CSE 2027-Fundamental of Data Analysis

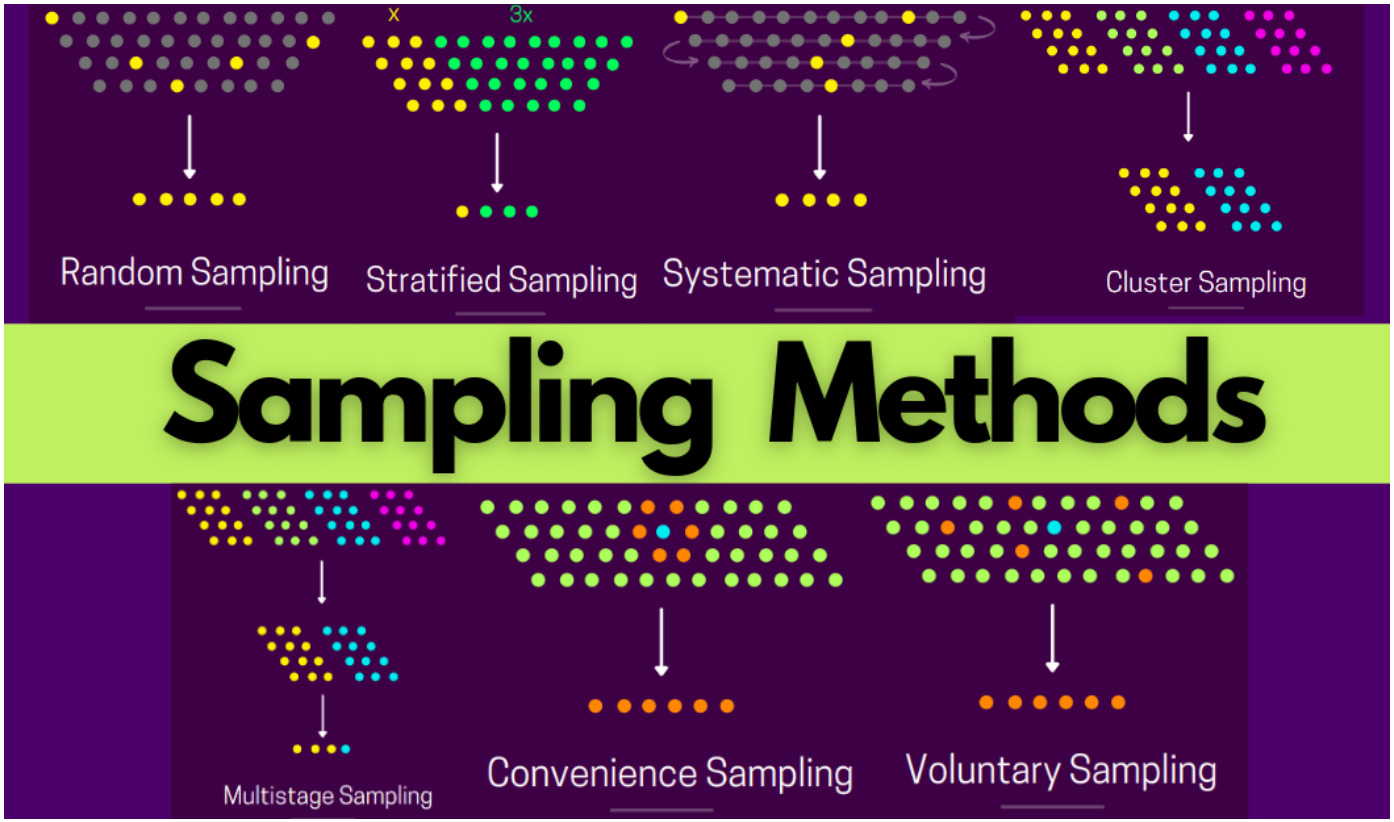
Module: 2: Statistical functions

Sampling Techniques: Fundamental Definitions, Important sampling distributions concept of standard error, Descriptive Statistics, Inferential Statistics (T test, Z test), Probability Uses In Business and Calculating Probability from a Contingency Tables.

Definition

- **Sampling** is the process of selecting a subset*(a predetermined number of observations)* from a larger population. It's a pretty common technique where in, we run experiments and draw conclusions about the population, without the need of having to study the entire population. We will go through two types of sampling methods:
- ***Probability Sampling*** *—Here we choose a sample based on the theory of probability.*
- ***Non-Probability Sampling*** *— Here we choose a sample based on non-random criteria, and not every member of the population has a chance of being included.*

