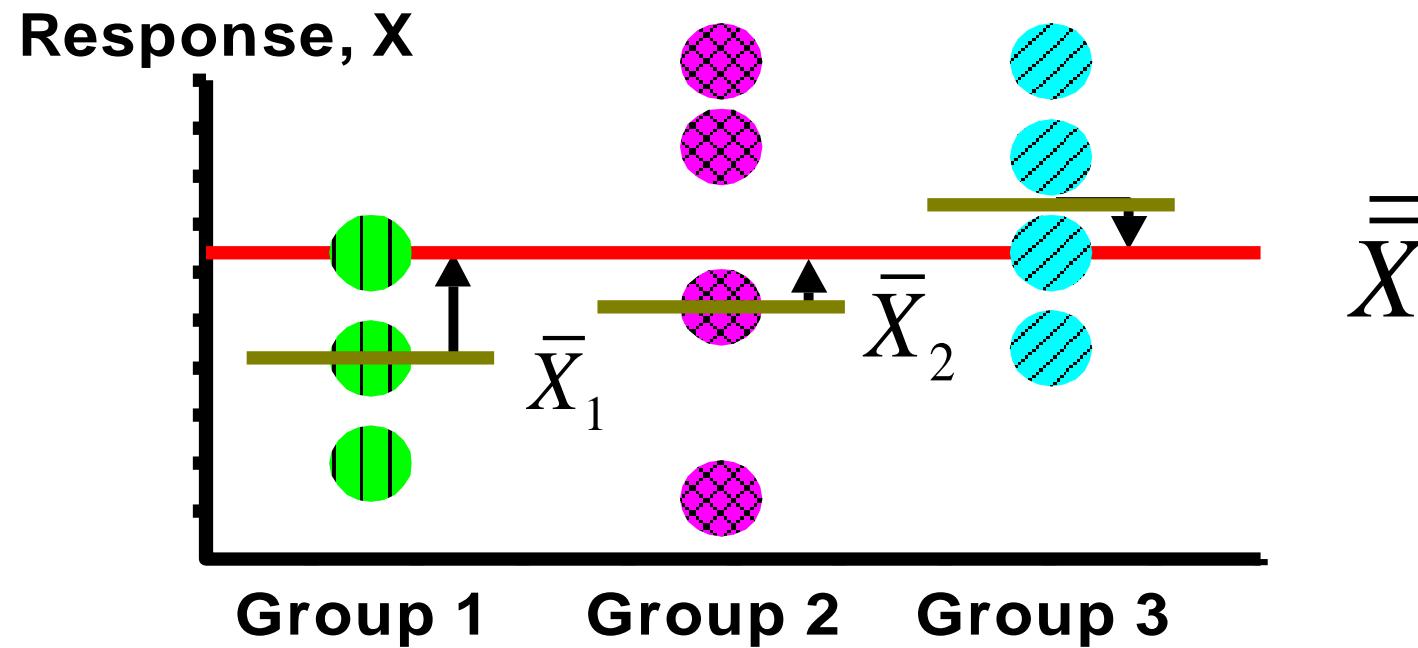
$$MSA = \frac{SSA}{c - 1}$$

Mean Square Among =  
SSA/degrees of freedom

# Among-Group Variation



$$SSA = n_1(\bar{X}_1 - \bar{\bar{X}})^2 + n_2(\bar{X}_2 - \bar{\bar{X}})^2 + \cdots + n_c(\bar{X}_c - \bar{\bar{X}})^2$$



# Within-Group Variation



$$SST = SSA + SSW$$

$$SSW = \sum_{j=1}^c \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2$$

Where:

$SSW$  = Sum of squares within groups

$c$  = number of groups

$n_j$  = sample size from group  $j$

$\bar{X}_j$  = sample mean from group  $j$

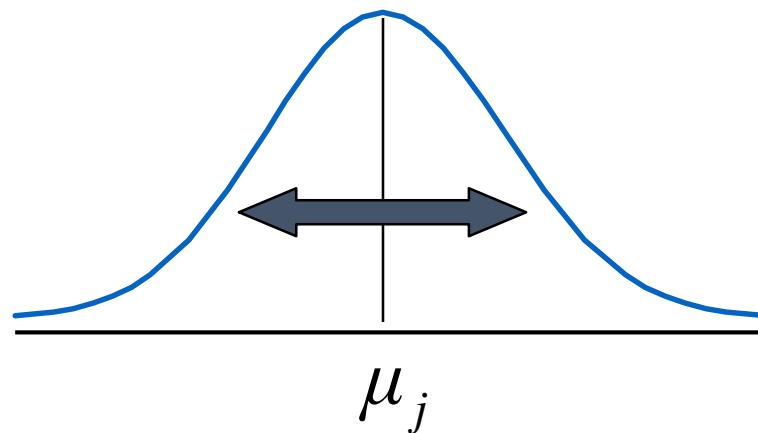
$X_{ij}$  =  $i^{\text{th}}$  observation in group  $j$

# Within-Group Variation



$$SSW = \sum_{j=1}^c \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2$$

Summing the variation within each group and then adding over all groups



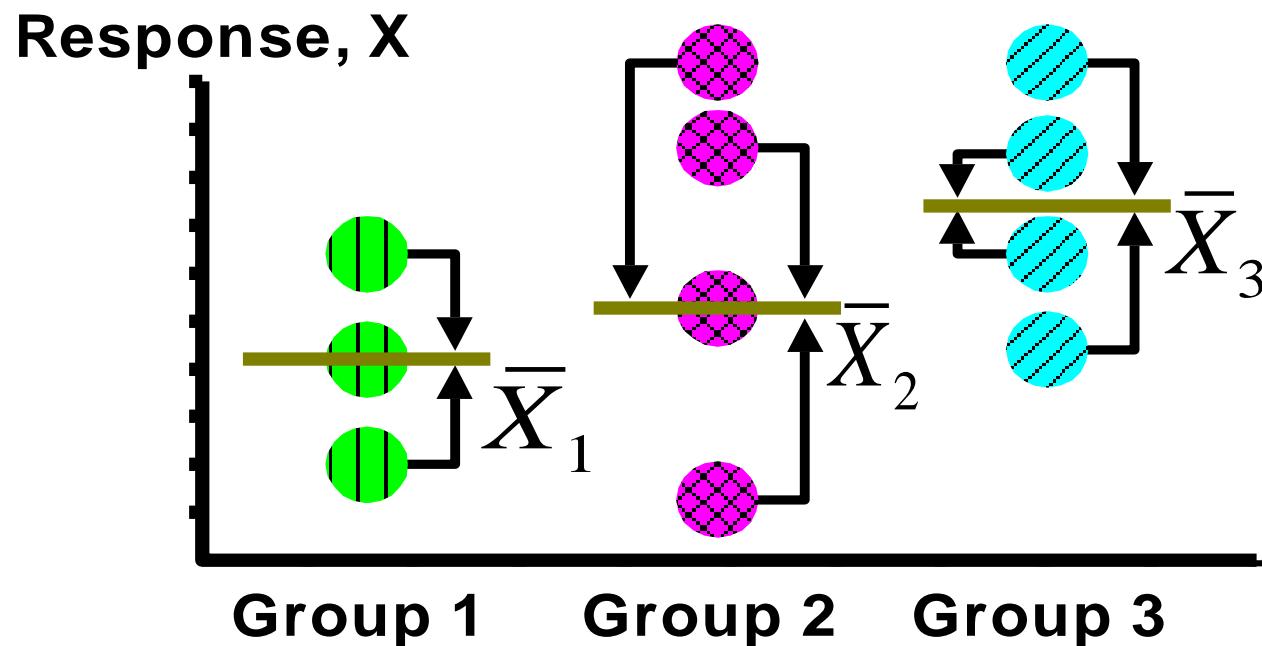
$$MSW = \frac{SSW}{n - c}$$

Mean Square Within =  
SSW/degrees of freedom

# Within-Group Variation



$$SSW = (X_{11} - \bar{X}_1)^2 + (X_{12} - \bar{X}_2)^2 + \cdots + (X_{cn_c} - \bar{X}_c)^2$$



# Obtaining the Mean Squares



The Mean Squares are obtained by dividing the various sum of squares by their associated degrees of freedom

$$MSA = \frac{SSA}{c - 1}$$

Mean Square Among  
(d.f. = c-1)

$$MSW = \frac{SSW}{n - c}$$

Mean Square Within  
(d.f. = n-c)

$$MST = \frac{SST}{n - 1}$$

Mean Square Total  
(d.f. = n-1)

# One-Way ANOVA Table



Source of Variation	Degrees of Freedom	Sum Of Squares	Mean Square (Variance)	F
Among Groups	$c - 1$	SSA	$MSA = \frac{SSA}{c - 1}$	$F_{STAT} =$
Within Groups	$n - c$	SSW	$MSW = \frac{SSW}{n - c}$	$\frac{MSA}{MSW}$
Total	$n - 1$	SST		

c = number of groups

n = sum of the sample sizes from all groups

df = degrees of freedom

# One-Way ANOVA - F Test Statistic



$$H_0: \mu_1 = \mu_2 = \dots = \mu_c$$

$H_1$ : At least two population means are different

- Test statistic

$$F_{STAT} = \frac{MSA}{MSW}$$

$MSA$  is mean squares **among** groups

$MSW$  is mean squares **within** groups

- Degrees of freedom

- $df_1 = c - 1$       ( $c$  = number of groups)

- $df_2 = n - c$       ( $n$  = sum of sample sizes from all populations)

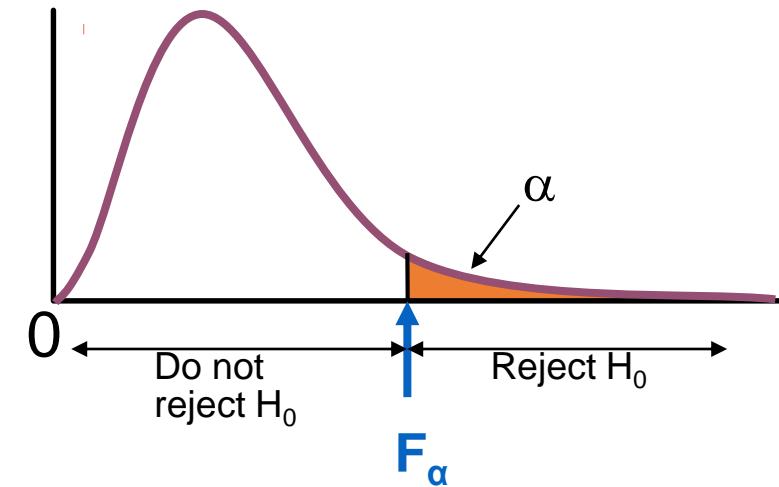
# Interpreting One-Way ANOVA F Statistic



- The F statistic is the ratio of the **among** estimate of variance and the **within** estimate of variance
  - The ratio must always be positive
  - $df_1 = c - 1$  will typically be small
  - $df_2 = n - c$  will typically be large

## Decision Rule:

- Reject  $H_0$  if  $F_{\text{STAT}} > F_\alpha$ , otherwise do not reject  $H_0$



# One-Way ANOVA F Test Example



- You want to see if three different golf clubs yield different distances.
- You randomly select five measurements from trials on an automated driving machine for each club.
- At the 0.05 significance level, is there a difference in mean distance?

