Unit 2
Univariate Analysis

Dr. S. Thenmozhi

Contents

- What is Statistics?
- Population and Sample
- Descriptive and Inferential statistics
- Variables and Data
- Uni-variate Data Analysis
- Data Visualization
- Measures of Center & Spread
- Normal distribution
- Z-Scores
- Tail Test

News on measures of center

- Vehicle population in Bengaluru has crossed more than 72 lakh, and on an average close to 2,000 new vehicles hit the roads daily. – New Indian Express
- House prices are almost five times the average local salary
- Sea levels rising in Antartica have been nearly a third faster than the global average

Statistics

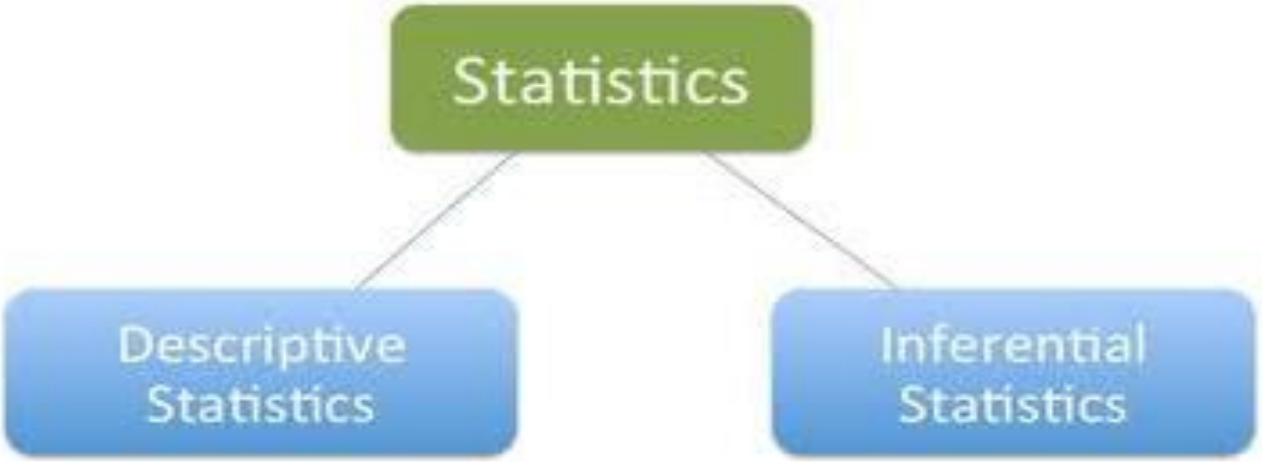
- A collection of methods for planning experiments, obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting and drawing conclusions based on the data.
  - Planning and carrying out research activities
  - Summarizing and exploring data
  - Making predictions and generalizing the phenomenon represented by data

Example

- Calculating the average length of downtime of a computer
- Data on No. of persons attending a seminar
- Evaluating effectiveness of products
- Predicting rainfall for the year
- Studying the vibrations of airplane wings

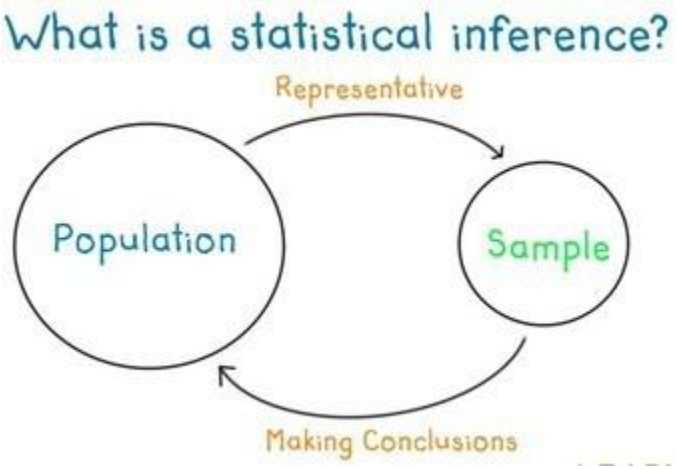
Types of Statistics

- Descriptive statistics
  - Consists of methods of organizing and summarizing information
  - includes the construction of graphs, charts, and tables and the calculation of various descriptive measures such as averages, measures of variation, and percentiles.
- Inferential Statistics
  - Consist of methods for drawing and measuring the reliability of conclusions about population based on information obtained from a sample of the population.
  - includes methods like point estimation, interval estimation and hypothesis testing which are all based on probability theory



Presenting, organizing and summarizing data

Drawing conclusions about a population based on data observed in a sample



Example

- Consider event of tossing dice. The dice is rolled 100 times and the results are forming the sample data.
- Descriptive statistics is finding out the frequencies in the sample data
- Inferential statistics can now be used to verify whether the dice is a fair or not

Data

- Data
  - Data is a collection of facts, such as numbers, words, measurements, observations or even just descriptions of things
  - Attributes that have been collected for observations
- Observation
  - The data collected for a particular instance
  - In DBMS perspective, a record is an observation

- Population
  - Population is the broader group of people to whom you intend to generalize the results of your study
  - The complete collection of all observations
- Sample
  - Sample is the group of individuals who actually participate in your study
  - A sub-collection of elements drawn from a population

Population or Sample

- **There are 120 people in your local football club**

Population	Sample
Asking age for all the people (all 120)	Choose some people and asking for age (may be 30)
Accurate	Not accurate, but may be good enough
Hard to do	Easier

BASIS FOR COMPARISON	POPULATION	SAMPLE
Meaning	Population refers to the collection of all elements possessing common characteristics, that comprises universe.	Sample means a subgroup of the members of population chosen for participation in the study.
Includes	Each and every unit of the group.	Only a handful of units of population.
Characteristic	Parameter	Statistic
Data collection	Complete enumeration or census	Sample survey or sampling
Focus on	Identifying the characteristics.	Making inferences about population.

Sampling methods

- Simple Random sampling
  - Which warrants that each subgroup of the population of size n has an equal probability of being picked as the sample
  - Sample with replacement
  - Sample without replacement
- Stratified Sampling
  - the data population is divided into groups, and samples are taken from each group.
  - The partitioning of the population into groups is called stratum.
- Two different samples from the same population will vary from each other as well. This phenomenon is known as **sampling variation**

Variables

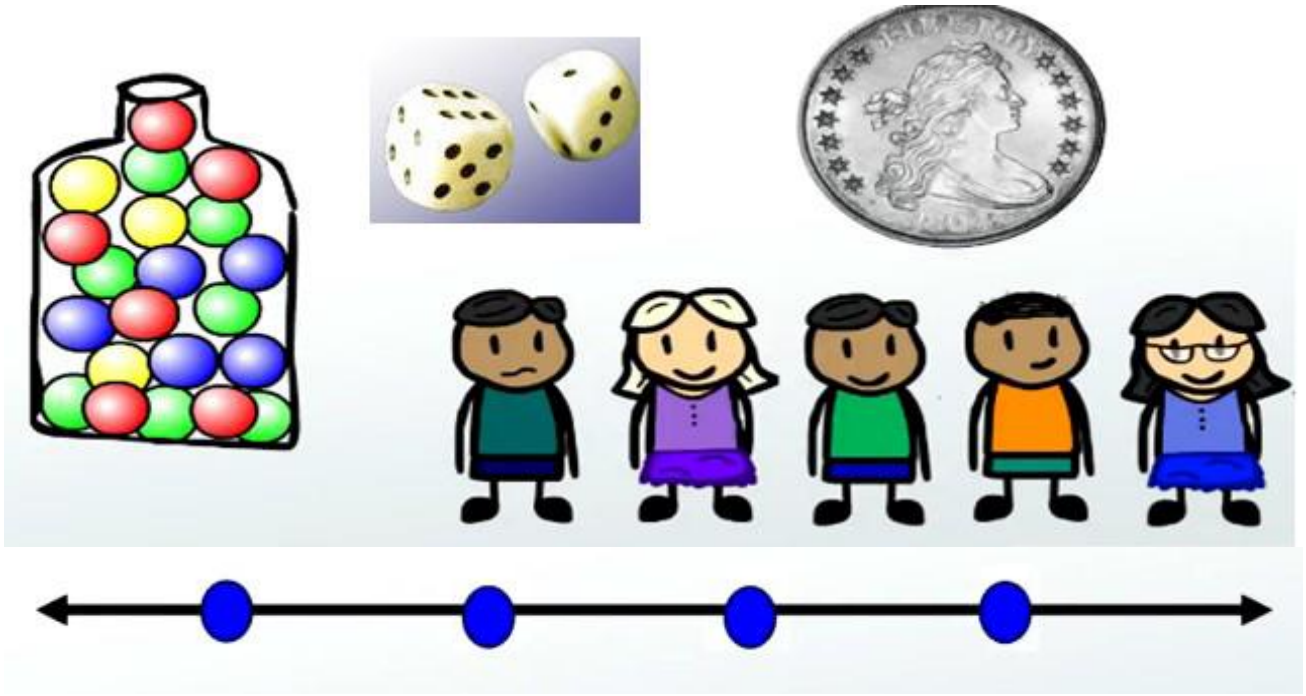
- A variable is an attribute that describes a person, place, thing, or idea
- The value of the variable can "vary" from one entity to another.
- A measurement describing some characteristic of a population.
- Eg: Age, Gender, business income, expenses, country of birth, capital expenditure, class grades, eye colour, vehicle type etc

Types of Variables

- **Quantitative variables**
- Numbers representing counts or measurements.
- Types: Discrete , Continuous
- **Discrete data** - data result when the number of possible values is either a finite number or a 'countable' number of possible values. 0, 1, 2, 3, . . .
- Example: The number of eggs that hens lay
- **Continuous data** - infinitely many possible values that correspond to some continuous scale that covers a range of values without gaps, interruptions, or jumps.
- Example: The amount of milk that a cow produces; e.g. 2.5 litres per day.