Sampling techniques



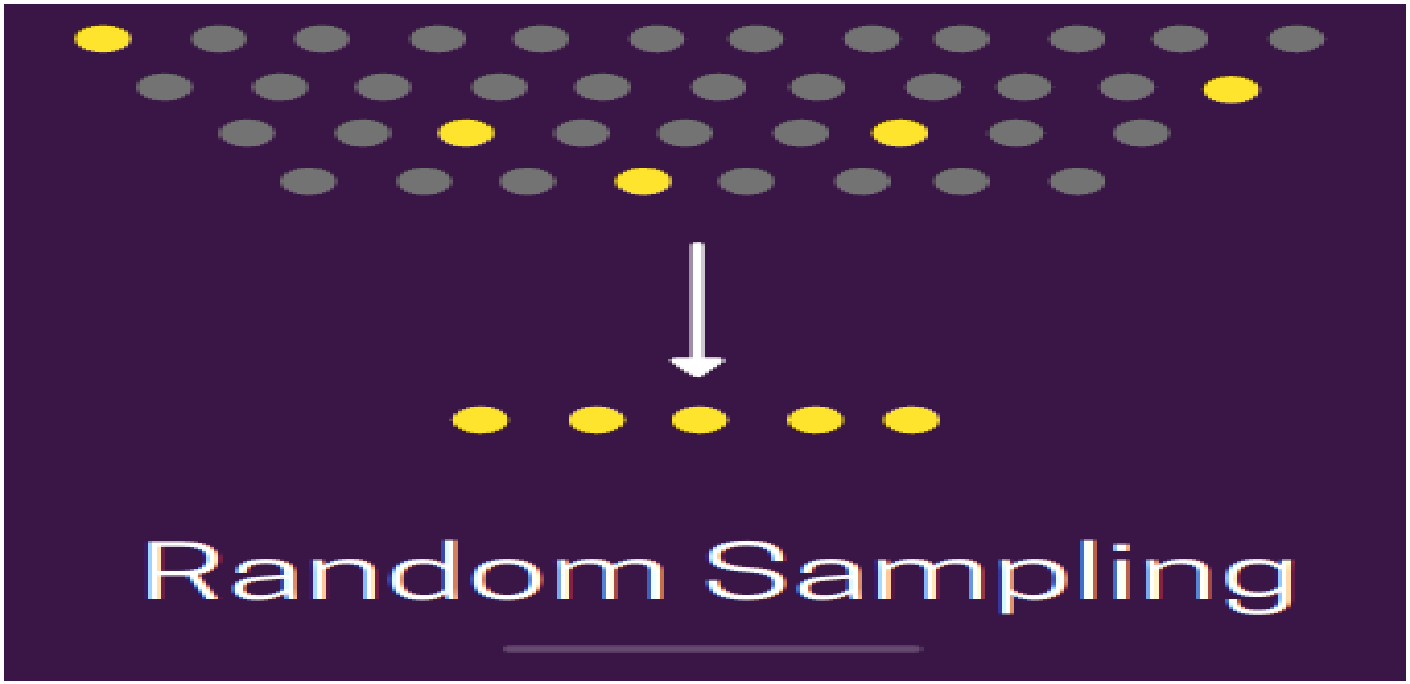
Sampling types

- Probability Sampling
  - Random Sampling
  - Stratified Sampling
  - Cluster Sampling
  - Systematic Sampling
  - Multistage Sampling
- Non –Probability Sampling
  - Convenience Sampling
  - Voluntary Sampling
  - Snowball Sampling

Random Sampling

- Under Random sampling, every element of the population has an equal probability of getting selected. *Below fig. shows the pictorial view of the same — All the points collectively represent the entire population wherein every point has an equal chance of getting selected.*

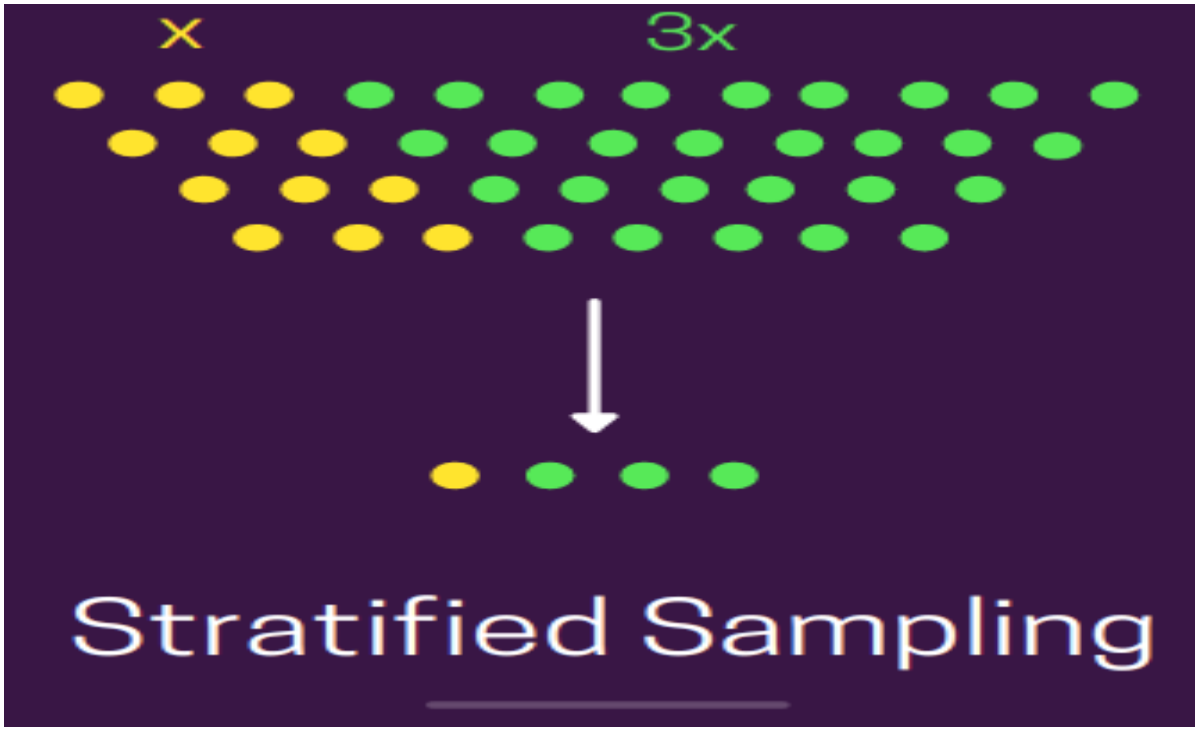
Random Sampling



Stratified Sampling

- Under stratified sampling, we **group the entire population into subpopulations** by some common property. *For example — Class labels in a typical ML classification task.* We then randomly sample from those groups individually, such that the **groups are still maintained in the same ratio** as they were in the entire population. *Below fig. shows a pictorial view of the same — We have two groups with a count ratio of x and 3x based on the color, we randomly sample from yellow and green sets separately and represent the final set in the same ratio of these groups.*