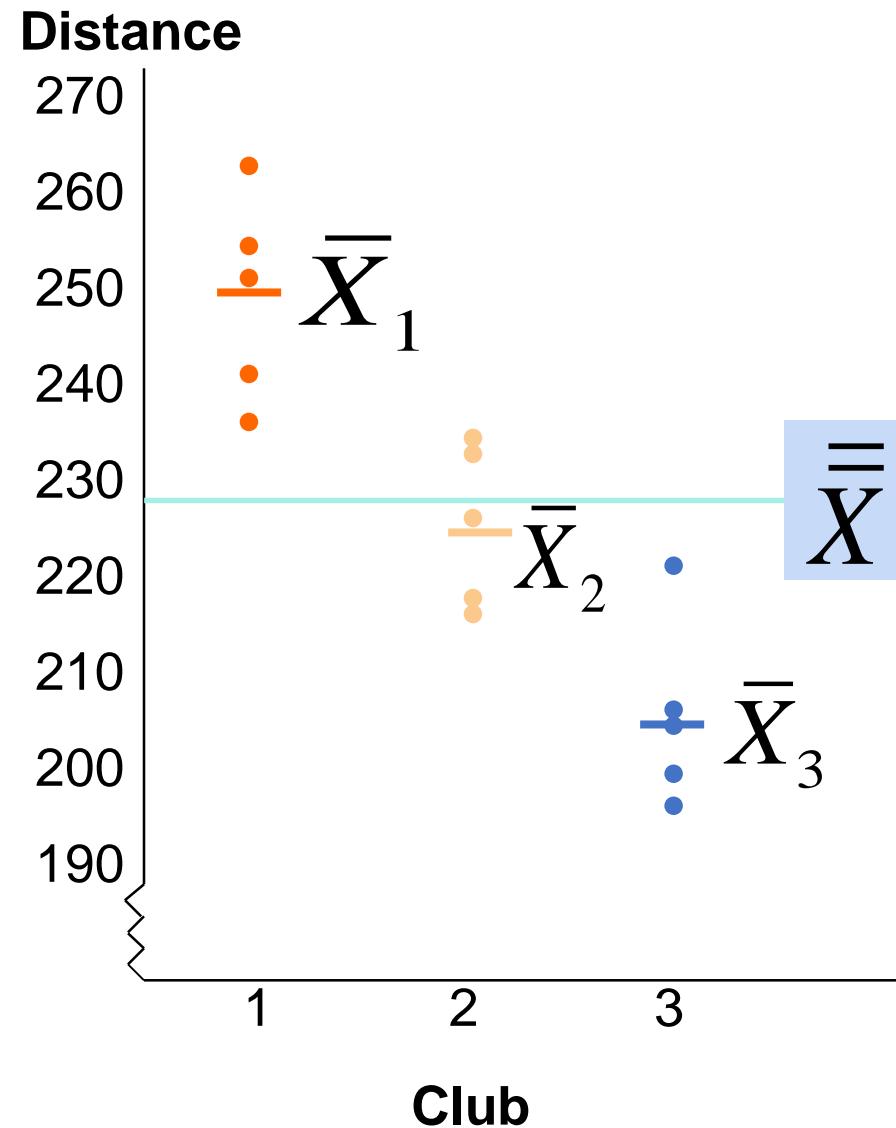
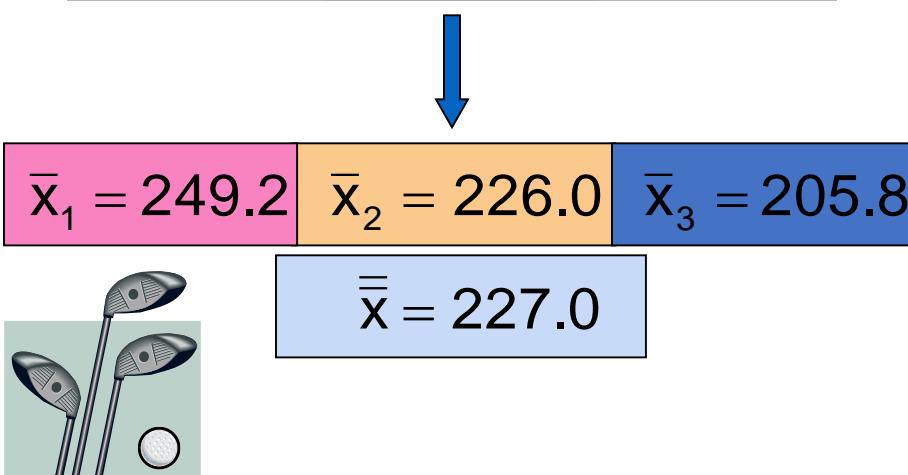
<b>Club 1</b>	<b>Club 2</b>	<b>Club 3</b>
254	234	200
263	218	222
241	235	197
237	227	206
251	216	204



# One-Way ANOVA Example: Scatter Plot



Club 1	Club 2	Club 3
254	234	200
263	218	222
241	235	197
237	227	206
251	216	204



# One-Way ANOVA Example Computations



<u>Club 1</u>	<u>Club 2</u>	<u>Club 3</u>		$\bar{X}_1 = 249.2$	$n_1 = 5$
254	234	200		$\bar{X}_2 = 226.0$	$n_2 = 5$
263	218	222		$\bar{X}_3 = 205.8$	$n_3 = 5$
241	235	197			
237	227	206			
251	216	204		$\bar{X} = 227.0$	$n = 15$

$c = 3$

A small illustration of three golf clubs standing upright next to a single golf ball.

$$SSA = 5 (249.2 - 227)^2 + 5 (226 - 227)^2 + 5 (205.8 - 227)^2 = 4716.4$$

$$SSW = (254 - 249.2)^2 + (263 - 249.2)^2 + \dots + (204 - 205.8)^2 = 1119.6$$

$$\downarrow$$
$$MSA = 4716.4 / (3-1) = 2358.2$$

$$MSW = 1119.6 / (15-3) = 93.3$$

$$\left. \right\} F_{STAT} = \frac{2358.2}{93.3} = 25.275$$

# One-Way ANOVA Example Solution

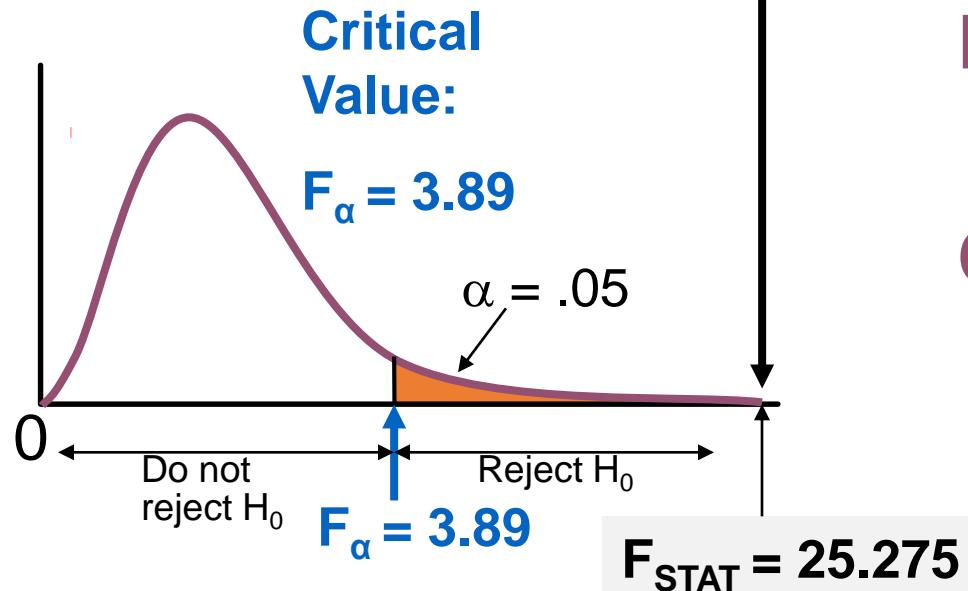


$$H_0: \mu_1 = \mu_2 = \mu_3$$

$$H_1: \mu_j \text{ not all equal}$$

$$\alpha = 0.05$$

$$df_1 = 2 \quad df_2 = 12$$



## Test Statistic:

$$F_{STAT} = \frac{MSA}{MSW} = \frac{2358.2}{93.3} = 25.275$$

## Decision:

Reject  $H_0$  at  $\alpha = 0.05$

## Conclusion:

There is evidence that at least one  $\mu_j$  differs from the rest

# One-Way ANOVA Excel Output

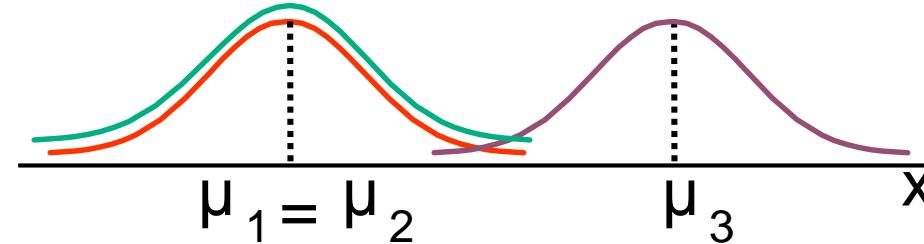


SUMMARY						
Groups	Count	Sum	Average	Variance		
Club 1	5	1246	249.2	108.2		
Club 2	5	1130	226	77.5		
Club 3	5	1029	205.8	94.2		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	4716.4	2	2358.2	25.275	4.99E-05	3.89
Within Groups	1119.6	12	93.3			
Total	5836.0	14				