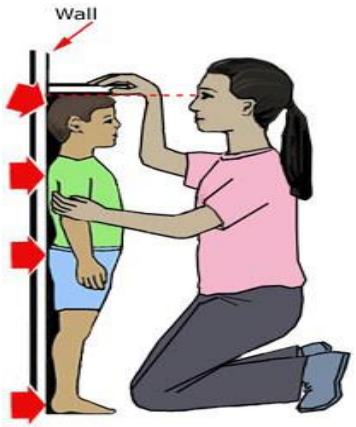
Discrete Data

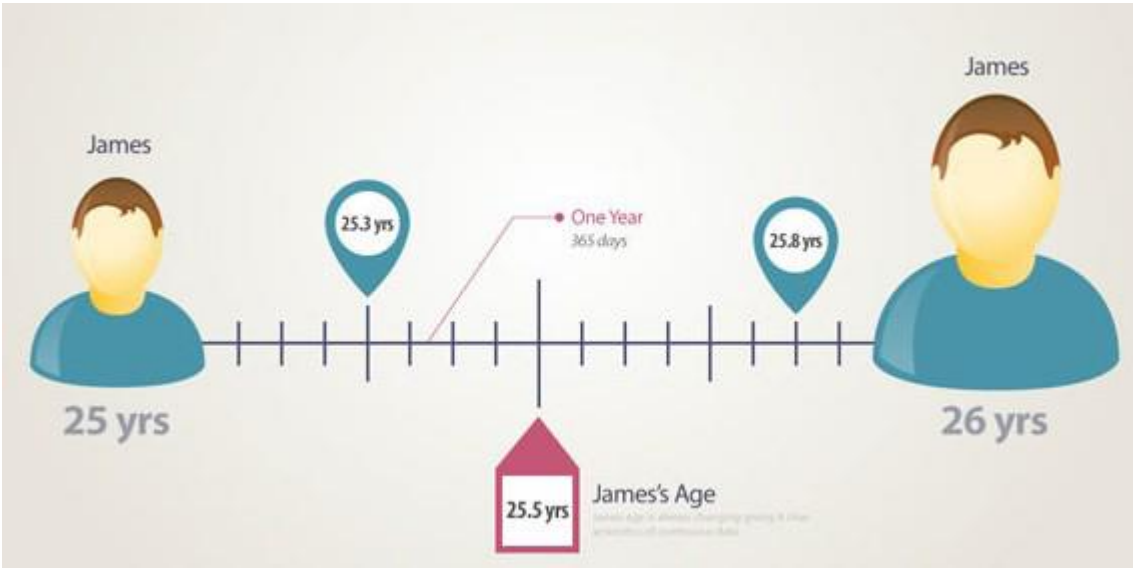
- Discrete Data can only take certain values
- Discrete Data – isolated points on the number line



Continuous Data

- Continuous Data can take any value (within a range).

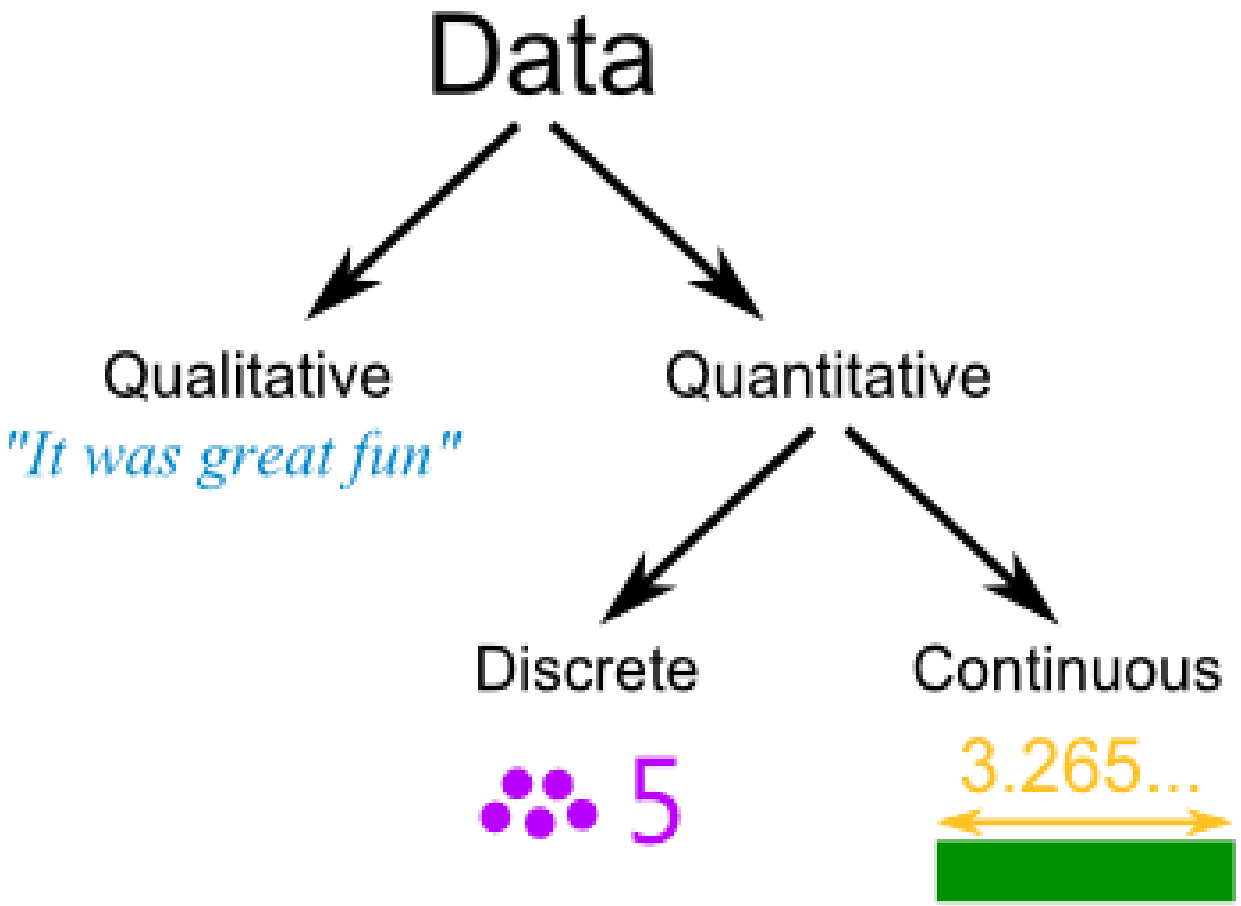




- **Qualitative Variable (or categorical)**
  - Can be separated into different categories that are distinguished by some nonnumeric characteristics.

Example: genders (male/female) of professional athletes.

Quantitative	Qualitative 
• 16 ounces of coffee • Temperature is 150°F • Cost $2.95 • Equals 50% of my daily coffee intake (16 oz. out of my 32 oz. daily coffee intake) • Cup is 6 inches tall	• Robust aroma • Columbian, fair-trade, organic • White cup and white lid • I like the taste, even though it's a little strong. • I'd recommend this coffee to a friend


Data
Qualitative
"It was great fun"
Quantitative
Discrete
5
Continuous
3.265...

Quiz

- Which of these is an example of a categorical variable?
  - Flavour of a soft drink ordered by each customer at a fast food restaurant
  - Height measured in inches , for each student in a class
  - Points scored by each player in a team

- The entire group of interest for a statistical conclusion is called the
  - Population
  - Sample
  - data

- A sub-group that is representative of a population is called
  - Sample
  - Category
  - data

- Statistical inference is
  - The process of estimates and conclusions carefully based on data from a sample
  - The process of estimates and conclusions carefully based on data from entire population
  - Pictorial display that summarizes data

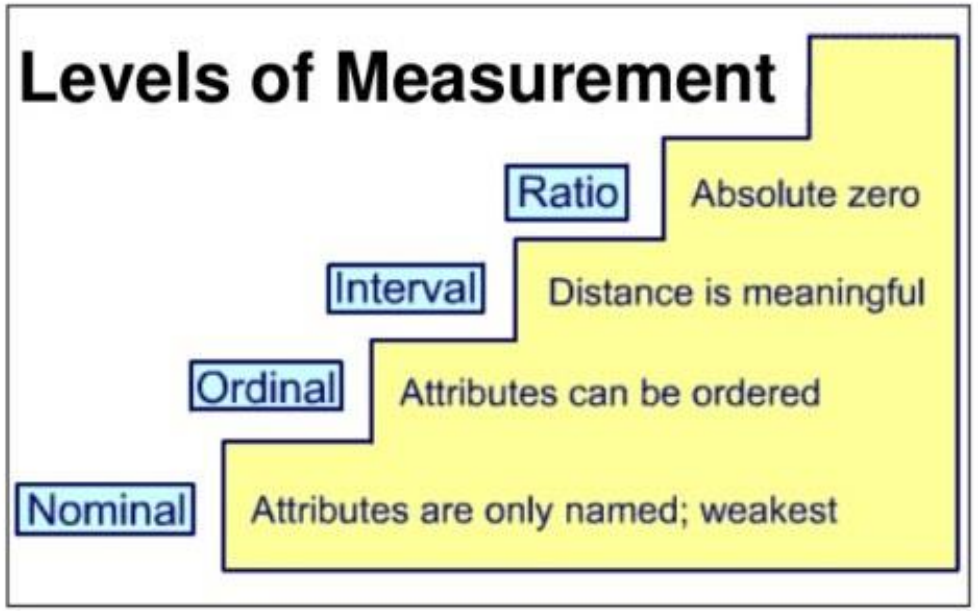
- Two types of statistical variables are
  - Categorical and descriptive
  - Categorical and numerical
  - Descriptive and numerical

Levels of Measurement

Another way to classify data is to use levels of measurement.