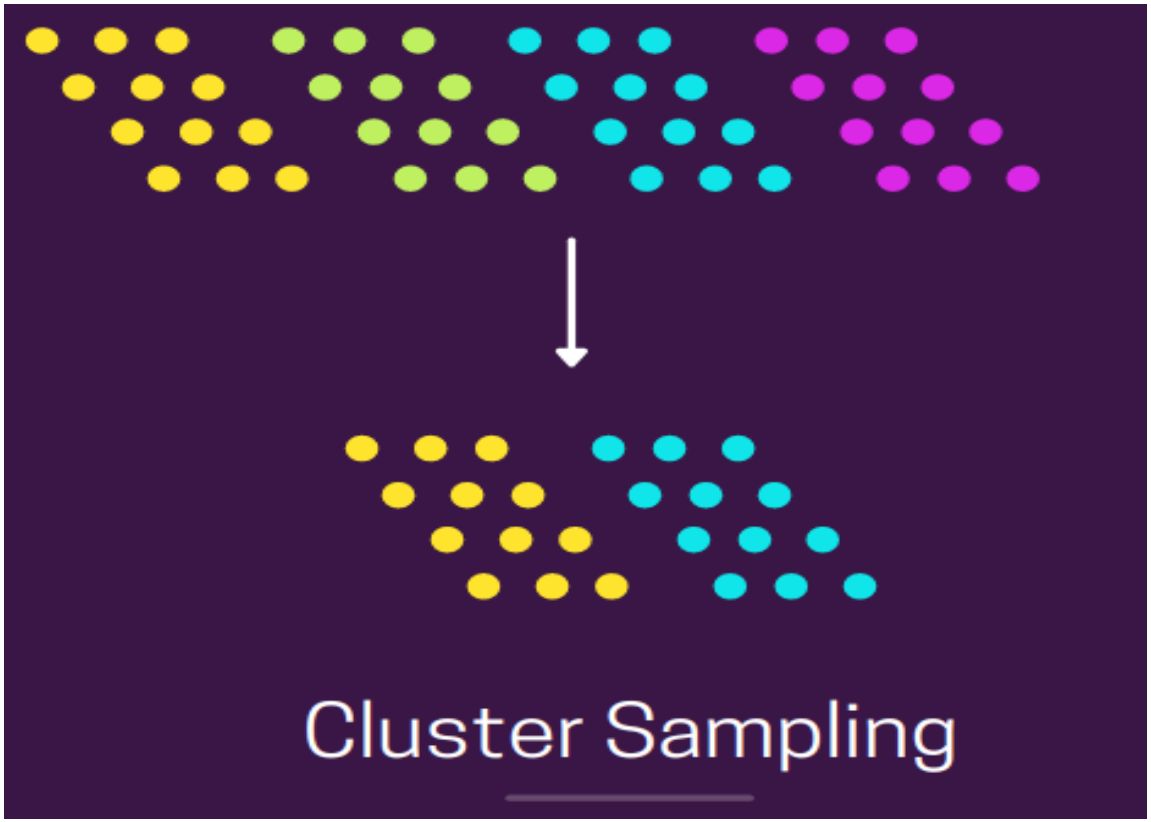
Stratified Sampling



Cluster Sampling

- In Cluster sampling, we **divide the entire population into subgroups**, wherein, each of those subgroups has similar characteristics to that of the population when considered in totality. Also, instead of sampling individuals, we **randomly select the entire subgroups**. *As can be seen in the below fig. that we had 4 clusters with similar properties (size and shape), we randomly select two clusters and treat them as samples*.

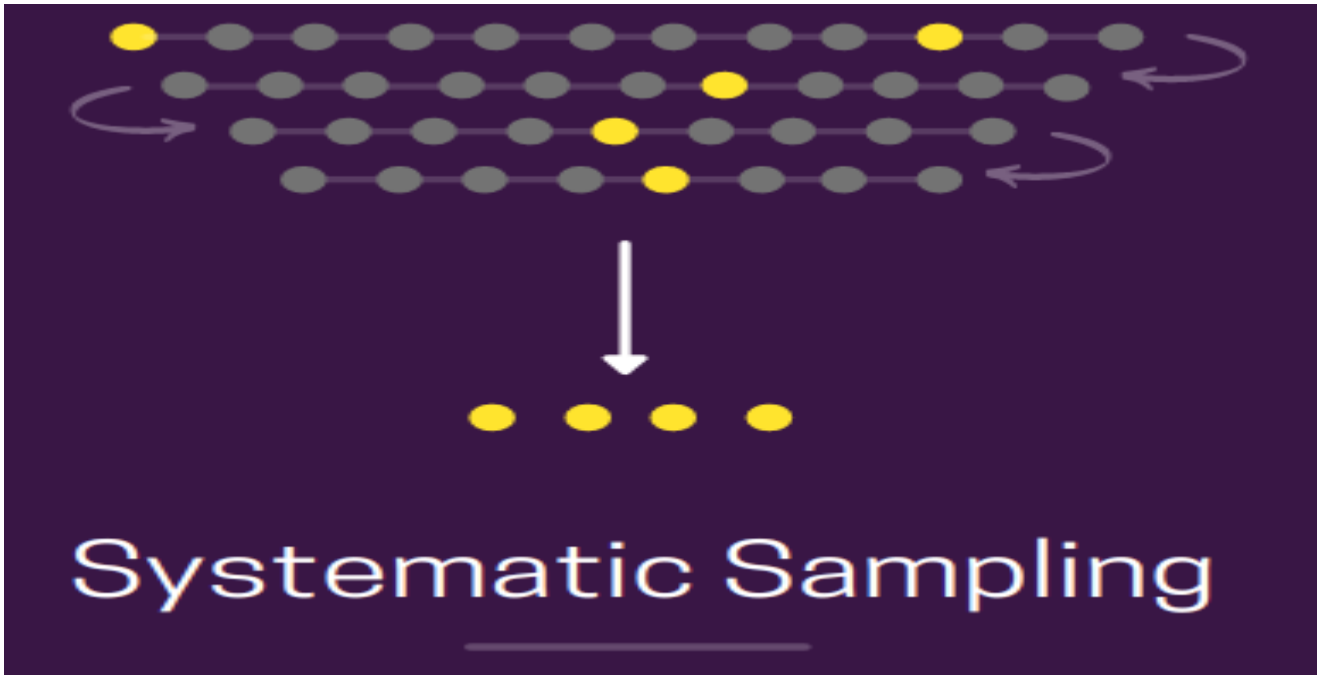
Cluster Sampling



Systematic Sampling

- Systematic sampling is about sampling items from the population at **regular predefined intervals***(basically fixed and periodic intervals)*. *For example — Every 5th element, 21st element and so on.* This sampling method tends to be more effective than the vanilla random sampling method in general. *Below fig. shows a pictorial view of the same — We sample every 9th and 7th element in order and then repeat this pattern.*