# The Tukey-Kramer Procedure



- Tells **which** population means are significantly different
  - e.g.:  $\mu_1 = \mu_2 \neq \mu_3$
  - Done after rejection of equal means in ANOVA
- Allows paired comparisons
  - Compare absolute mean differences with critical range



# Tukey-Kramer Critical Range



$$\text{Critical Range} = Q_\alpha \sqrt{\frac{\text{MSW}}{2} \left( \frac{1}{n_j} + \frac{1}{n_{j'}} \right)}$$

where:

$Q_\alpha$  = Upper Tail Critical Value from Studentized Range Distribution with  $c$  and  $n - c$  degrees of freedom

MSW = Mean Square Within

$n_j$  and  $n_{j'}$  = Sample sizes from groups  $j$  and  $j'$

# The Tukey-Kramer Procedure: Example



<b>Club 1</b>	<b>Club 2</b>	<b>Club 3</b>
254	234	200
263	218	222
241	235	197
237	227	206
251	216	204

$$\bar{x}_1 = 249.2 \quad \bar{x}_2 = 226.0 \quad \bar{x}_3 = 205.8$$

$$\bar{\bar{x}} = 227.0$$

1. Compute absolute mean differences:

$$|\bar{x}_1 - \bar{x}_2| = |249.2 - 226.0| = 23.2$$

$$|\bar{x}_1 - \bar{x}_3| = |249.2 - 205.8| = 43.4$$

$$|\bar{x}_2 - \bar{x}_3| = |226.0 - 205.8| = 20.2$$

2. Find the  $Q_\alpha$  value from the table with  $c = 3$  and  $(n - c) = (15 - 3) = 12$  degrees of freedom:

$$Q_\alpha = 3.77$$

# The Tukey-Kramer Procedure: Example



## 3. Compute Critical Range:

$$\text{Critical Range} = Q_{\alpha} \sqrt{\frac{\text{MSW}}{2} \left( \frac{1}{n_j} + \frac{1}{n_{j'}} \right)} = 3.77 \sqrt{\frac{93.3}{2} \left( \frac{1}{5} + \frac{1}{5} \right)} = 16.285$$

## 4. Compare:

$$|\bar{x}_1 - \bar{x}_2| = 23.2$$

$$|\bar{x}_1 - \bar{x}_3| = 43.4$$

$$|\bar{x}_2 - \bar{x}_3| = 20.2$$

5. All of the absolute mean differences are greater than critical range. Therefore there is a significant difference between each pair of means at 5% level of significance.

# ANOVA Assumptions Levene's Test



- Tests the assumption that the variances of each population are equal.
- First, define the null and alternative hypotheses:
  - $H_0: \sigma^2_1 = \sigma^2_2 = \dots = \sigma^2_c$
  - $H_1: \text{Not all } \sigma^2_j \text{ are equal}$
- Second, compute the absolute value of the difference between each value and the median of each group.
- Third, perform a one-way ANOVA on these absolute differences.

# Levene Homogeneity Of Variance Test Example



$$H_0: \sigma^2_1 = \sigma^2_2 = \sigma^2_3$$

$H_1:$  Not all  $\sigma^2_j$  are equal

## Calculate Medians

Club 1	Club 2	Club 3
--------	--------	--------

237	216	197
-----	-----	-----

241	218	200
-----	-----	-----

251	227	204
-----	-----	-----

**Median**

254	234	206
-----	-----	-----

263	235	222
-----	-----	-----

## Calculate Absolute Differences

Club 1	Club 2	Club 3
--------	--------	--------

14	11	7
----	----	---

10	9	4
----	---	---

0	0	0
---	---	---

3	7	2
---	---	---

12	8	18
----	---	----

# Factorial Design : Two-Way ANOVA



- Examines the effect of
  - **Two factors of interest** on the dependent variable
    - e.g., compare Mileage across different brands of cars and in different States
  - **Interaction between the different levels** of these two factors
    - e.g., Does the effect of one particular brand depend on the State in which ?

# Two-Way ANOVA



- Assumptions
  - Populations are normally distributed
  - Populations have equal variances
  - Independent random samples are drawn

# Two-Way ANOVA - Sources of Variation



**Two Factors of interest: A and B**

r = number of levels of factor A

c = number of levels of factor B

n' = number of replications for each combination of factor levels.

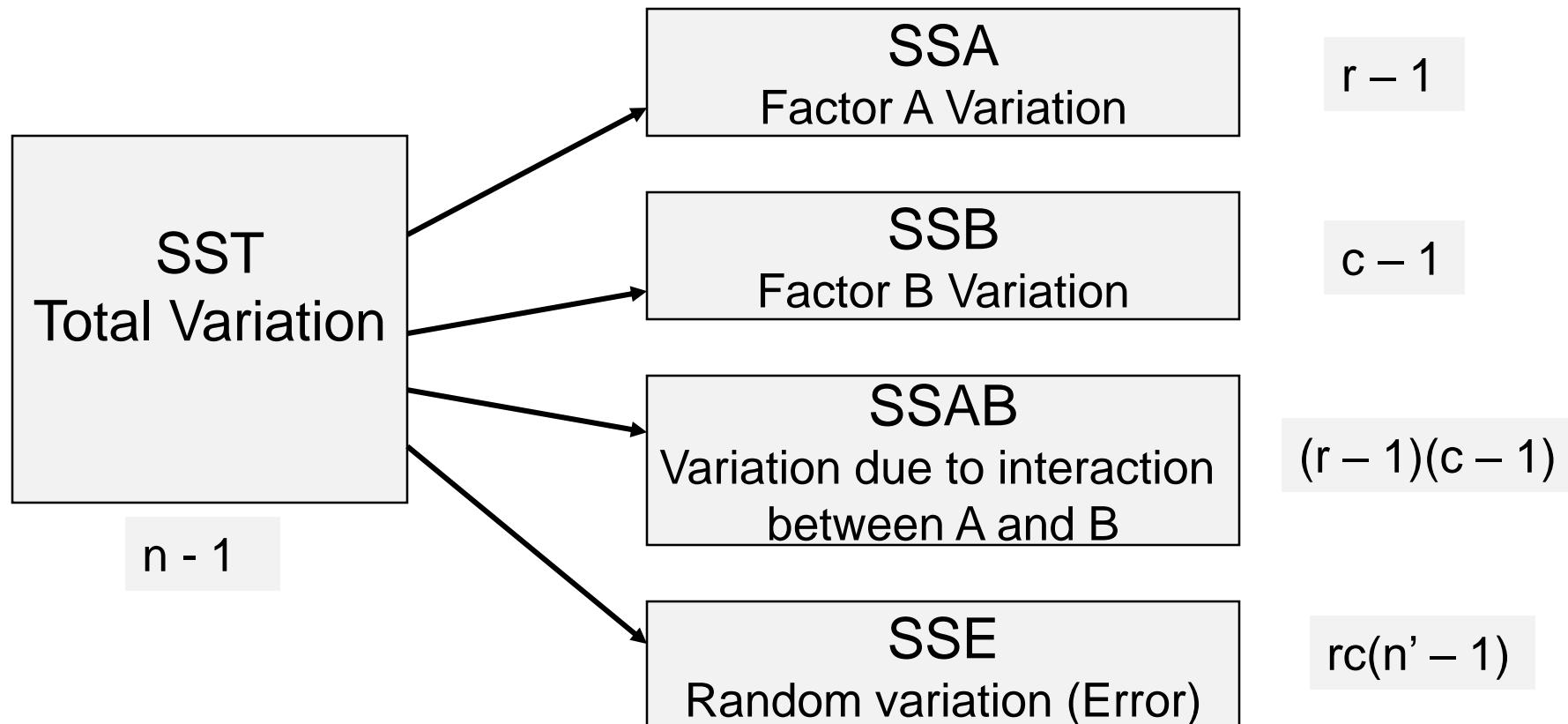
n = total number of observations in all cells

$X_{ijk}$  = value of the  $k^{\text{th}}$  observation of level i of factor A and level j of factor B

# Two-Way ANOVA Sources of Variation



$$SST = SSA + SSB + SSAB + SSE$$



# Two-Way ANOVA Equations



Total Variation:

$$SST = \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^{n'} (X_{ijk} - \bar{\bar{X}})^2$$

Factor A Variation:

$$SSA = cn' \sum_{i=1}^r (\bar{X}_{i..} - \bar{\bar{X}})^2$$

Factor B Variation:

$$SSB = rn' \sum_{j=1}^c (\bar{X}_{.j.} - \bar{\bar{X}})^2$$

# Two-Way ANOVA Equations



Interaction Variation:

$$SS_{AB} = n' \sum_{i=1}^r \sum_{j=1}^c (\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{\bar{X}})^2$$

Sum of Squares Error:

$$SSE = \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^{n'} (X_{ijk} - \bar{X}_{ij.})^2$$

# Two-Way ANOVA Equations



where:

$$\bar{\bar{X}} = \frac{\sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^{n'} X_{ijk}}{rcn'} = \text{Grand Mean}$$

$$\bar{X}_{i..} = \frac{\sum_{j=1}^c \sum_{k=1}^{n'} X_{ijk}}{cn'} = \text{Mean of } i^{\text{th}} \text{ level of factor A } (i=1, 2, \dots, r)$$

$$\bar{X}_{.j.} = \frac{\sum_{i=1}^r \sum_{k=1}^{n'} X_{ijk}}{rn'} = \text{Mean of } j^{\text{th}} \text{ level of factor B } (j=1, 2, \dots, c)$$

$$\bar{X}_{ij.} = \sum_{k=1}^{n'} \frac{X_{ijk}}{n'} = \text{Mean of cell ij}$$

r = number of levels of factor A  
c = number of levels of factor B  
n' = number of replications in each cell