Four of these levels are

- Ordinal
- Nominal
- Interval
- Ratio



❖ nominal level of measurement

characterized by data that consist of names, labels, or categories only. The data cannot be arranged in an ordering scheme (such as low to high). the numbers in the variable are used only to classify the data. In this level of measurement, words, letters, and alpha-numeric symbols can be used.

Example: survey responses yes, no, undecided

❖ ordinal level of measurement

involves data that may be arranged in some order, but differences between data values either cannot be determined or are meaningless

Example: Course grades A, B, C, D, or F

❖ interval level of measurement

like the ordinal level, with the additional property that the difference between any two data values is meaningful. However, there is no natural zero starting point (where *none* of the quantity is present)

Example: Years 1000, 2000, 1776, and 1492

❖ ratio level of measurement

the interval level modified to include the natural zero starting point (where zero indicates that *none* of the quantity is present). For values at this level, differences and ratios are meaningful.

Example: Prices of college textbooks ($0 represents no cost)

	Level of measurement			
Type of descriptive statistic	**Nominal**	**Ordinal**	**Interval**	**Ratio**
Mode, counts, frequency	Yes	Yes	Yes	Yes
Median, minimum, maximum, range.	No	Yes	Yes	Yes
Mean, variance, standard deviation.	No	No	Yes	Yes

Summary - Levels of Measurement

- **Nominal** - categories only
- **Ordinal** - categories with some order
- **Interval** - differences but no natural starting point
- **Ratio** - differences and a natural starting point

Quiz – Identify variable based on value and level of measurement

- No. on back of players
- Grades of a student
- Rank of a student
- Temperature
- Age
- Rating of a movie
- Electrical current readings
- Calendar dates
- Street numbers
- Gender of employee

Quiz – Identify variable type and Level

- No. on back of players – Quantitative (Discrete), Nominal
- Grades of a student – Qualitative, Ordinal
- Rank of a student – Quantitative(Discrete), Ordinal
- Temperature –Quantitative(Continuous), Interval/Ratio
- Age- Quantitative(continuous), Interval
- Rating of a movie – Quantitative(discrete), Ordinal
- Electrical current readings – Quantitative(Discrete),Ratio
- Calendar dates – Quantitative(Discrete), Interval
- Street numbers – Quantitative (Discrete), Ordinal
- Gender of employee – Qualitative, Nominal

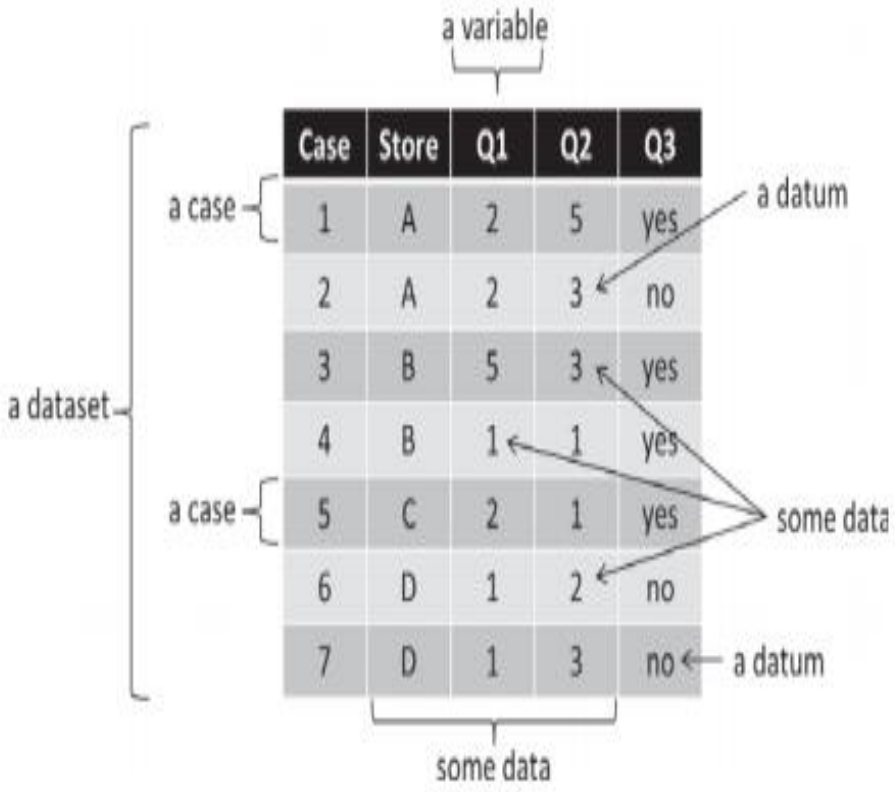
Confused????

Ask these Questions in the same order

		If You Answer...	
		Yes	No
1	Are the labels meaningful?	Go to Question 2.	There's no measurement!
2	Do the labels have an order?	Go to Question 3.	It's nominal.
3	Are the distances between values meaningful?	Go to Question 4.	It's ordinal.
4	Is there a meaningful zero?	It's ratio.	It's interval.

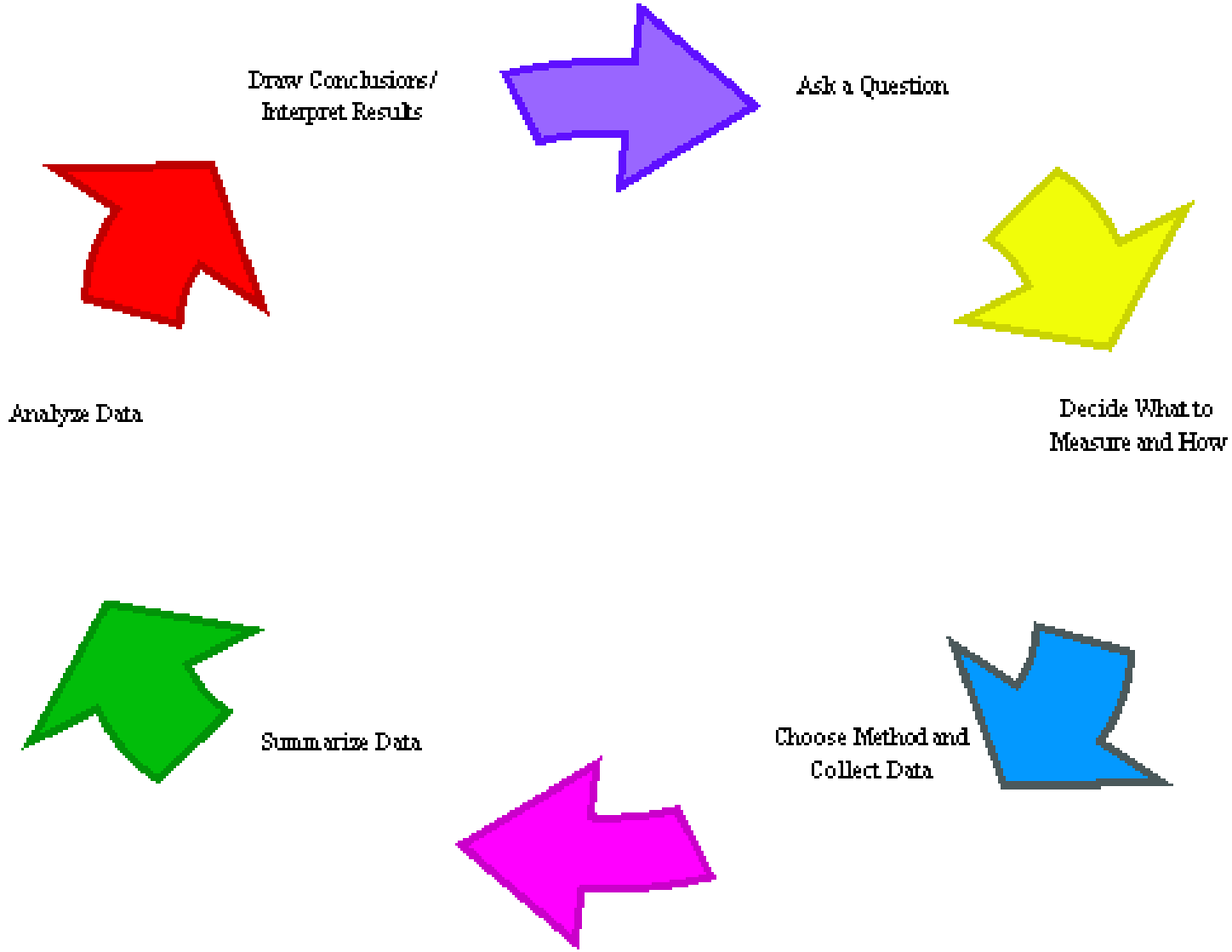
Organization of Data for Statistics

- The rows and columns can be in the form of
  - Spread sheets
  - Databases
  - Text files (Comma Separated Values)
  - Data Frames



Statistical Data Analysis

- Begin
- **Formulate the Research Problem**
- Define Population and Sample
- Collect the data
- **Do Descriptive Data Analysis**
- **Use Appropriate statistical methods to solve the research problem**
- **Report the results / Draw conclusions**
- End


Draw Conclusions/
Interpret Results
Ask a Question
Analyze Data
Decide What to
Measure and How
Summarize Data
Choose Method and
Collect Data

Univariate and Bivariate Analysis

- **Univariate**: During the study, if we consider only one variable then it is univariate analysis
  - Eg: Frequency distribution of gender
- **Bivariate :** Conduct a study that examines the relationship between two variables
  - Eg: Suppose we want to find the relationship between the height and weight of high school students.