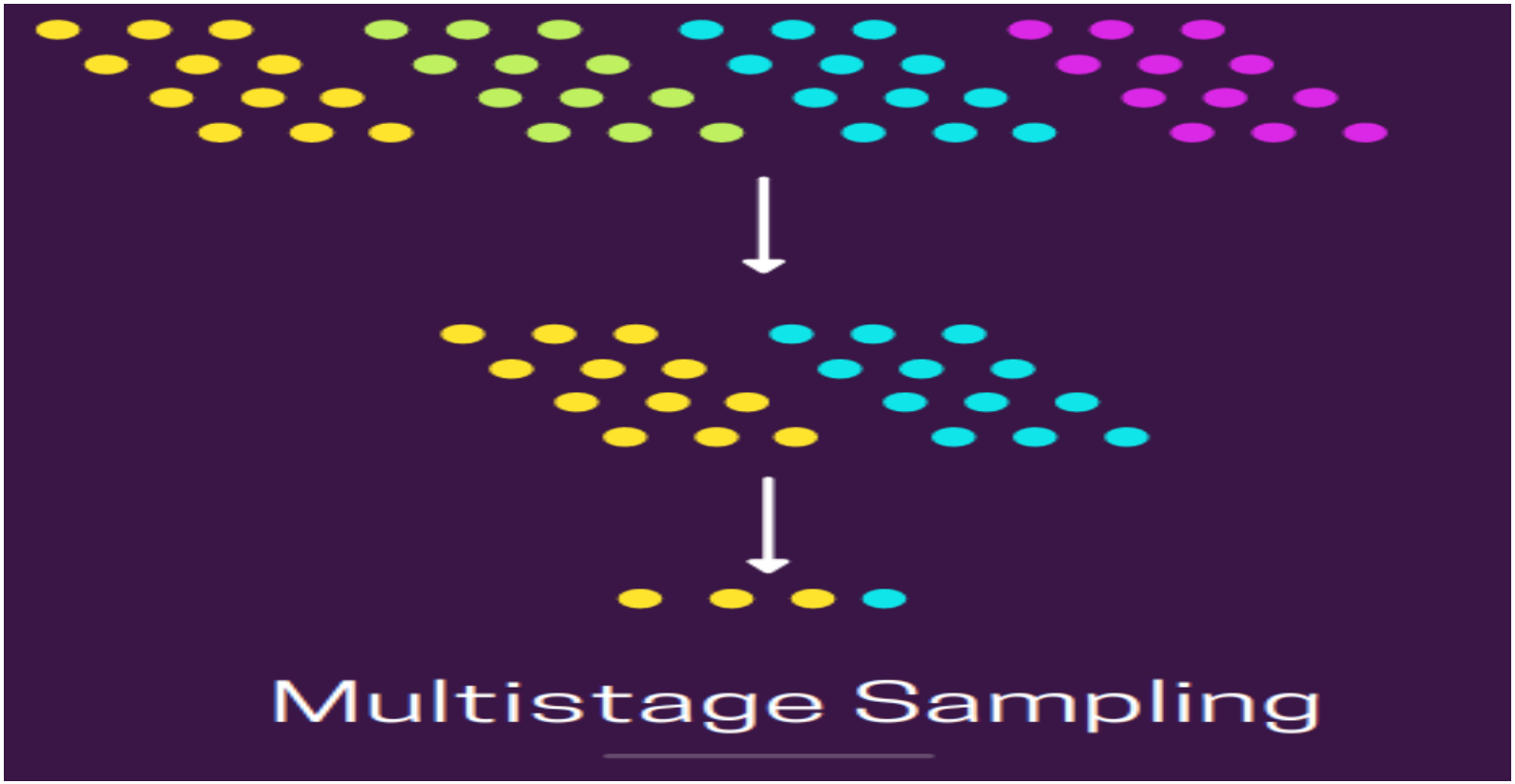
Systematic Sampling



Multistage sampling

- Under Multistage sampling, we **stack multiple sampling methods** one after the other. For example, at the first stage, cluster sampling can be used to choose clusters from the population and then we can perform random sampling to choose elements from each cluster to form the final set. *Below fig. shows a pictorial view of the same —*

Multistage Sampling



Convenience Sampling

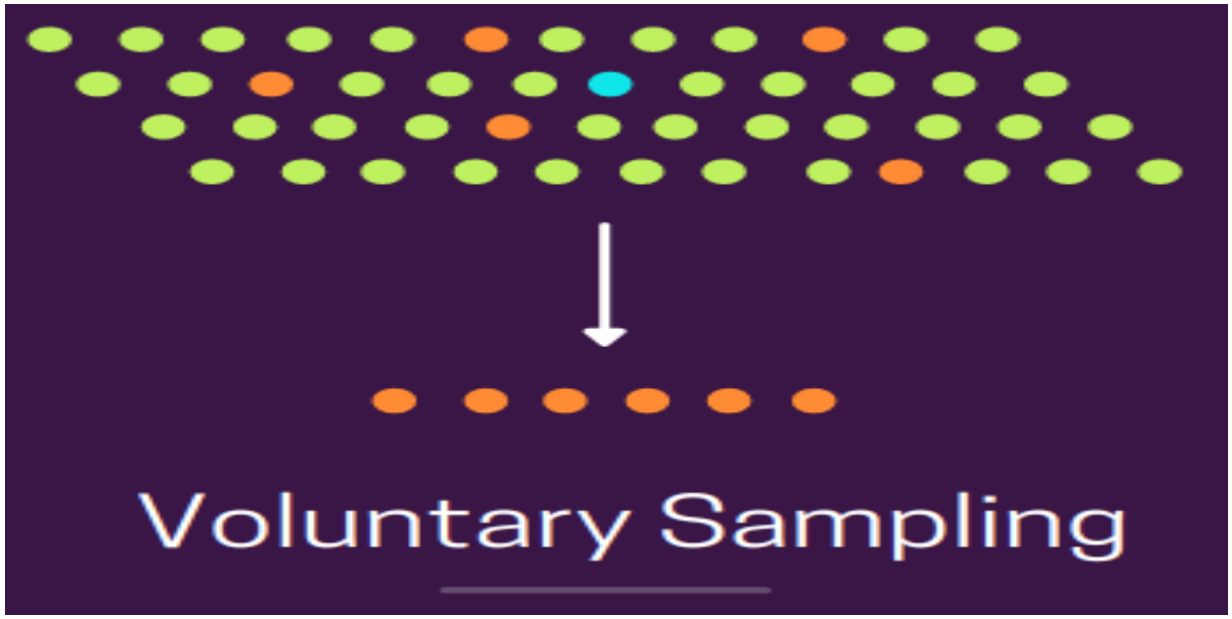
- Under convenience sampling, the researcher includes only those **individuals who are most accessible and available to participate in the study**. *Below fig. shows the pictorial view of the same — Blue dot is the researcher and orange dots are the most accessible set of people in orange's vicinity.*

Convenience Sampling

Voluntary Sampling

- Under Voluntary sampling, **interested people usually take part by themselves** by filling in some sort of survey forms. A good example of this is the you tube survey about "Have you seen any of these ads", which has been recently shown a lot. Here, the **researcher who is conducting the survey has no right to choose anyone.** *Below fig. shows the pictorial view of the same — Blue dot is the researcher, orange one's are those who voluntarily agreed to take part in the study.*