# Mean Square Calculations



$$MSA = \text{Mean square factor A} = \frac{SSA}{r - 1}$$

$$MSB = \text{Mean square factor B} = \frac{SSB}{c - 1}$$

$$MSAB = \text{Mean square interaction} = \frac{SSAB}{(r - 1)(c - 1)}$$

$$MSE = \text{Mean square error} = \frac{SSE}{rc(n' - 1)}$$

# Two-Way ANOVA : The F Test Statistics



$H_0: \mu_{1..} = \mu_{2..} = \mu_{3..} = \dots = \mu_{r..}$

$H_1: \text{Not all } \mu_{i..} \text{ are equal}$

## F Test for Factor A Effect

$$F_{STAT} = \frac{MSA}{MSE}$$

Reject  $H_0$  if  
 $F_{STAT} > F_\alpha$

$H_0: \mu_{.1} = \mu_{.2} = \mu_{.3} = \dots = \mu_{.c}$

$H_1: \text{Not all } \mu_{.j} \text{ are equal}$

## F Test for Factor B Effect

$$F_{STAT} = \frac{MSB}{MSE}$$

Reject  $H_0$  if  
 $F_{STAT} > F_\alpha$

$H_0: \text{the interaction of A and B is equal to zero}$

$H_1: \text{interaction of A and B is not zero}$

## F Test for Interaction Effect

$$F_{STAT} = \frac{MSAB}{MSE}$$

Reject  $H_0$  if  
 $F_{STAT} > F_\alpha$

# Two-Way ANOVA - Summary Table



Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F
Factor A	<b>SSA</b>	$r - 1$	<b>MSA</b> $= SSA/(r - 1)$	$\frac{MSA}{MSE}$
Factor B	<b>SSB</b>	$c - 1$	<b>MSB</b> $= SSB / (c - 1)$	$\frac{MSB}{MSE}$
AB (Interaction)	<b>SSAB</b>	$(r - 1)(c - 1)$	<b>MSAB</b> $= SSAB / (r - 1)(c - 1)$	$\frac{MSAB}{MSE}$
Error	<b>SSE</b>	$rc(n' - 1)$	<b>MSE =</b> $SSE/rc(n' - 1)$	
Total	<b>SST</b>	$n - 1$		

# Features of Two-Way ANOVA - F Test

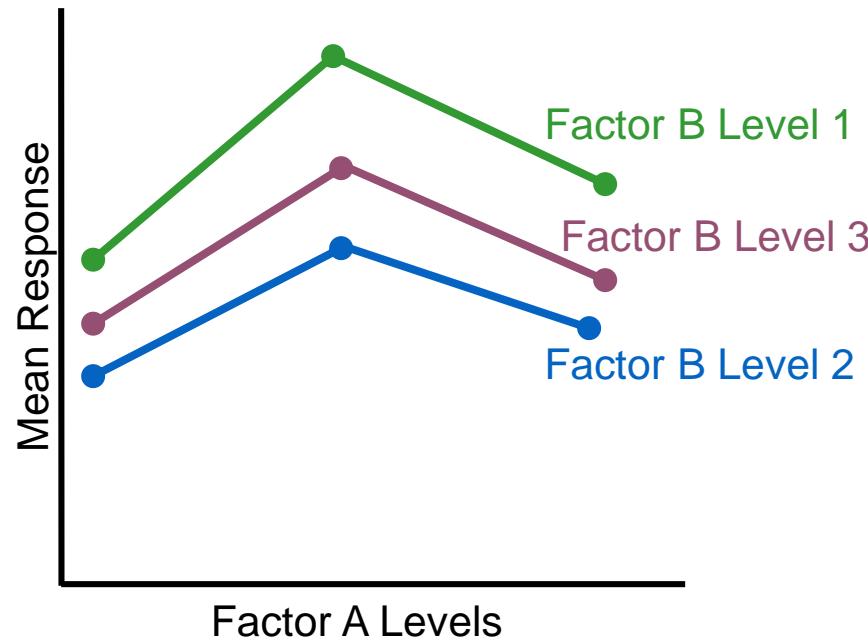


- Degrees of freedom always add up
  - $n-1 = rc(n'-1) + (r-1) + (c-1) + (r-1)(c-1)$
  - Total = error + factor A + factor B + interaction
- The denominators of the F Test are always the same but the numerators are different
- The sums of squares always add up
  - $SST = SSE + SSA + SSB + SSAB$
  - Total = error + factor A + factor B + interaction

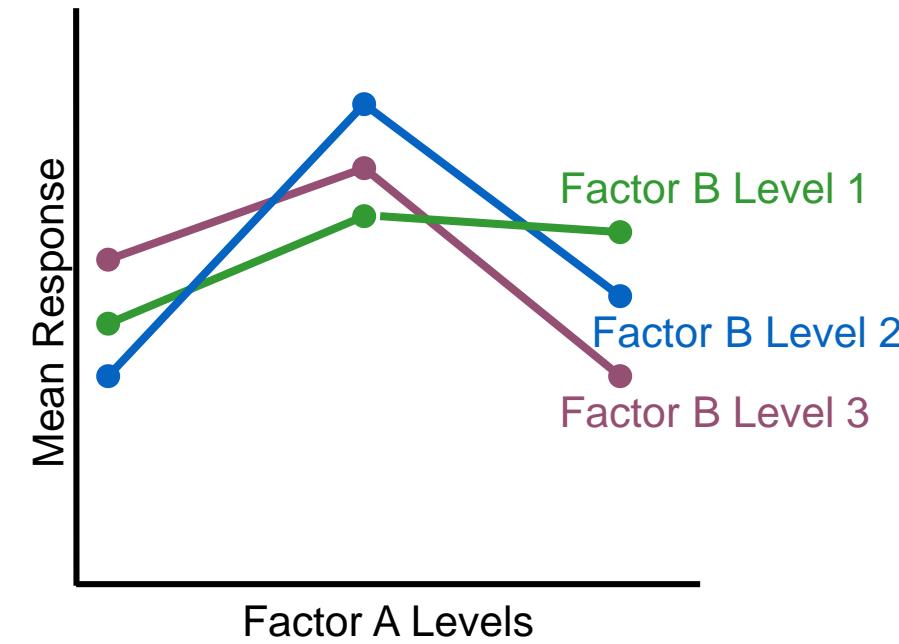
# Examples: Interaction vs. No Interaction



- No interaction: line segments are parallel



- Interaction is present: some line segments not parallel



# Multiple Comparisons: The Tukey Procedure



- Unless there is a significant interaction, you can determine the levels that are significantly different using the Tukey procedure
- Consider all absolute mean differences and compare to the calculated **critical range**

- Example: Absolute differences  
for factor A, assuming three levels:

$$|\bar{X}_{1..} - \bar{X}_{2..}|$$

$$|\bar{X}_{1..} - \bar{X}_{3..}|$$

$$|\bar{X}_{2..} - \bar{X}_{3..}|$$

# Multiple Comparisons: The Tukey Procedure



- Critical Range for Factor A:

$$\text{Critical Range} = Q_\alpha \sqrt{\frac{\text{MSE}}{c n'}}$$

(where  $Q_\alpha$  is from Table E.10 with r and  $rc(n'-1)$  d.f.)

---

- Critical Range for Factor B:

$$\text{Critical Range} = Q_\alpha \sqrt{\frac{\text{MSE}}{r n'}}$$

(where  $Q_\alpha$  is from Table E.10 with c and  $rc(n'-1)$  d.f.)



# Time Series Forecasting

# Time series Forecasting



- A collection of observations, each one being recorded at time t.

Year	1990	1991	1992	1993
Sale	71.2	74	73.5	76

- Example:
- Governments forecast unemployment rates, interest rates, and expected revenues from income taxes for policy purposes.
- Marketing executives forecast demand, sales, and consumer preferences for strategic planning
- College administrators forecast enrollments to plan for facilities and for faculty recruitment
- Retail stores forecast demand to control inventory levels, hire employees and provide training

# Components of Time series data



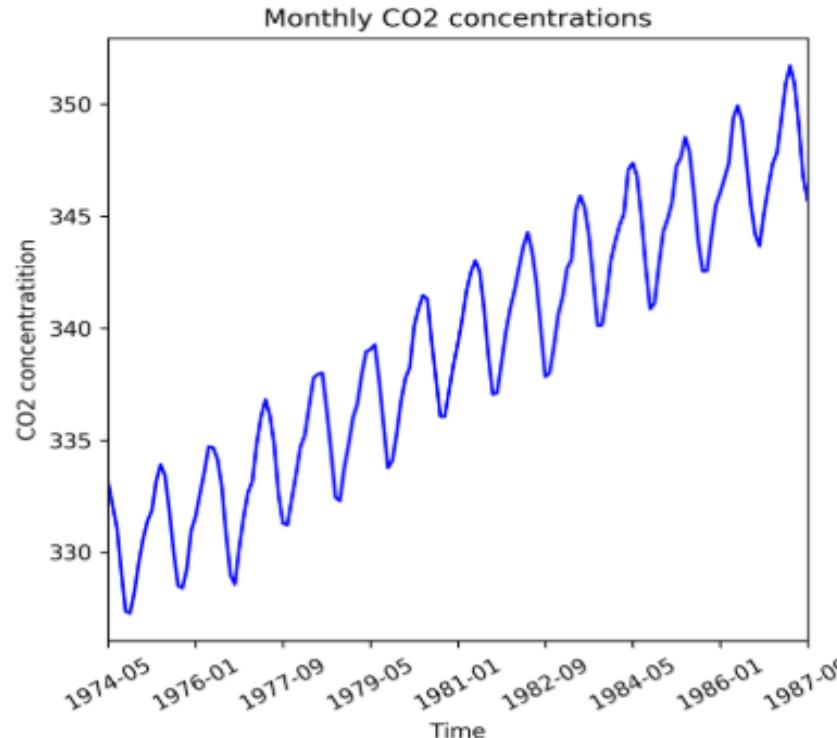
- General trend or secular trend
- Seasonality
- Cyclical movements
- Unexpected variations or irregular fluctuations