Univariate Analysis -Qualitative data

- Frequency – Count of observations on a particular category
- Relative Frequency – $\frac{\text{Frequency in a class}}{\text{Total no of observations}}$
- Cumulative relative frequency – summation of all the relative frequency
- Visualization method
  - Bar graph
  - Pie chart

Find the answer

- A Process manufacturers produces bearings for combustion engine. The bearings thickness are between 1.486 and 1.490 mm are classified as conforming which means they meet the specification. Bearings thicker than this are reground and bearings thinner than this are scrapped. In a sample of 1000 bearings 910 are conforming 53 are reground and 37 are scrapped. Find the sample proportions.

- Suppose, of the 140 children, 20 lived in owner occupied houses, 70 lived in council houses and 50 lived in private rented accommodation. Find the relative frequency of each class and cumulative relative frequency percentage.

Univariate Analysis – Quantitative Data

- Measures of Center / Central Tendency
  - Mean, Median, Mode
- Measures of Variation /Dispersion /Spread
  - Range, Interquartile range, SD
- Data visualization
  - Histogram
  - Box plot

Measures of Central Tendency

Measures of Central Tendency:

A single number to serve as a representative value around which all the numbers in the set tend to cluster.

Sometimes it is referred to as a “middle” number of the data.

Three types of measures of central tendency:

- Mean (average)
- Median (middle)
- Mode (most)

Measures of Central Tendency

Mean

The **mean** (**arithmetic mean** or **average**) of a set of data is found by adding up all the items and then dividing by the sum of the number of items.

The mean of a sample is denoted by $\overline{x}$ (read "*x* bar").

The mean of a complete population is denoted by μ (the lower case Greek letter *mu*).

The **mean** of n data items $x_1, x_2, \ldots, x_n$, is given by the formula

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \qquad \text{or} \qquad \overline{x} = \frac{\sum x}{n}$$

Measures of Central Tendency

Example:

Ten students were polled as to the number of siblings in their individual families.

The raw data is the following set: {3, 2, 2, 1, 3, 6, 3, 3, 4, 2}.

Find the mean number of siblings for the ten students.

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

$$\bar{x} = \frac{3+2+2+1+3+6+3+3+4+2}{10}$$

$$\bar{x} = \frac{29}{10}$$

$$\bar{x} = 2.9 \text{ siblings}$$

Measures of Central Tendency

Mean

Practice:

7 participants in bike race had the following finishing times in minutes: 28,22,26,29,21,23,24.

What is the mean?

8 participants in bike race had the following finishing times in minutes: 28,22,26,29,21,23,24,50.

What is the mean?

Measures of Central Tendency

Weighted Mean

The **weighted mean** of n numbers $x_1, x_2, \ldots, x_n$, that are weighted by the respective factors $f_1, f_2, \ldots, f_n$ is given by the formula:

$$\bar{w} = \frac{\sum (x \cdot f)}{\sum f}.$$

Eg: your SGPA calculation is a weighted mean

Each subject is associated with credits which are the weights/factors and your score is the x value.

SGPA = (4*9+4*10)/Sum of weights/factors

Measures of Central Tendency

Trimmed Mean

A **trimmed mean** is a method of averaging that removes a small designated percentage of the largest and smallest values before calculating the **mean**.

It is measure of centre that is designed to be unaffected by outliers

1. Arrange sample values in order
2. Trim equal number of values from either end
3. Compute mean for the remaining

If p% of the data are trimmed from each end, the result is called p% trimmed mean. Commonly used p% are 5%, 10%, 20%

Measures of Central Tendency

Example:

Example: Find the trimmed 20% mean for the following test scores: 60, 81, 83, 91, 99.

Step 1: Trim the top and bottom 20% from the data. That leaves us with the middle three values:

60, 81, 83, 91, 99.

Step 2: Find the mean with the remaining values. The mean is (81 + 83 + 91) / 3) = 85.

Measures of Central Tendency

Median

Another measure of central tendency, is the **median**.

This measure divides a group of numbers into two parts, with half the numbers below the median and half above it.

The median is not as sensitive to extreme values as the mean.

To find the **median** of a group of items:

1. Rank the items.

2. If the number of items is ***odd***, the median is the middle item in the list. ((n+1)/2)

3. If the number of items is ***even***, the median is the mean of the two middle numbers.(Avg(n/2,n/2 +1))

Measures of Central Tendency

Median

Example:

Ten students in a math class were polled as to the number of siblings in their individual families and the results were:

3, 2, 2, 1, 1, 6, 3, 3, 4, 2.

Find the median number of siblings for the ten students.

Position of the median: 10/2 = 5

Between the 5^{th} and 6^{th} values

Data in order: 1, 1, 2, 2, 2, 3, 3, 3, 4, 6

Median = (2+3)/2 = 2.5 siblings

Measures of Central Tendency

Median

Example:

Nine students in a math class were polled as to the number of siblings in their individual families and the results were:

3, 2, 2, 1, 6, 3, 3, 4, 2.

Find the median number of siblings for the ten students.

Position of the median: (9+1)/2 = 5

The 5^{th} value

In order: 1, 2, 2, 2, 3, 3, 3, 4, 6

Median = 3 siblings

Measures of Central Tendency

Median

Practice:

7 participants in bike race had the following finishing times in minutes: 28,22,26,29,21,23,24.

What is the median?

8 participants in bike race had the following finishing times in minutes: 28,22,26,29,21,23,24,50.

What is the median?

Mean Vs Median

	Mean	Median
Definition	The mean is the arithmetic average of a set of numbers, or distribution.	The median is described as the numeric value separating the higher half and lower half of a sample, a population, or a probability distribution.
Applicability	The mean is used for normal distributions.	The median is generally used for skewed distributions.
Relevance to the data set	The mean is not a robust tool since it is largely influenced by outliers.	The median is better suited for skewed distributions to derive at central tendency since it is much more robust and sensible.

Measures of Central Tendency

Mode

The **mode** of a data set is the value that occurs the most often.

If a distribution has two modes, then it is called **bimodal**.

In a large distribution, this term is commonly applied even when the two modes do not have exactly the same frequency

Example:

Ten students in a math class were polled as to the number of siblings in their individual families and the results were: 3, 2, 2, 1, 3, 6, 3, 3, 4, 1. Find the mode for the number of siblings.

3, 2, 2, 1, 3, 6, 3, 3, 4, 1

The mode for the number of siblings is 3.

Measures of Central Tendency

Central Tendency from Stem-and-Leaf Displays

The mean can be calculated from the data presented in a Stem-and-Leaf display.

The median and mode are easily identified when the "leaves" are **ranked** (in numerical order) on their "stems."

Find the median and mode using stem and Leaf Displays for the following data.

15, 56,58, 28,29,29,36, 43,46,
37,40, 42,16,20,27,42,
51, 58, 42, 36,37

Median: (21+1)/2 = 11th term

The median is 37.

Mode is 42

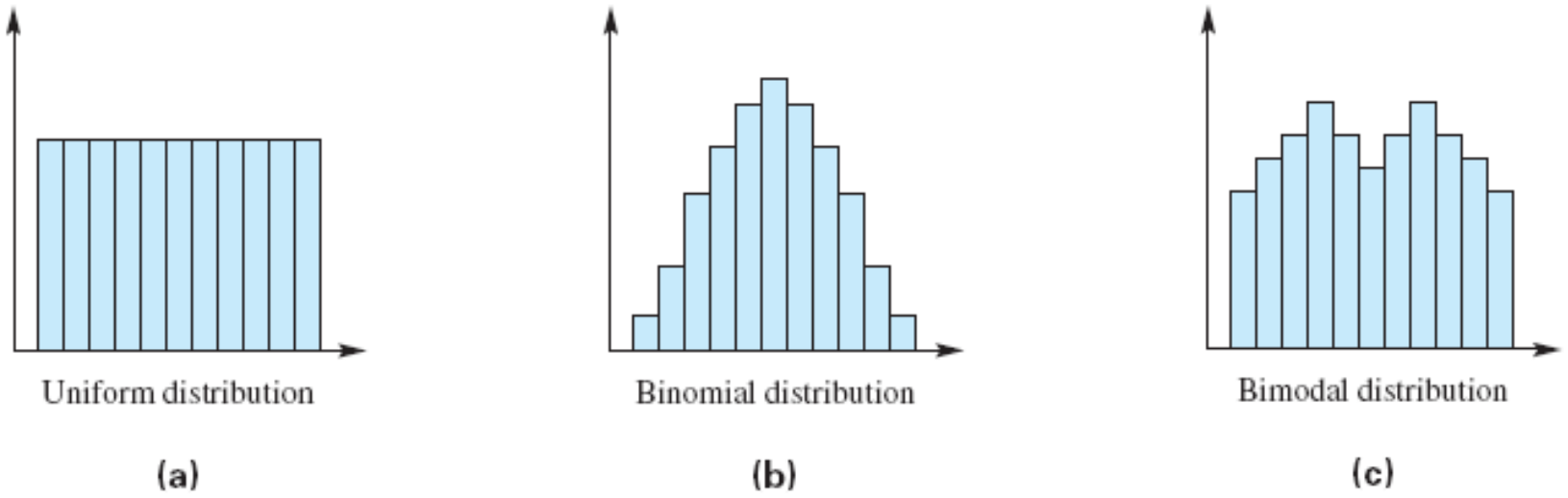
1	5	6				
2	0	7	8	9	9	
3	6	6	7	7		
4	0	2	2	2	3	6
5	1	6	8	8		

Measures of Central Tendency

Symmetry in Data Sets

The analysis of a data set often depends on whether the distribution is **symmetric** or **non-symmetric**.