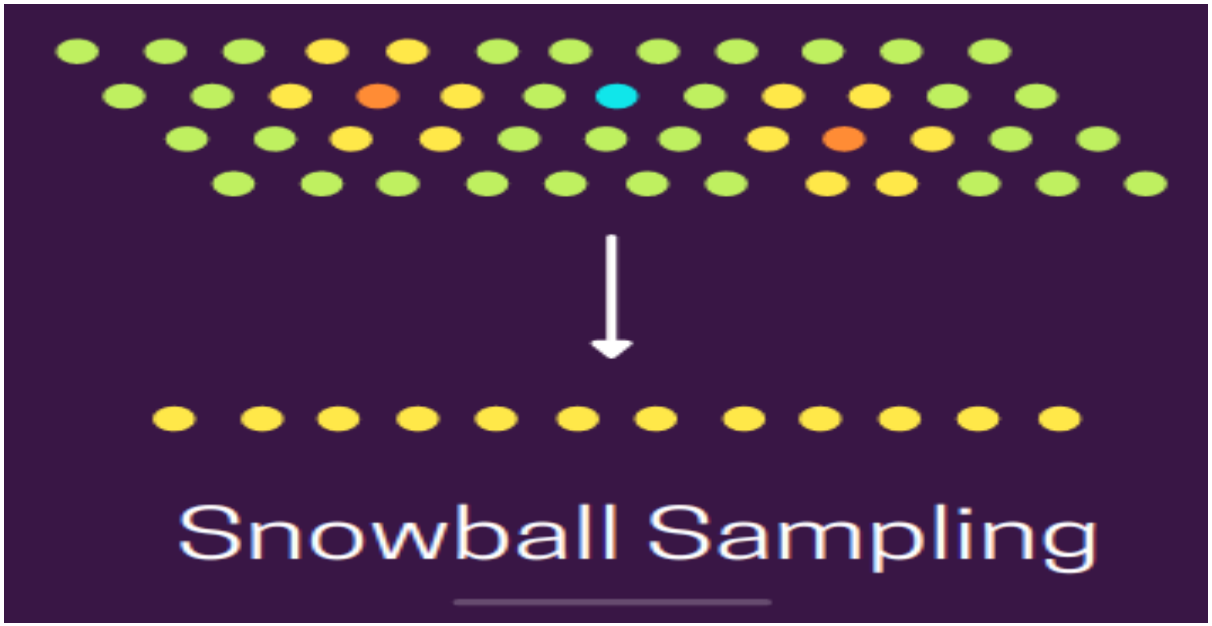
Voluntary Sampling



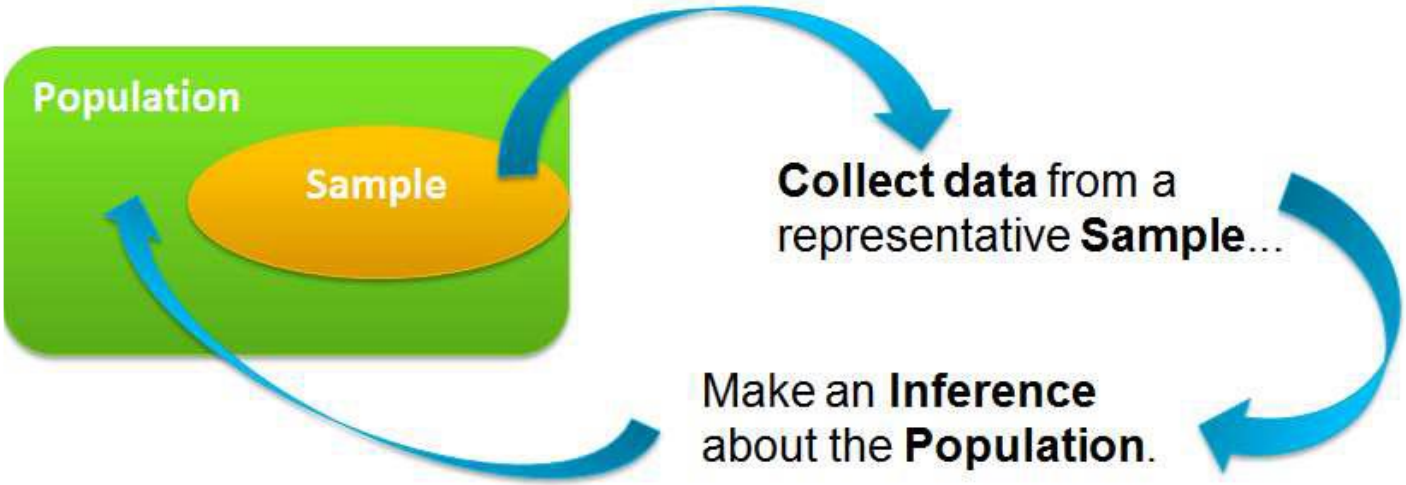
Snowball Sampling

- Under Snowball sampling, th**e final set is chosen via other participants**, i.e. The researcher asks other known contacts to find people who would like to participate in the study. *Below fig. shows the pictorial view of the same — Blue dot is the researcher, orange ones are known contacts(of the researcher), and yellow ones (orange's contacts) are other people that got ready to participate in the study.*