# Structure of Time series data

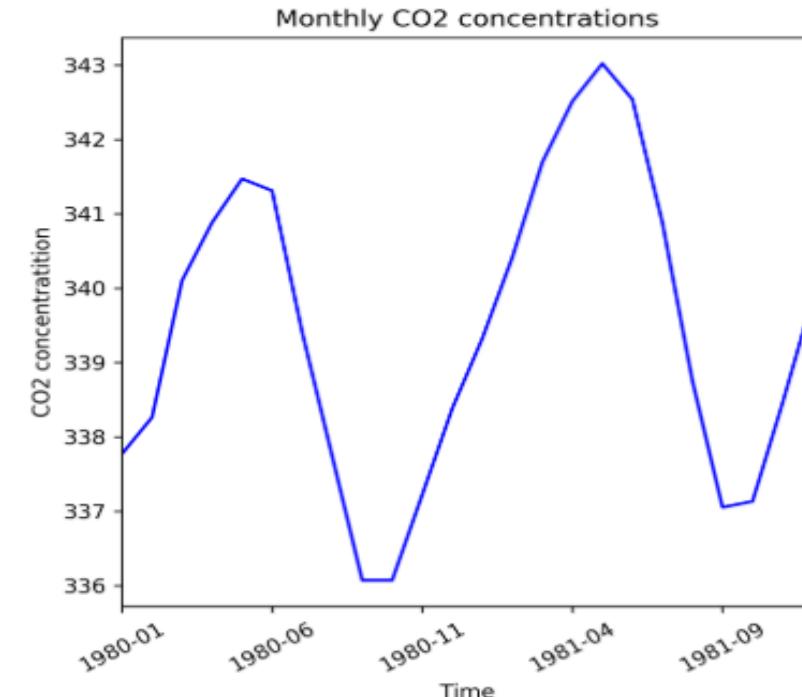


- General trend
  - When a time series exhibits an upward or downward movement in the long run, it is said to have a general trend.
- Example
  - CO<sub>2</sub> concentrations in air measured during 1974 through 1987:

**Time series of CO<sub>2</sub> readings with an upward trend**



**Shorter run of CO<sub>2</sub> readings time series which is not able to reveal general trend**

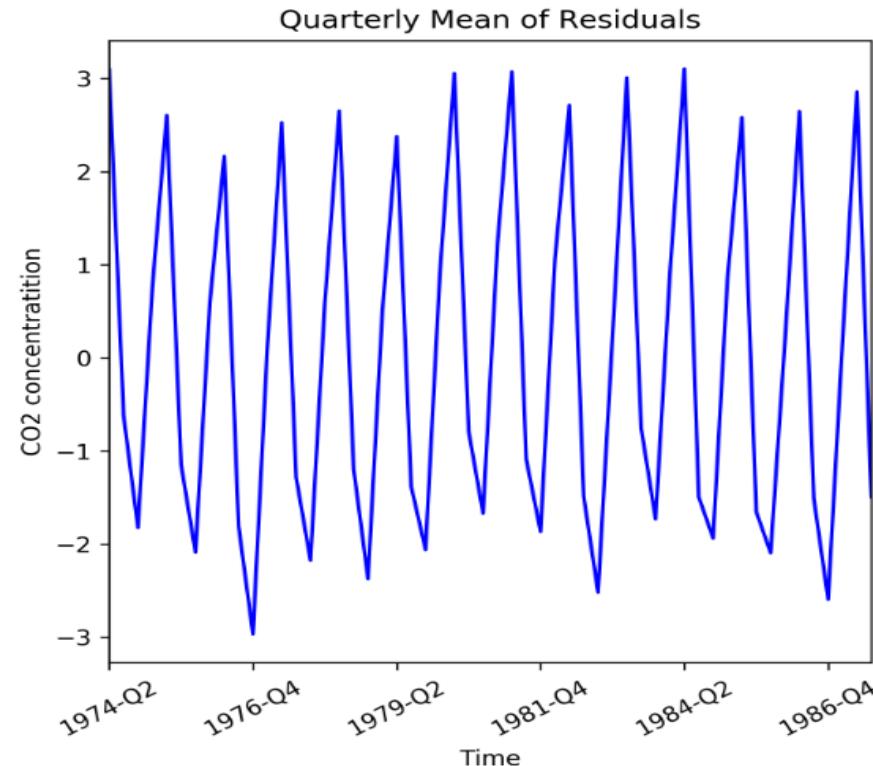


# Structure of Time series data



## ➤ Seasonality

- Seasonality refers to periodic fluctuations in time series data that happens at regular periods. While traditionally used to literally mean seasons (e.g. Spring, Summer, Autumn, Winter), it can occur during any time period, like hours, days, or weeks.

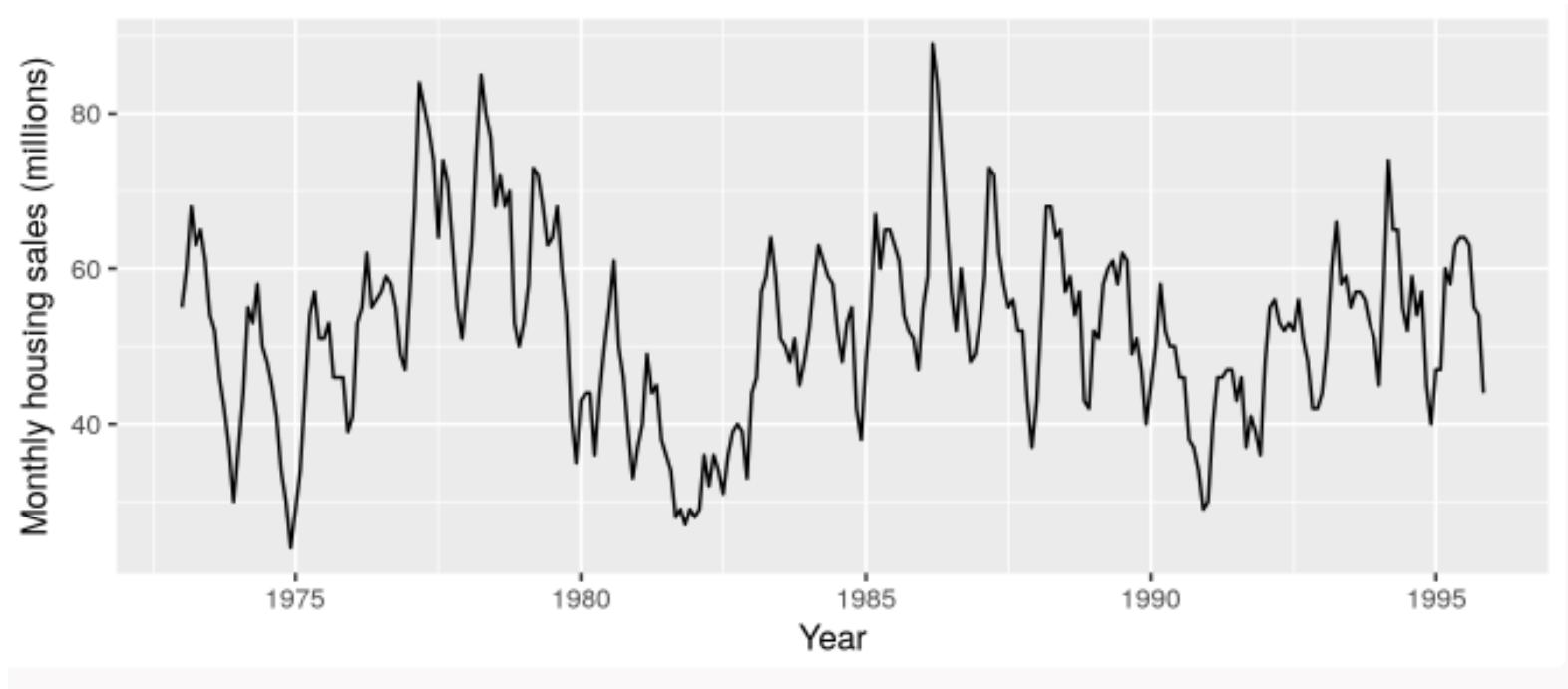


# Structure of Time series data



## ➤ Cyclical movements

- The cyclical component of a time series refers to (regular or periodic) fluctuations around the trend, excluding the irregular component, revealing a succession of phases of expansion and contraction.
- Example: Economic data affected by business cycles with a period varying between about 5 and 7 years.

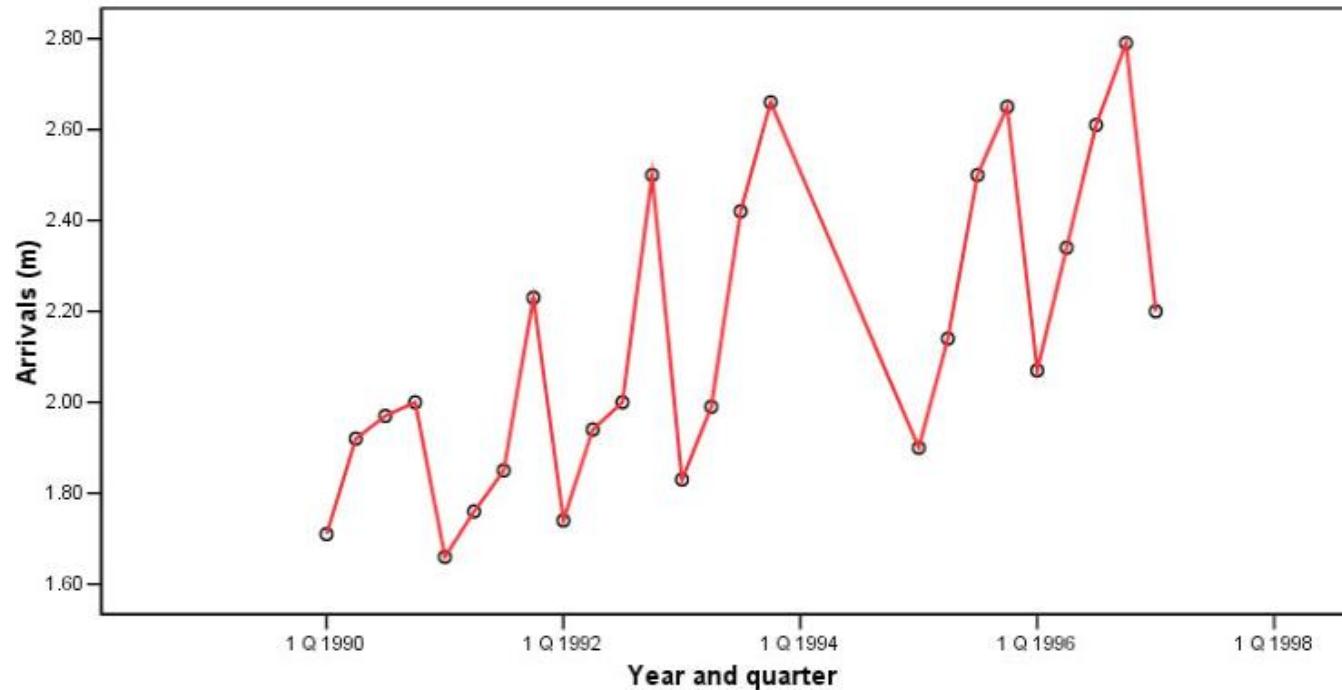


# Unexpected variations or irregular fluctuations



## ➤ Unexpected variations

- It is the irregular movements of the data over a period of time.
- Variations excluding trend, seasonal and cyclical variations are irregular fluctuations.
- Following chart tracks the flight arrivals at Gatwick airport per quarter over 6 years (measured in passenger millions).