Symmetric distribution: the pattern of frequencies from a central point is the same (or nearly so) from the left and right.



Some symmetric distributions

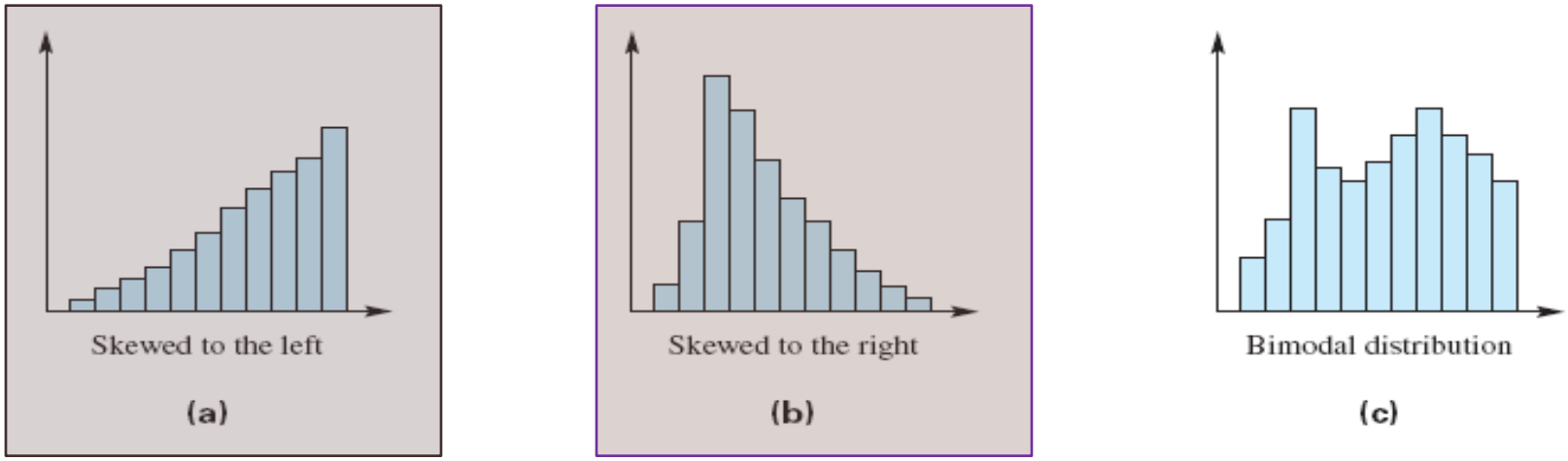
Measures of Central Tendency

Symmetry in Data Sets

Non-symmetric distribution: the patterns from a central point from the left and right are different.

Skewed to the left: a tail extends out to the left.

Skewed to the right: a tail extends out to the right.



Some non-symmetric distributions

Measures of Dispersion

Dispersion is another analytical method to study data.

A main use of dispersion is to compare the amounts of spread in two (or more) data sets.

A common technique in inferential statistics is to draw comparisons between populations by analyzing samples that come from those populations.

Two of the most common measures of dispersion are the *range* and the *standard deviation*.

Range

For any set of data, the **range** of the set is given by the following formula:

Range = (greatest value in set) – (least value in set).

Measures of Dispersion

Range

Example:

The two sets below have the same mean and median (7). Find the range of each set.

Set A	1	2	7	12	13
Set B	5	6	7	8	9

Range of Set A: $13 - 1 = 12$

Range = max - min

Range of Set A: $9 - 5 = 4$

Note: When range is more dispersion is more

Measures of Dispersion

Standard Deviation

One of the most useful measures of dispersion is the ***standard deviation*.**

It is based on ***deviations from the mean*** of the data.

Find the deviations from the mean for all data values of the sample 1, 2, 8, 11, 13.

The mean is 7.

To find each deviation, subtract the mean **from** each data value.

Data Value	1	2	8	11	13
Deviation	– 6	– 5	1	4	6

The sum of the deviations is always equal to zero.

Measures of Dispersion

Standard Deviation

Calculating the Sample Standard Deviation

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

The **sample standard deviation** is found by calculating the square root of the **variance**.

The **variance** is found by summing the squares of the deviations and dividing that sum by $n - 1$ (since it is a sample instead of a population).

The sample standard deviation is denoted by the letter *s*.

The standard deviation of a population is denoted by σ.

Measures of Dispersion

Standard Deviation

Calculating the Sample Standard Deviation

1. Calculate the mean of the numbers.

2. Find the deviations from the mean.

3. Square each deviation.

4. Sum the squared deviations.

5. Divide the sum in Step 4 by $n - 1$.

6. Take the square root of the quotient in Step 5.

Measures of Dispersion

Standard Deviation

Calculating the Sample Standard Deviation

Example:

Find the standard deviation of the sample set {1, 2, 8, 11, 13}.

$\bar{x} = (1+2+8+11+13)/5 \quad = 7$

Data Value	1	2	8	11	13
Deviation	– 6	– 5	1	4	6
$(\text{Deviation})^2$	36	25	1	16	36

Sum of the $(\text{Deviations})^2$ = 36 + 25 + 1 + 16 + 36 = 114

Measures of Dispersion

Standard Deviation

Calculating the Sample Standard Deviation

Sum of the (Deviations)^2 = 36 + 25 + 1 + 16 + 36 = 114

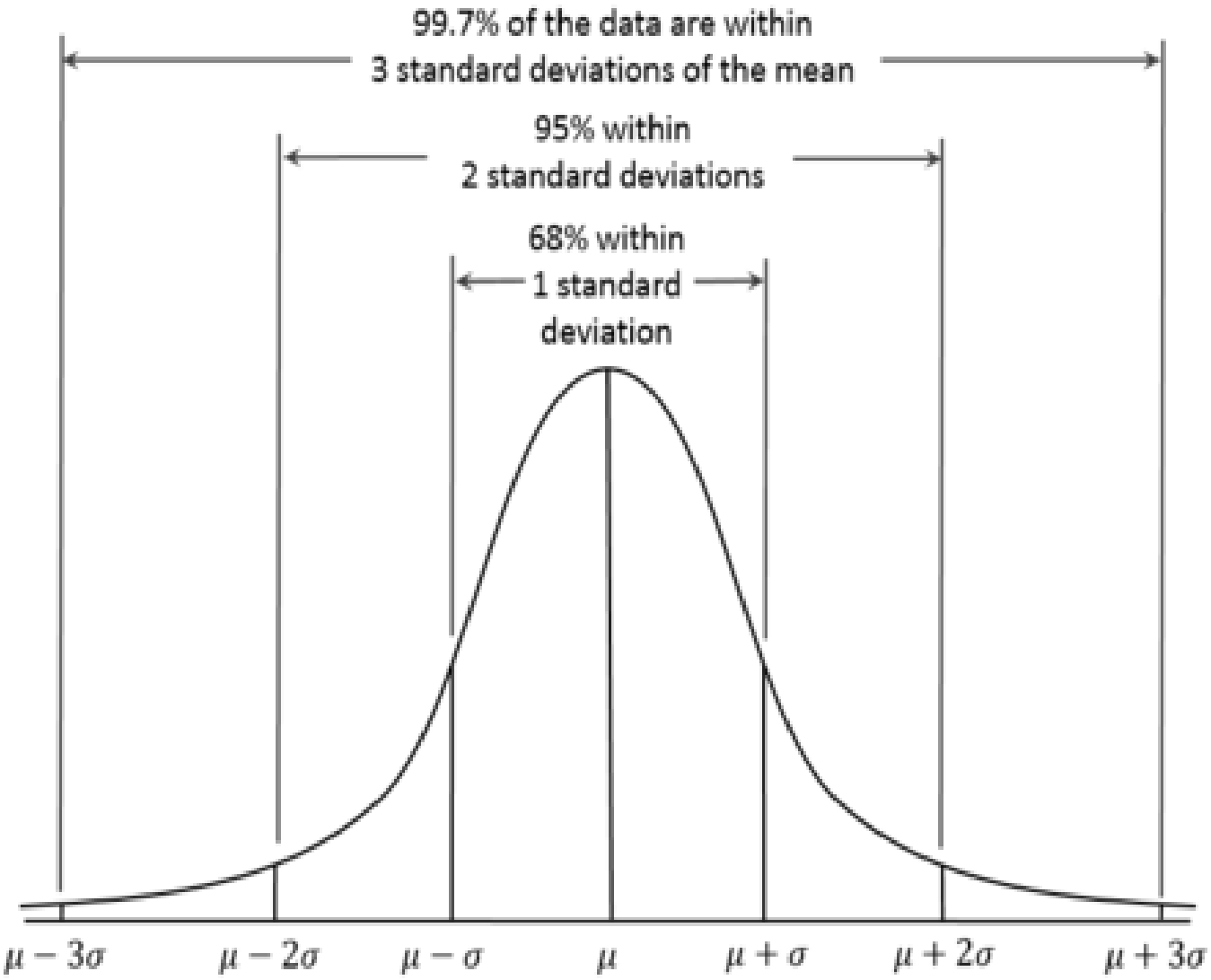
Divide 114 by $n - 1$ with $n = 5$:

$$\frac{114}{5 - 1} = 28.5$$

Take the square root of 28.5:

$$\sqrt{28.5} = 5.34$$

The sample standard deviation of the data is 5.34.