Snowball Sampling

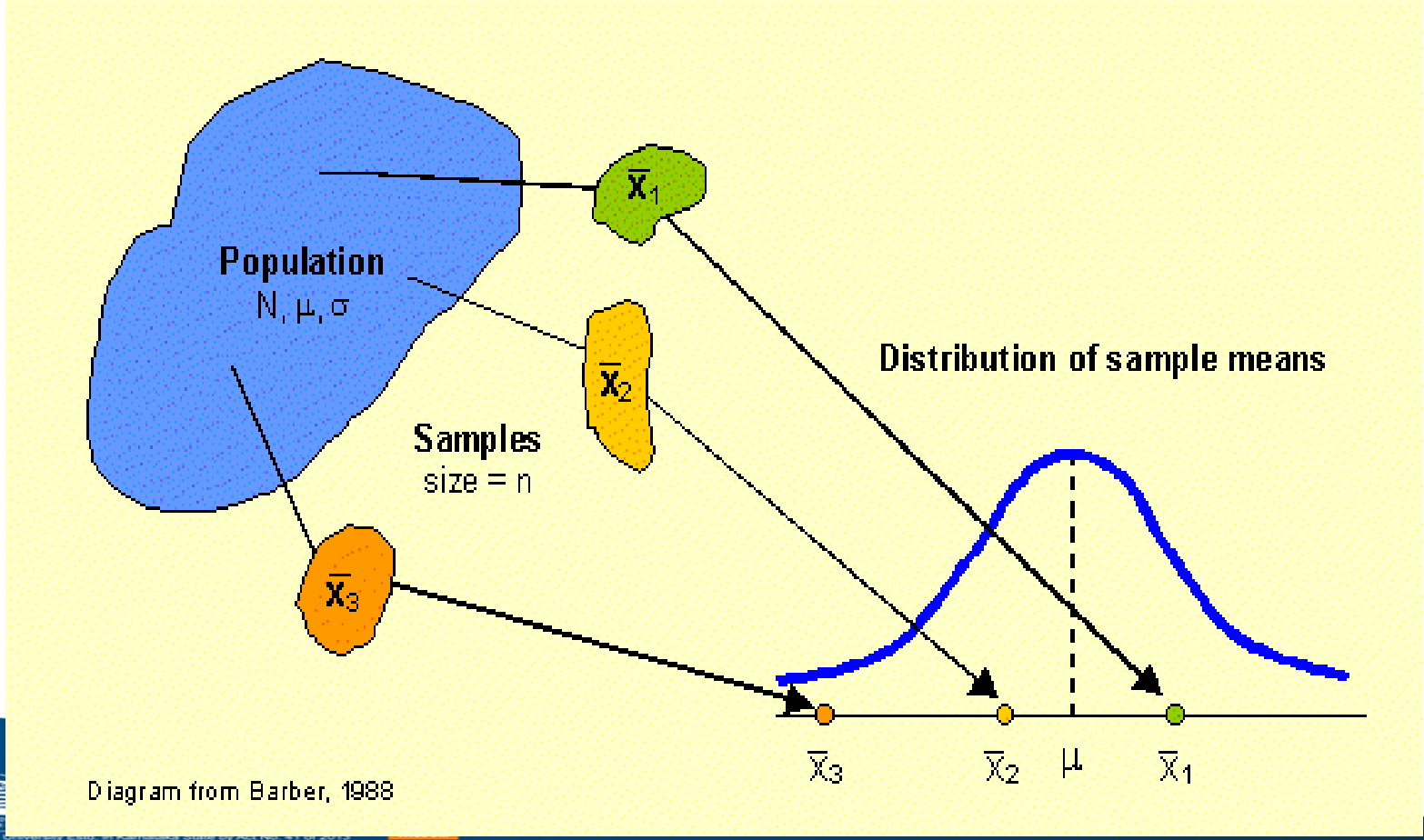


Sample-Outlook

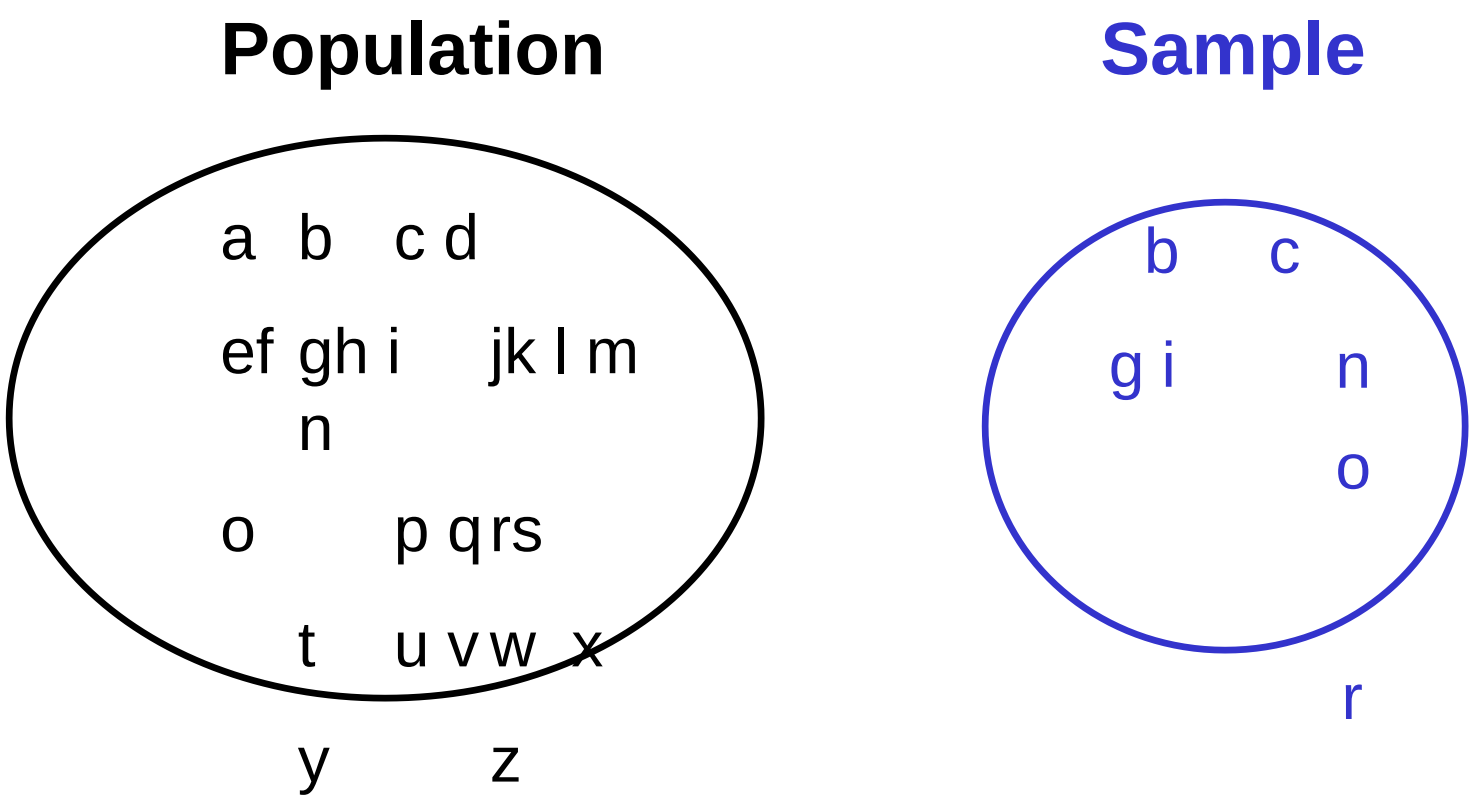


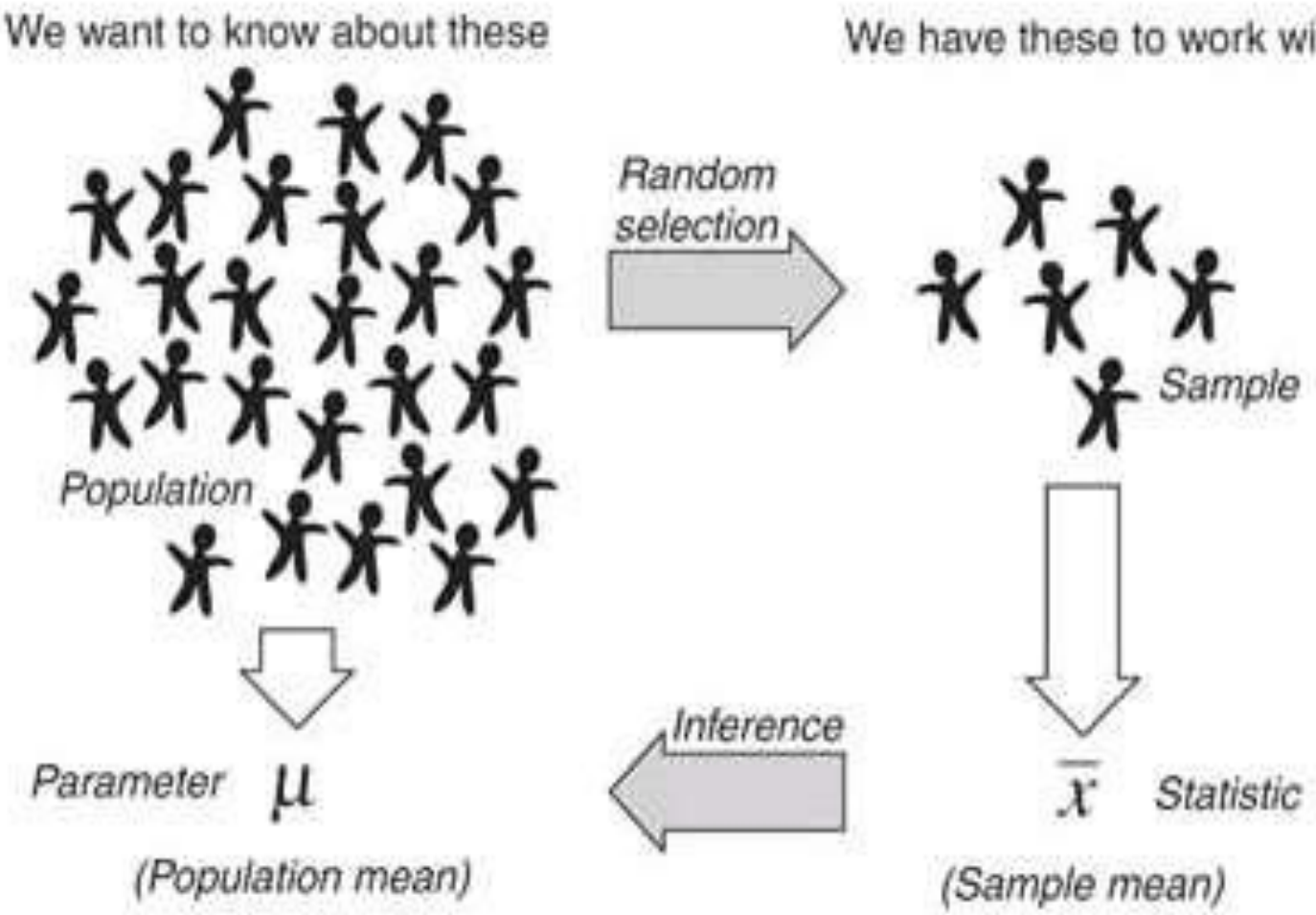
What is a Sampling Distribution?

If we take many random samples of equal size and calculate the mean value from each sample, we would begin to form a frequency distribution. The distribution of a statistic computed for each of many random samples is called a sampling distribution.



Population Vs Sample





The mean of a sampling distribution is called the expected value of the mean: it is the mean expected of the population.

Variance

The variance describes the spread of the data and measures how much the values of a variable differ from the mean. For variables that represent only a sample of some population and not the population as a whole, the variance formula is

$$\sigma^2 = \frac{\sum(xi - \bar{x})^2}{N - 1}$$

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

where the mean is

$\bar{x} = (3 + 4 + 4 + 5 + 5 + 5 + 6 + 6 + 6 + 7 + 7 + 8 + 9)/ 13$

$\bar{x} = 5.8$

- To calculate variance, we substitute the values into the variance formula:

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

= 2.86

Standard Deviation

- The standard deviation is the square root of the variance.
- The standard deviation is the most widely used measure of the deviation of a variable.
- The higher the value, the more widely distributed the variable's data values are around the mean.