# Moving Average Methods



## ➤ Moving Averages

- Mean of time series data from several consecutive periods.
- The name 'moving' indicates it is continually recomputed as new data becomes available.
- **Example:**
- If you have sales data for a twenty-year period, you can calculate a five-year moving average, a four-year moving average, a three-year moving average and so on.
- Stock market analysts will often use a 50 or 200 day moving average to help them see trends in the stock market and (hopefully) forecast where the stocks are headed.

# Moving Average Method



- **Formula:** Five-year moving average

- First average:

$$MA(5) = \frac{Y_1 + Y_2 + Y_3 + Y_4 + Y_5}{5}$$

- Second average:

$$MA(5) = \frac{Y_2 + Y_3 + Y_4 + Y_5 + Y_6}{5}$$

- etc.

# Moving Average Method - Formula

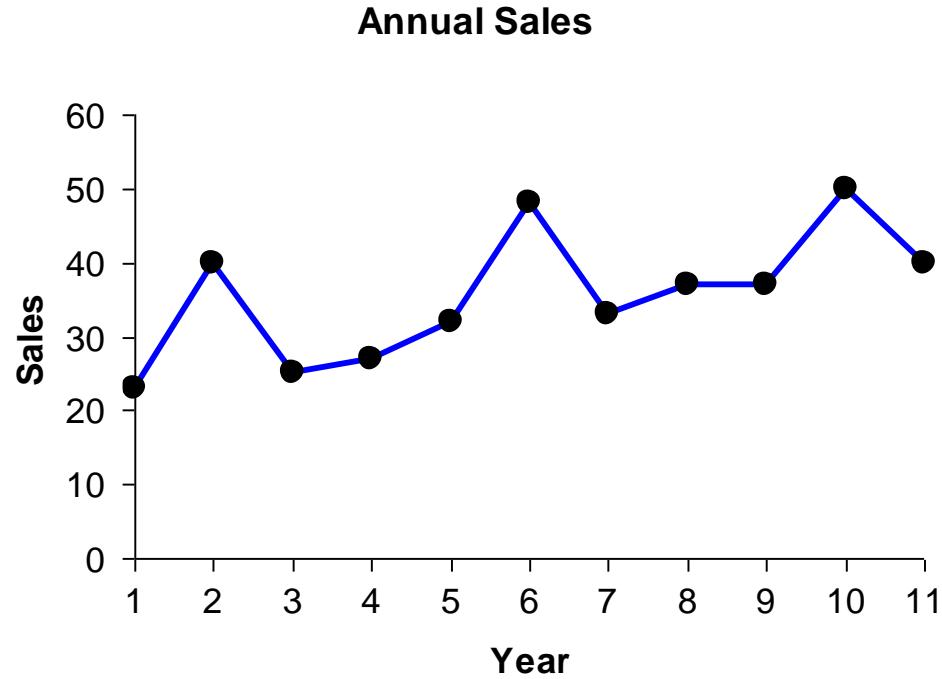


Years (t)	Variable (Y)	4-year Moving Averages	4-year Moving Averages Centered
$t_1$	$Y_1$	-----	-----
$t_2$	$Y_2$	$\frac{Y_1 + Y_2 + Y_3 + Y_4}{4} = a_1$	-----
$t_3$	$Y_3$	$\frac{Y_2 + Y_3 + Y_4 + Y_5}{4} = a_2$	$\frac{a_1 + a_2}{2} = A_1$
$t_4$	$Y_4$	$\frac{Y_3 + Y_4 + Y_5 + Y_6}{4} = a_3$	$\frac{a_2 + a_3}{2} = A_2$
$t_5$	$Y_5$	:	:
:	:	:	:

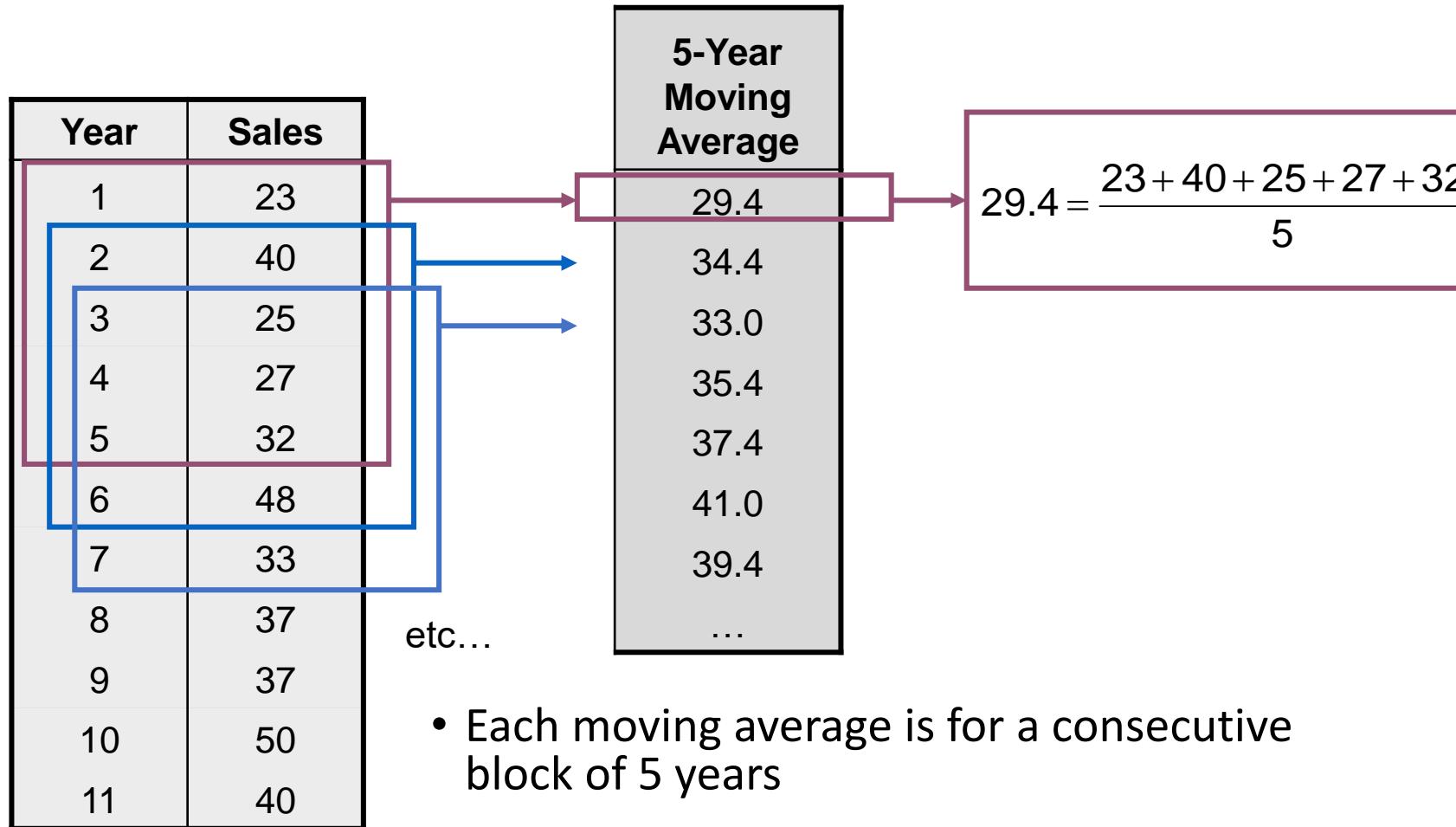
# Moving Average Method



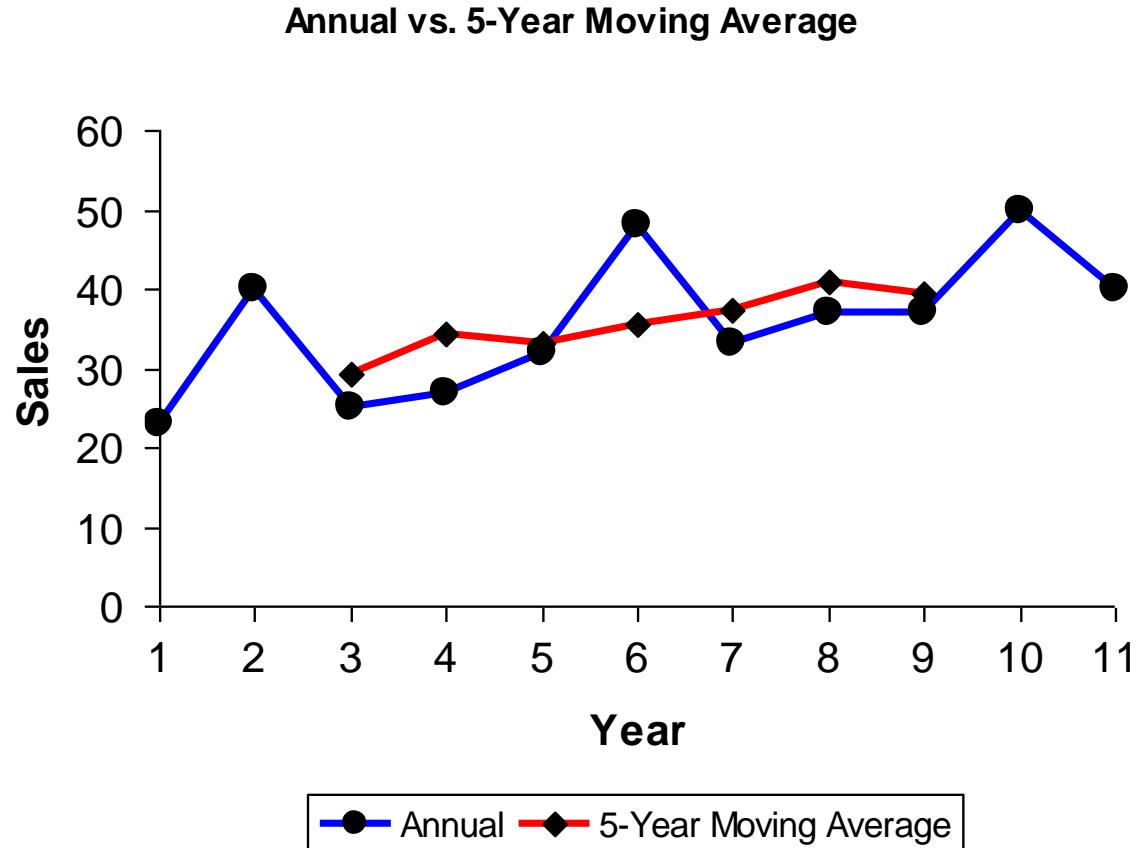
Year	Sales
1	23
2	40
3	25
4	27
5	32
6	48
7	33
8	37
9	37
10	50
11	40
etc...	etc...