99.7% of the data are within
3 standard deviations of the mean
95% within
2 standard deviations
68% within
1 standard
deviation
$\mu - 3\sigma$
$\mu - 2\sigma$
$\mu - \sigma$
μ
$\mu + \sigma$
$\mu + 2\sigma$
$\mu + 3\sigma$

Measures of Dispersion

Standard Deviation

Example: Interpreting Measures

Two companies, *A* and *B*, sell small packs of sugar for coffee. The mean and standard deviation for samples from each company are given below. Which company consistently provides more sugar in their packs? Which company fills its packs more consistently?

Company *A*	Company *B*
$\bar{x}_A = 1.013$ tsp	$\bar{x}_B = 1.007$ tsp
$s_A = .0021$	$s_B = .0018$

Measures of Dispersion

Standard Deviation

Example: Interpreting Measures

Company *A*	Company *B*
$\bar{x}_A = 1.013$ tsp	$\bar{x}_B = 1.007$ tsp
$s_A = .0021$	$s_B = .0018$

Which company consistently provides more sugar in their packs?

The sample mean for Company A is greater than the sample mean of Company B.

The inference can be made that Company A provides more sugar in their packs.

Measures of Dispersion

Standard Deviation

Example: Interpreting Measures

Company A	Company B
$\bar{x}_A = 1.013$ tsp	$\bar{x}_B = 1.007$ tsp
$s_A = .0021$	$s_B = .0018$

Which company fills its packs more consistently?

The standard deviation for Company B is less than the standard deviation for Company A.

The inference can be made that Company B fills their packs more closer to their mean than Company A.

Measures of Dispersion

Variance

Variance is the square of the standard deviation or

1. Work out the Mean (the simple average of the numbers)
2. Then for each number: subtract the Mean and square the result (the squared difference).
3. Then work out the average of those squared differences

Data Value	1	2	8	11	13
Deviation	-6	-5	1	4	6
	36	25	1	16	36

Sum of the $(\text{Deviations})^2 = 36 + 25 + 1 + 16 + 36 = 114$

Average of the $(\text{Deviations})^2 = 114/4 = 28.5$

The Normal Distribution

Definition and Properties of a Normal Curve

A **normal curve** is a symmetric, bell-shaped curve.

Any random continuous variable whose graph has a bell shape is said to have a **normal distribution**.

On a **normal curve** the horizontal axis is labeled with the mean and the specific data values of the standard deviations.

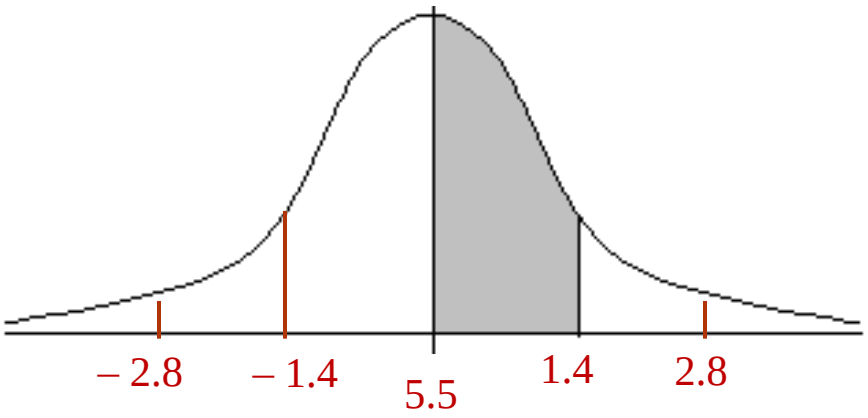
If the horizontal axis is labeled using the number of standard deviations from the mean, rather than the specific data values, then the curve the **standard normal curve**

The Normal Distribution

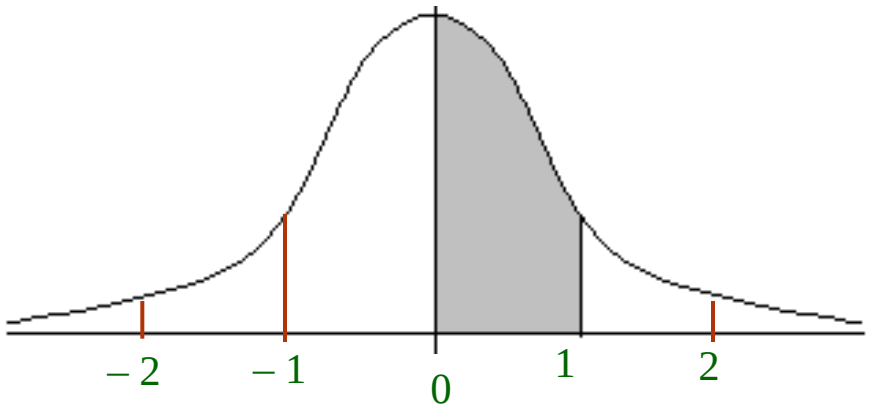
Sample Statistics

$\bar{x} = 5.5, s = 1.4$

Normal Curve

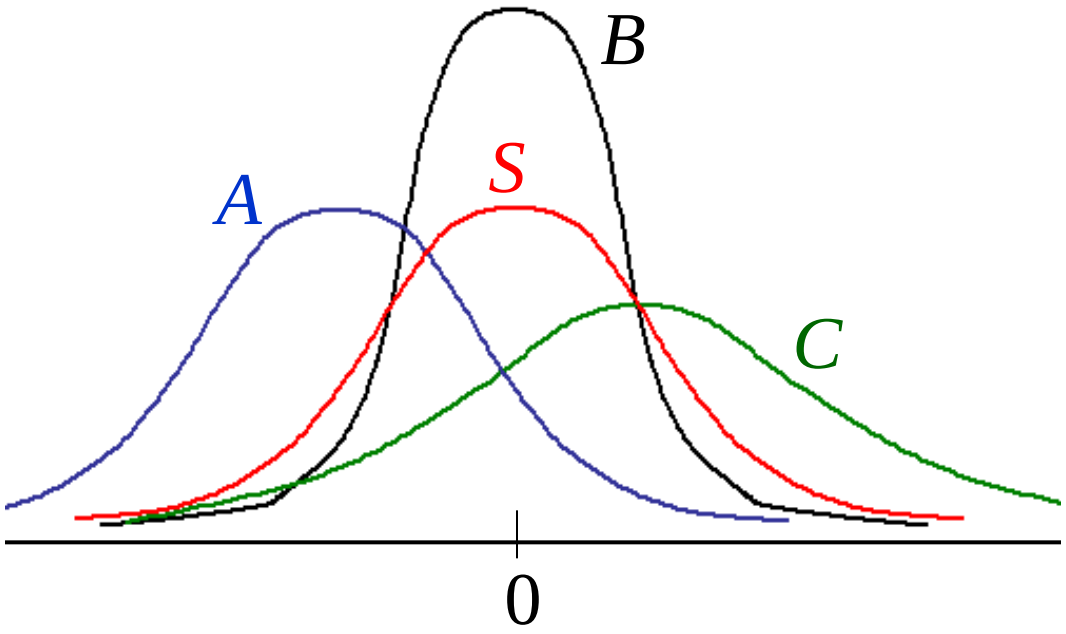


Standard Normal Curve



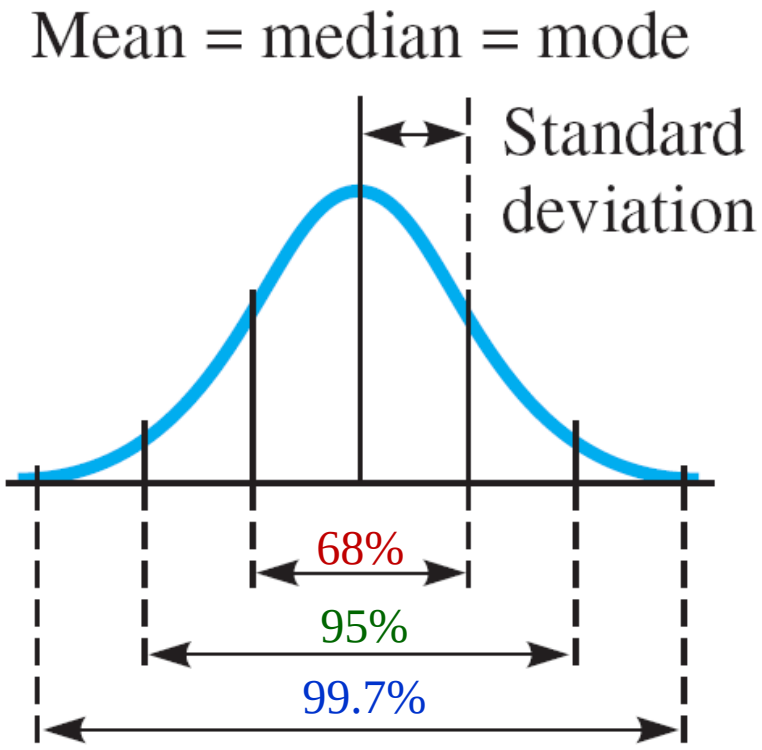
The Normal Distribution

Normal Curves



The Normal Distribution

Empirical Rule



Measures of Dispersion

Quartiles

For any set of data (ranked in order from least to greatest):

The **second quartile, Q_2 (50%)** is the median.