- standard deviation is calculated as $\sqrt{2.86}$ or 1.69.

Example:

- A variable has a mean value of 45 with a standard deviation value of 6. Approximately 68% of the observations should be in the range 39–51 (45 ± one standard deviation) and approximately 95% of all observations fall within two standard deviations of the mean (between 33 and 57).

Sampling Distribution

- The **sampling distribution** of a statistic is the distribution of values taken by the statistic in **all** possible samples of the same size from the same population.
- In practice, it's difficult to take all possible samples of size *n* to obtain the actual sampling distribution of a statistic. Instead, we can use simulation to imitate the process of taking many, many samples.
- One of the uses of probability theory in statistics is to obtain sampling distributions without simulation.

Developing a Sampling Distribution

- **Assume there is a population …**
- Population size N=4
- Random variable, X,
  is age of individuals
- Values of X:

18, 20, 22, 24 (years)

Developing a Sampling Distribution

$$\mu = \frac{\sum X_i}{N}$$

$$= \frac{18 + 20 + 22 + 24}{4} = 21$$

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N-1}} = 2.5819$$

Sampling Distribution

$$\sigma = \sqrt{\frac{\Sigma (x_i - \mu)^2}{N-1}}$$

$$= \sqrt{\frac{(18-21)^2 + (20-21)^2 + (22-21)^2 + (24-21)^2}{3}}$$

$$= \sqrt{\frac{(-3)^2 + (-1)^2 + (1)^2 + (3)^2}{3}} = \sqrt{\frac{20}{3}} = 2.5819$$