# Moving Average Method



# Moving Average Method - Graph



# Moving Average Method - Example



Example: Compute 5-year, 7-year and 9-year moving averages for the following data.

Years	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
Values	2	4	6	8	10	12	14	16	18	20	22

Years	Values	5-Year Moving		7-Year Moving		9-Year Moving	
		Total	Average	Total	Average	Total	Average
1990	2	---	---	---	---	---	---
1991	4	---	---	---	---	---	---
1992	6	30	6	---	---	---	---
1993	8	40	8	56	8	---	---
1994	10	50	10	70	10	90	10
1995	12	60	12	84	12	108	12
1996	14	70	14	98	14	126	14
1997	16	80	16	112	16	---	---
1998	18	90	18	---	---	---	---
1999	20	---	---	---	---	---	---
2000	22	---	---	---	---	---	---

# Univariate data



The term refers to a time series that consists of single observations recorded sequentially over equal time increments.

Examples :

Monthly CO2 concentrations and Stock value

CO2	Year	Month
333.13	1974	5
332.09	1974	6
331.10	1974	7
329.14	1974	8
327.36	1974	9
327.29	1974	10

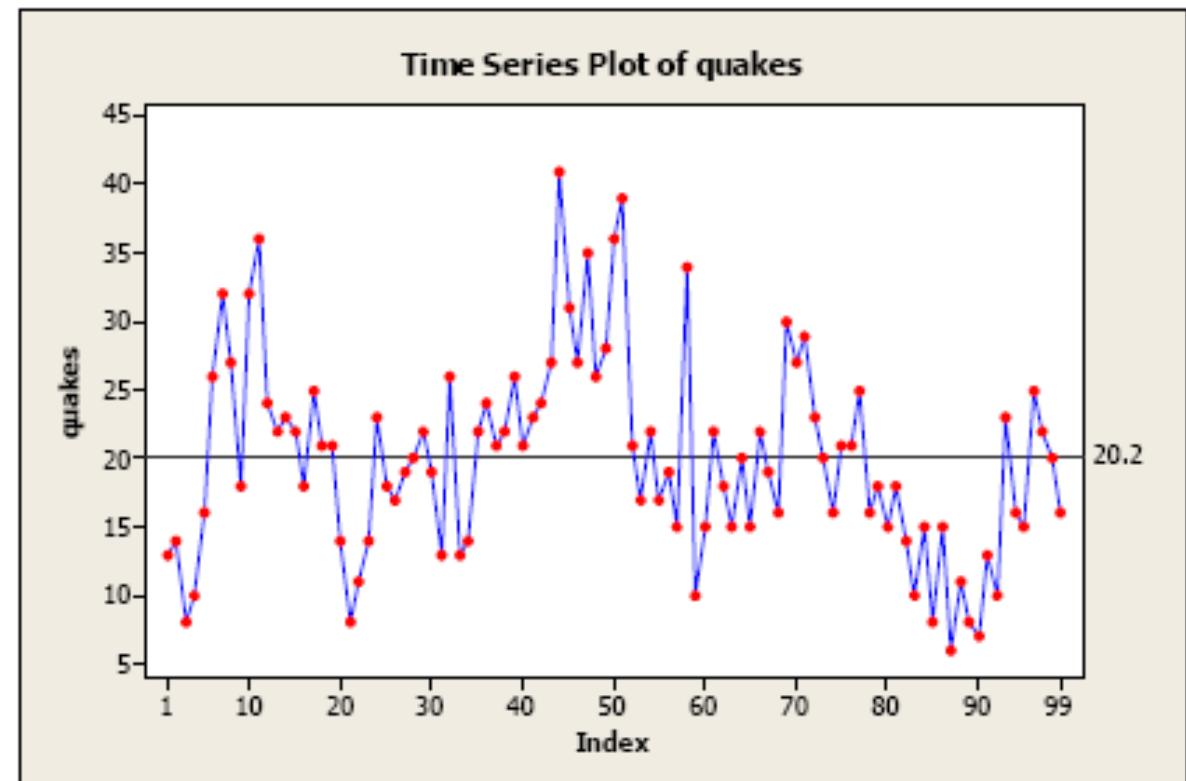
The idea behind this is to use the past values to forecast future trends

# Example



The following plot is a time series plot of the annual number of earthquakes in the world with seismic magnitude over 7.0, for a 99 consecutive years. By a time series plot, we simply mean that the variable is plotted against time.

- There is **no consistent trend** (upward or downward) over the entire time span. The series appears to slowly wander up and down. The horizontal line drawn at quakes = 20.2 indicates the mean of the series. Notice that the series tends to stay on the same side of the mean (above or below) for a while and then wanders to the other side.
- Almost by definition, there is **no seasonality** as the data are annual data.
- There are **no obvious outliers**.
- It's difficult to judge whether the variance is constant or not.



# Application of univariate time series in terms of forecasting



- Forecasting inflation rate or unemployment rate in the near future could be of interest to the government.
- Firms may be interested in demand for their product or the market share of the product
- Forecasting gold or silver prices by the jewel merchant

# Patterns emerging in time series data



- Is there a trend, meaning that, on average, the measurements tend to increase (or decrease) over time?
- Is there seasonality, meaning that there is a regularly repeating pattern of highs and lows related to calendar time such as seasons, quarters, months, days of the week, and so on?
- Are there outliers? With time series data, your outliers are far away from your other data.
- Is there constant variance over time, or is the variance non-constant?
- Are there any abrupt changes to either the level of the series or the variance?

# White noise



- A series is called white noise when it is purely random in nature.

Let  $\{\varepsilon_t\}$  denote a series.

Then it has zero mean  $[\sum (\varepsilon_t) = 0]$

Has a constant variance  $[V(\varepsilon_t) = \sigma^2]$

The scatter plot of such series across time will indicate no pattern and hence forecasting the future values of such series is not possible

# Auto Regressive Model



An AR model is one in which  $Y$  depends on its own past values  $Y_{t-1}$ ,  $Y_{t-2}$ ,  $Y_{t-3}$ , etc  
Thus,

$$Y_t = f(Y_{t-1}, Y_{t-2}, Y_{t-3}, \dots, \varepsilon_t)$$

A common representation of auto regressive model AR(p) is

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \dots + \beta_p Y_{t-p} + \varepsilon_t$$

# Moving Average Model



A moving average model is one when  $Y_t$  depends only on the random error terms

$$Y_t = f(\varepsilon_{t-1}, \varepsilon_{t-2}, \varepsilon_{t-3}, \dots)$$

The representation of moving average model MA(q) is

$$Y_t = \beta_0 + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \phi_3 \varepsilon_{t-3} + \dots + \phi_q \varepsilon_{t-q}$$

# Auto Regressive Moving Average Model



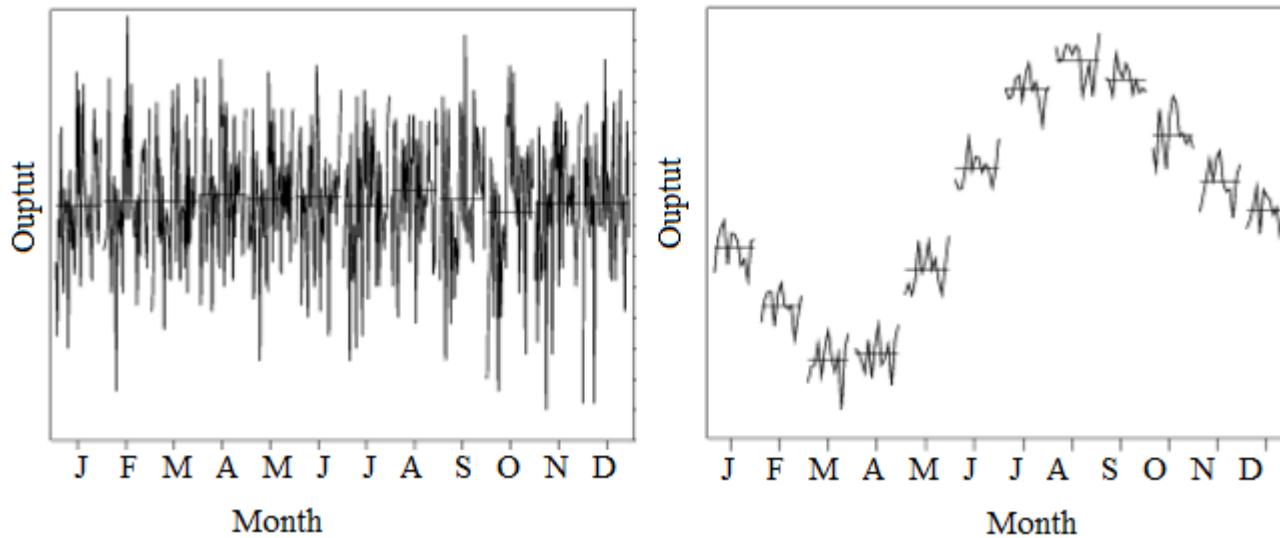
The time series may be represented as a mix of both AR and Ma models referred as ARMA(p,q)

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \dots + \beta_p Y_{t-p} + \varepsilon_t + \\ \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \phi_3 \varepsilon_{t-3} + \dots + \phi_q \varepsilon_{t-q} .$$

# Stationary series



A time series has stationarity if a shift in time doesn't cause a change in the shape of the distribution. Basic properties of the distribution like the mean , variance and covariance are constant over time.