The **first quartile, Q_1 (25%)** is the median of items below Q_2.

The **third quartile, Q_3 (75%)** is the median of items above Q_2.

Measures of Dispersion

Quartiles

Example: Quartiles

The following are test scores (out of 100) for a particular math class.

44	56	58	62	64	64	70	72	72	72
74	74	75	78	78	79	80	82	82	84
86	87	88	90	92	95	96	96	98	100

Find the three quartiles.

N=30 -> even elements
Avg(n/2,n/2 +1)
Avg(15,16)th position

The average of the 15th and 16th items represents the 2nd quartile (Q_2) or the median = 78.5

50% of the scores were below 78.5.

Measures of Dispersion

Quartiles

Example: Quartiles

The following are test scores (out of 100) for a particular math class.

44	56	58	62	64	64	70	72	72	72
74	74	75	78	78	79	80	82	82	84
86	87	88	90	92	95	96	96	98	100

Find the three quartiles.

Q_1= 25%

25% = 0.25

0.25(30)

7.5

The 8th item represents the 1st quartile (Q_1)

25% of the scores were below 72.

Measures of Dispersion

Quartiles

Example: Quartiles

The following are test scores (out of 100) for a particular math class.

44	56	58	62	64	64	70	72	72	72
74	74	75	78	78	79	80	82	82	84
86	87	88	90	92	95	96	96	98	100

Find the three quartiles.

Q_3= 75%

75% = 0.75

0.75(30)

22.5

The 23rd item represents the 3rd quartile (Q_3)

75% of the scores were below 88.

Measures of Dispersion

Box Plots

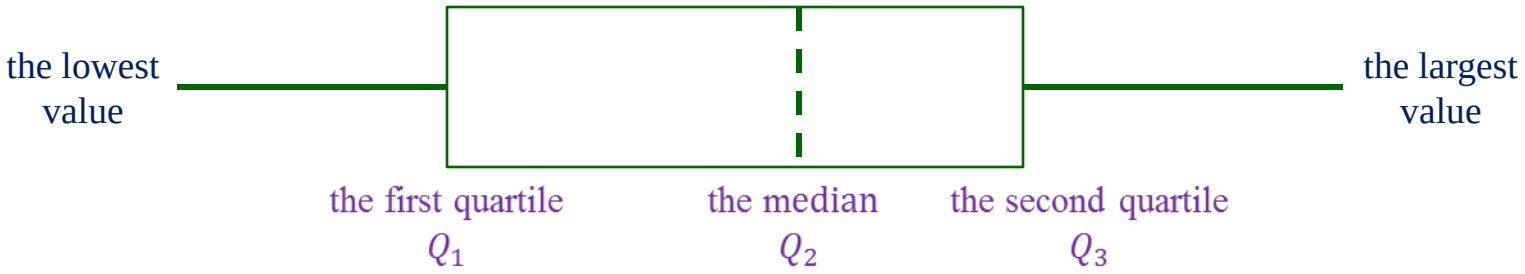
A box plot or a box and whisker plot is a visual display of five statistical measures.

The five statistical measures are:

the lowest value,

the first quartile, the median, the third quartile,

the largest value.



The Interquartile Range: $IR = Q_3 - Q_1$

Measures of Dispersion

Box Plots

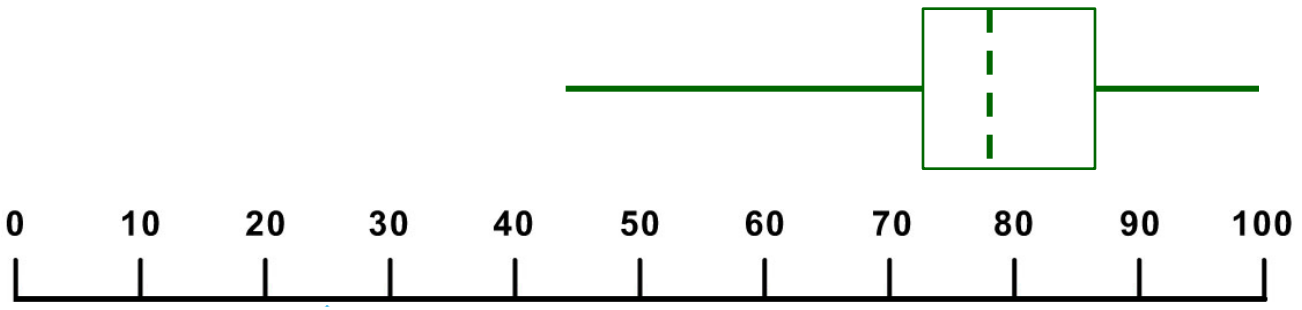
Example:

The following are test scores (out of 100) for a particular math class.

44	56	58	62	64	64	70	72	72	72
74	74	75	78	78	79	80	82	82	84
86	87	88	90	92	95	96	96	98	100

Q_1= 25% = 72 Q_2= 50% = median= 78.5 Q_3= 75%= 88

Lowest = 44 Largest = 100



The Interquartile Range: $IR = Q_3 - Q_1 = 88 - 72 = 16$

Practice Problems

Suppose we have the following data:

2, 3, 5, 6, 7, 9, 9, 11, 12, 15

1. What is the mean of these data?
2. What is the median?
3. Which is the mode?
4. What is the Standard Deviation?
5. What is the first quartile?
6. What is the third quartile?
7. What is the interquartile Range?

Suppose we have the following data:

2, 3, 5, 6, 7, 9, 9, 11, 12, 15

1. Add 1 to each item in the data and answer the following:

a) What is the median?

b) Which is the mode?

c) What is the Standard Deviation?

d) What is the first quartile?

e) What is the third quartile?

f) What is the interquartile Range?

Suppose we have the following data:

2, 3, 5, 6, 7, 9, 9, 11, 12, 15

2. Multiply each item by 2 in the data and answer the following:

a) What is the median?

b) Which is the mode?

c) What is the Standard Deviation?

d) What is the first quartile?

e) What is the third quartile?

f) What is the interquartile Range?

Points to remember

1) If a constant is added/ subtracted from each item in the sample then :

a) Mean increases or decreases by the same constant.

b) Median increases or decreases by the same constant.

c) **Standard Deviation remains the same.**

d) Q1 increases or decreases by the same constant.

e) Q3 increases or decreases by the same constant.

f) **IQR remains the same.**

Points to remember

2) If a constant is multiplied/divided to/from each item in the sample then :

a) Mean multiplied or divided by the same constant.

b) Median multiplied or divided by the same constant.

c) Standard Deviation multiplied or divided by the same constant.

d) Q1 multiplied or divided by the same constant.

e) Q3 multiplied or divided by the same constant.

f) IQR multiplied or divided by the same constant.

Quiz – Say True/False

1. For any list of numbers, half of them will be below the mean?

2. Is the sample mean always the most frequently occurring value?

3. Is the sample mean always equal to one of the values in the sample?

4. Is it possible for standard deviation of a list of numbers to be equal to 0?

5. Is it possible for standard deviation to be greater than mean?

Answer

1. For any list of numbers, half of them will be below the mean? **(False)**

2.Is the sample mean always the most frequently occurring value? **(False)**

3. Is the sample mean always equal to one of the values in the sample? **(False)**

4. Is it possible for standard deviation of a list of numbers to be equal to 0? **(True)**

5. Is it possible for standard deviation to be greater than mean? **(True)**

Question

A vendor converts the weights on the packages from pounds to kilograms

1 kg = 2.2 lb

a) How does this affect the mean weight of the packages?

b) How does this affect the standard deviation of the weights?

Answer

A vendor converts the weights on the packages from pounds to kilograms

1 kg = 2.2 lb

a) How does this affect the mean weight of the packages?

The mean will be divided by 2.2.

b) How does this affect the standard deviation of the weights?

The standard deviation will be divided by 2.2.

Question

The vendor begins using heavier packaging, which increases the weight of each package by 50g.

a) How does this affect the mean weight of the packages?

b) How does this affect the standard deviation of the weights?

The vendor begins using heavier packaging, which increases the weight of each package by 50g.

a) How does this affect the mean weight of the packages?

The mean will increase by 50 g.

b) How does this affect the standard deviation of the weights?

The standard deviation will be unchanged.

IQR Rule for outliers

- Arrange the elements in order
- Calculate first quartile (Q1), third quartile (Q3) and the interquartile range (IQR=Q3-Q1)
- Compute Q1–1.5 × IQR and Compute Q3+1.5 × IQR Anything outside this range is an outlier.

Z-Scores and Conversions

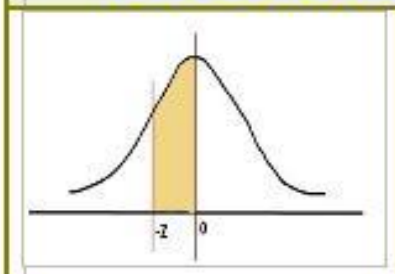
What is a Z Score?

A measure of an observation's distance from the mean.

The distance is measured in standard deviation units.

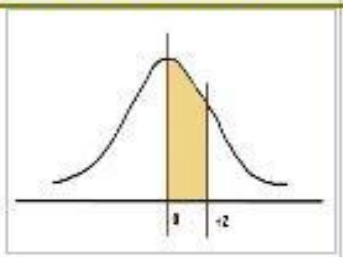
- If a z-score is zero, it's on the mean.
- If a z-score is positive, it's above the mean.
- If a z-score is negative, it's below the mean.
- If a z-score is 1, it's 1 SD above the mean.
- If a z-score is –2, it's 2 SDs below the mean.