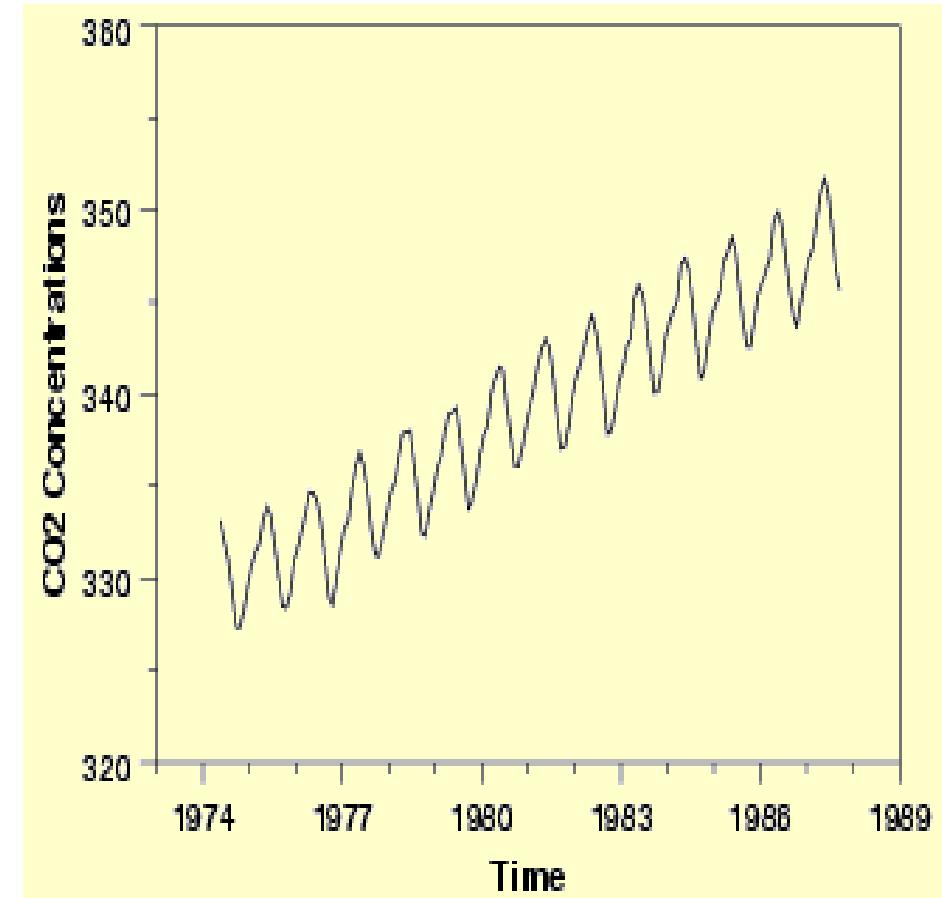
# Stationary series - Example



## ➤ Sample dataset:

CO<sub>2</sub> concentration in the year and month

CO <sub>2</sub>	Year	Month
333.13	1974	5
332.09	1974	6
331.1	1974	7
329.14	1974	8
327.36	1974	9
327.29	1974	10
328.23	1974	11
329.55	1974	12



➤ The initial run sequence plot of the data indicates a rising trend.

# Stationary Time series Transformation



The two most common ways to make a non-stationary time series curve stationary are:

- Differencing
- Log Transforming

# Stationary Time series Transformation



## Differencing:

In order to make your series stationary, you take a **difference between the data points**.

Differencing can help **stabilize the mean** of a time series by removing changes in the level of a time series, and so eliminating trend and seasonality.

- Data series

$X_1, X_2, X_3, X_4, X_5, X_6, \dots, X_n$

You series with difference of degree 1 becomes:

$$\text{If } d=0: y_t = Y_t$$

$$\text{If } d=1: y_t = Y_t - Y_{t-1}$$

$$\text{If } d=2: y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2}$$

## Data series

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2003	141	157	185	199	203	189	207	207	171	150	138	165
2004	145	168	197	208	210	209	238	238	199	168	152	196
2005	183	200	249	251	289	249	279	279	232	204	194	232
2006	215	239	270	279	307	305	322	339	263	241	229	272

## Data series after taking difference

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2003		16	28	14	4	-14	18	0	-36	-21	-12	27
2004	-20	23	29	11	2	-1	29	0	-39	-31	-16	44
2005	-13	17	49	2	38	-40	30	0	-47	-28	-10	38
2006	-17	24	31	9	28	-2	17	17	-76	-22	-12	43

- Data series after taking difference

$(X_2 - X_1, X_3 - X_2, X_4 - X_3, \dots, X_n - X_{n-1})$

# Stationary Time series Transformation



## Log transforming :

- Transformations such as logarithms can help to **stabilize the variance of a time series**.

## Data series

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2003	141	157	185	199	203	189	207	207	171	150	138	165
2004	145	168	197	208	210	209	238	238	199	168	152	196
2005	183	200	249	251	289	249	279	279	232	204	194	232
2006	215	239	270	279	307	305	322	339	263	241	229	272

## Data series after taking log

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2003	2.149219	2.195900	2.267172	2.298853	2.307496	2.276462	2.315970	2.315970	2.232996	2.176091	2.139879	2.217484
2004	2.161368	2.225309	2.294466	2.318063	2.322219	2.320146	2.376577	2.376577	2.298853	2.225309	2.181844	2.292256
2005	2.262451	2.301030	2.396199	2.399674	2.460898	2.396199	2.445604	2.445604	2.365488	2.309630	2.287802	2.365488
2006	2.332438	2.378398	2.431364	2.445604	2.487138	2.484300	2.507856	2.530200	2.419956	2.382017	2.359835	2.434569

## ARIMA model which can be written as a linear equation

$$Y_t = c + \phi_1 y_{d-t-1} + \phi_p y_{d-t-p} + \dots + \theta_1 e_{t-1} + \theta_q e_{t-q} + e_t$$

# ARIMA Steps



1. Visualize the time series
2. Stationarize the series
3. Plot ACF/PACF charts and find optimal parameters
4. Build the ARIMA model
5. Make Predictions

# ARIMA



ARIMA stands for auto-regressive integrated moving average

It is specified by these three order parameters: (p, d, q).

The process of fitting an ARIMA model is sometimes referred to as the Box-Jenkins method.

- Auto Regression (AR) – In auto-regression the values of a given time series data are regressed on their own lagged values, which is indicated by the “p” value in the model.
- Differencing (I-for Integrated) – This involves differencing the time series data to remove the trend and convert a non-stationary time series to a stationary one. This is indicated by the “d” value in the model. If  $d = 1$ , it looks at the difference between two time series entries
- Moving Average (MA) – The moving average nature of the model is represented by the “q” value which is the number of lagged values of the error term.

# Identification of p - ARIMA



Identify Autoregressive (AR) and Moving average (MA) processes by using the Autocorrelation function (ACF) and Partial Autocorrelation function (PACF).

## **Identifying the p order of AR model**

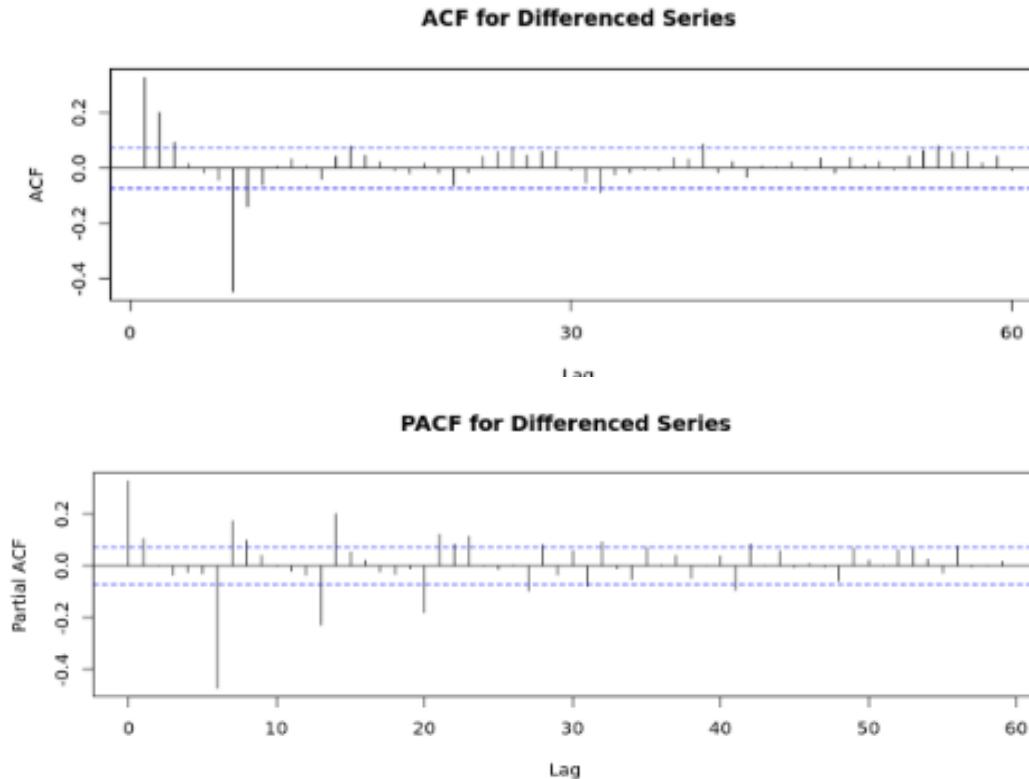
- For AR models, the ACF will dampen exponentially and the PACF will be used to identify the order (p) of the AR model.
- If we have one significant spike at lag 1 on the PACF, then we have an AR model of the order 1, i.e. AR(1).
- If we have significant spikes at lag 1, 2, and 3 on the PACF, then we have an AR model of the order 3, i.e. AR(3).

# Identification of q - ARIMA



- **Identifying the q order of MA model**
- For MA models, the PACF will dampen exponentially and the ACF plot will be used to identify the order of the MA process.
- If we have one significant spike at lag 1 on the ACF, then we have an MA model of the order 1, i.e. MA(1).
- If we have significant spikes at lag 1, 2, and 3 on the ACF, then we have an MA model of the order 3, i.e. MA(3).

# ACF & PACF



- There are significant auto correlations at lag 1 and 2 and beyond.
- Partial correlation plots show a significant spike at lag 1 ,2 and 7.
- This suggests that we might want to test models with AR or MA components of order 1, 2, or 7.
- A spike at lag 7 might suggest that there is a seasonal pattern present, perhaps as day of the week.

# ARIMA MODEL



- Build ARIMA Model for any dataset
- Predict from the ARIMA model which is build
- Akaike information criterion (AIC) is a fined technique based on in-sample fit to estimate the likelihood of a model to predict/estimate the future values.
- A good model is the one that has minimum AIC among all the other models.



Thank you