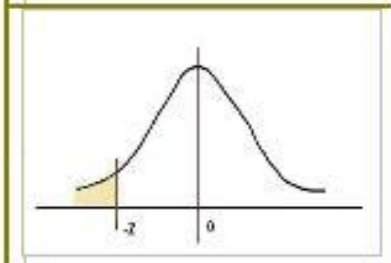
Procedure to find the Probability using Positive Z-score table



Case 1:
Area between 0 and any Z score

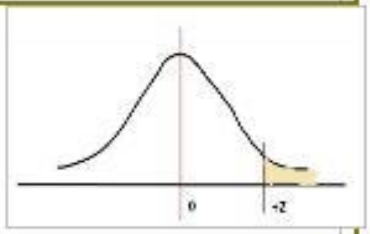
The area reading for the z value in the z-score table is the required probability.

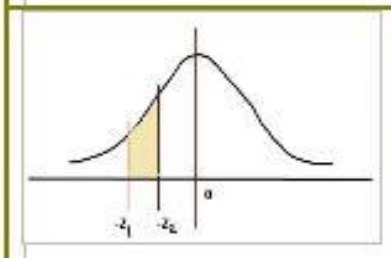




Case 2.
Area in any tail

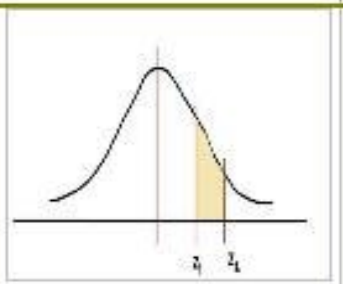
Find the area given for the z value in the z-score table. Subtract the area found from 0.5000 to find the probability.





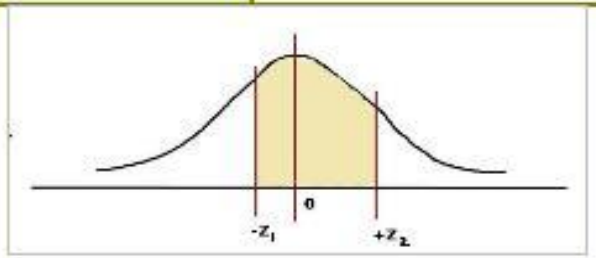
Case 3.
Area between two Z scores on the same side of the mean

Find the areas given for both the z values in the table Subtract the smaller area from the larger area to get the probability.



Case 4.
The area between two z values on opposite sides of the mean.

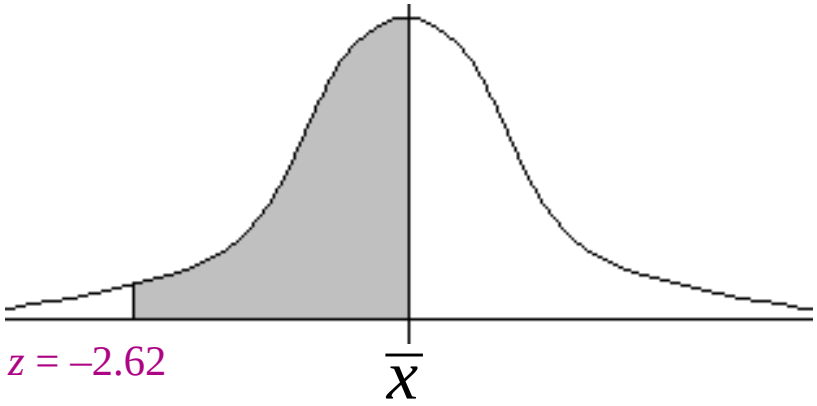
Find the areas given for both the z values given in the table.
Add the two areas found from the z-score table to arrive on the probability.



The Normal Distribution

Example: Applying the Normal Curve Table

Use the table to find the percent of all scores that lie between the mean and 2.62 standard deviations below the mean.



$z = -2.62$ Find 2.62 in the z column. The table entry is 0.4956

Therefore, 49.56% of all values lie between the mean and 2.62 standard deviations below the mean.

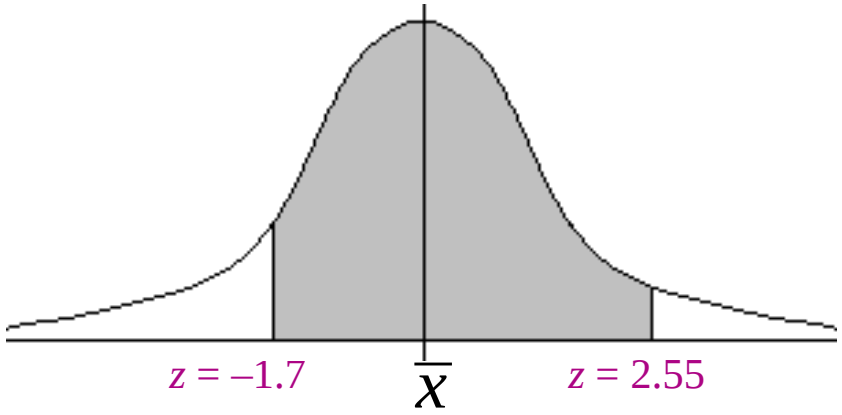
or

There is a 0.4956 probability that a randomly selected value will lie between the mean and 2.62 standard deviations below the mean.

The Normal Distribution

Example: Applying the Normal Curve Table

Find the percent of all scores that lie between the given z-scores.



$z = -1.7$ The table entry is 0.4554

$z = 2.55$ The table entry is 0.4946

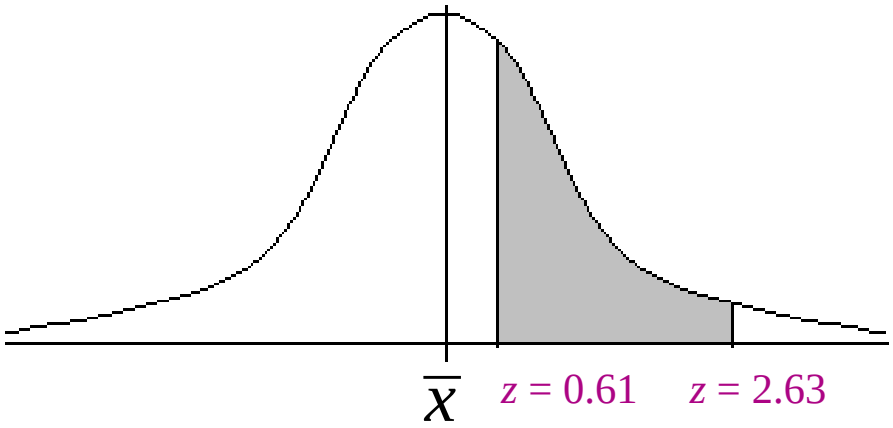
$0.4554 + 0.4946 = 0.95$

Therefore, 95% of all values lie between – 1.7 and 2.55 standard deviations.

The Normal Distribution

Example: Applying the Normal Curve Table

Find the probability that a randomly selected value will lie between the given z-scores.



$z = 0.61$ The table entry is 0.2291

$z = 2.63$ The table entry is 0.4957

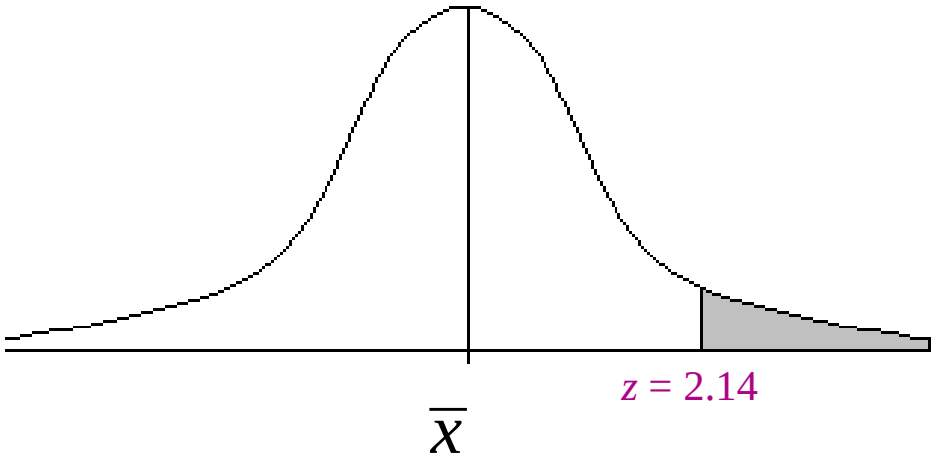
$0.4957 - 0.2291 = 0.2666$

There is a 0.2666 probability that a randomly selected value will lie between 0.61 and 2.63 standard deviations.

The Normal Distribution

Example: Applying the Normal Curve Table

Find the probability that a randomly selected value will lie above the given z-score.



$z = 2.14$ The table entry is 0.4838

Half of the area under the curve is 0.5000

$0.5000 - 0.4838 = 0.0162$

There is a 0.0162 probability that a randomly selected value will lie 2.14 standard deviations.

Practice

- The test scores of students in a class test has a mean of 70 and with a standard deviation of 12. What is the probable percentage of students scored more than 85?

- An organization made a survey on the monthly salary of their clerical level employees, in dollars. The data revealed the mean as 4000 with a standard deviation of $600. Find what percentage of employees are in the salary bracket [3000, 4500].