$$\sigma = 2.5819$$

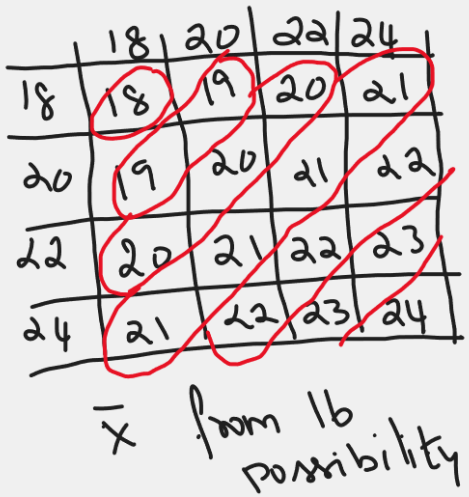
Developing a Sampling Distribution

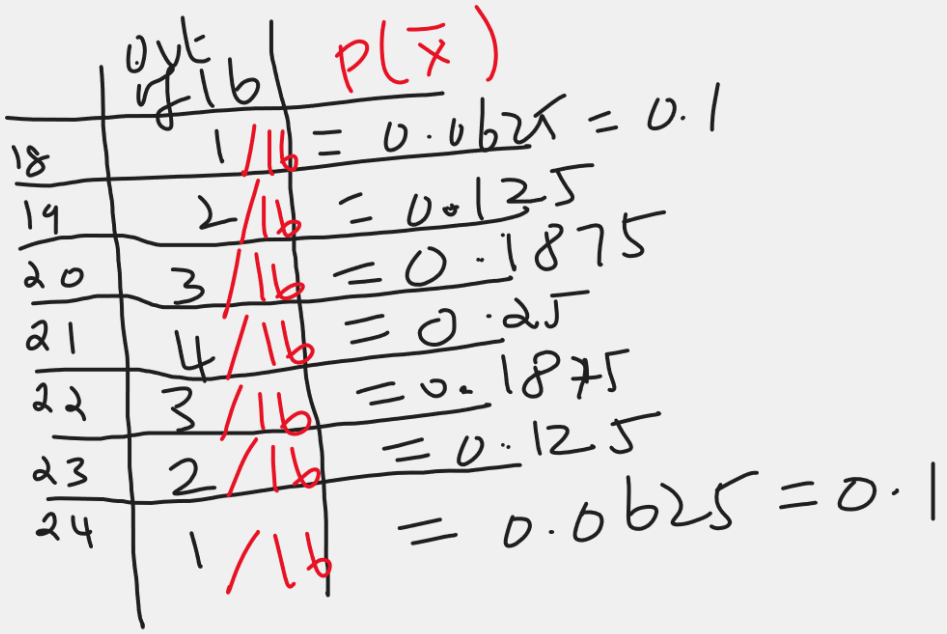
1st Obs	2nd Observation			
	18	20	22	24
18	18,18	18,20	18,22	18,24
20	20,18	20,20	20,22	20,24
22	22,18	22,20	22,22	22,24
24	24,18	24,20	24,22	24,24

16 possible samples (sampling with replacement)

Obs	18	20	22	24
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24

Sampling Distribution





Developing a Sampling Distribution

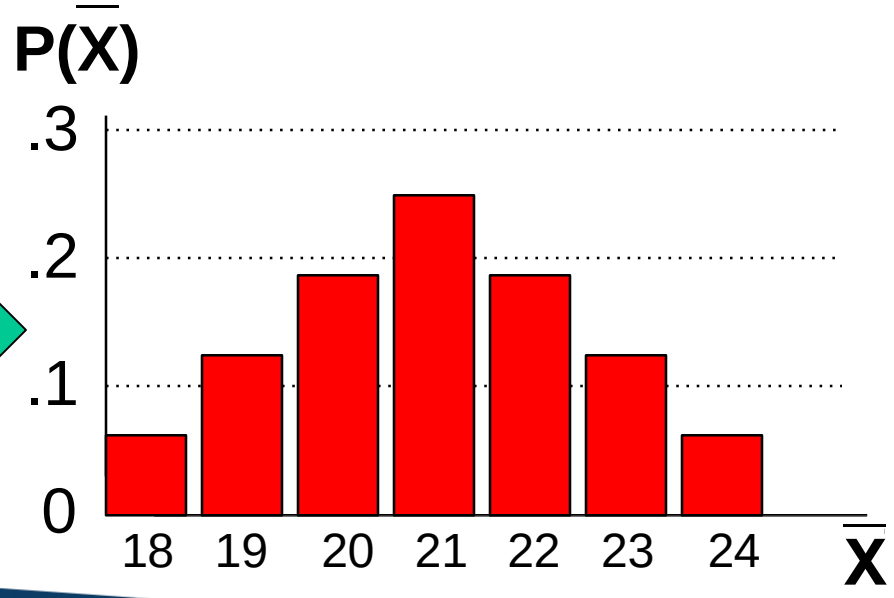
(continued)

Sampling Distribution of All Sample Means

16 Sample Means

1st Obs	2nd Observation			
	18	**20**	**22**	**24**
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24

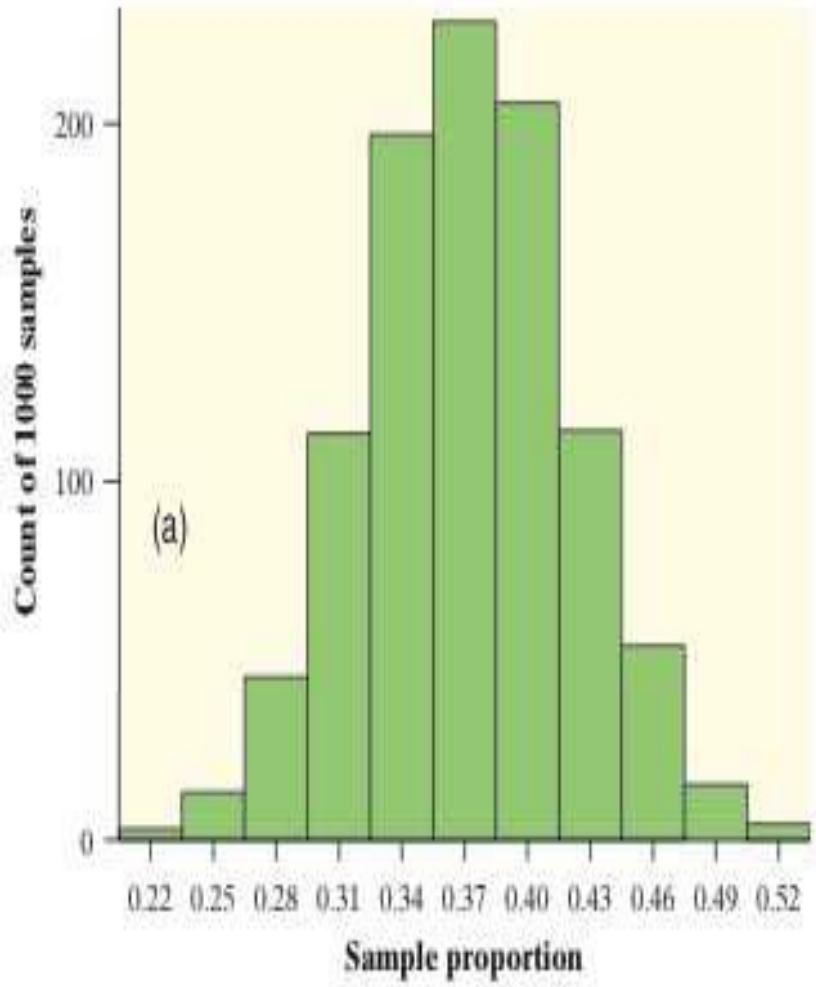
Sample Means Distribution

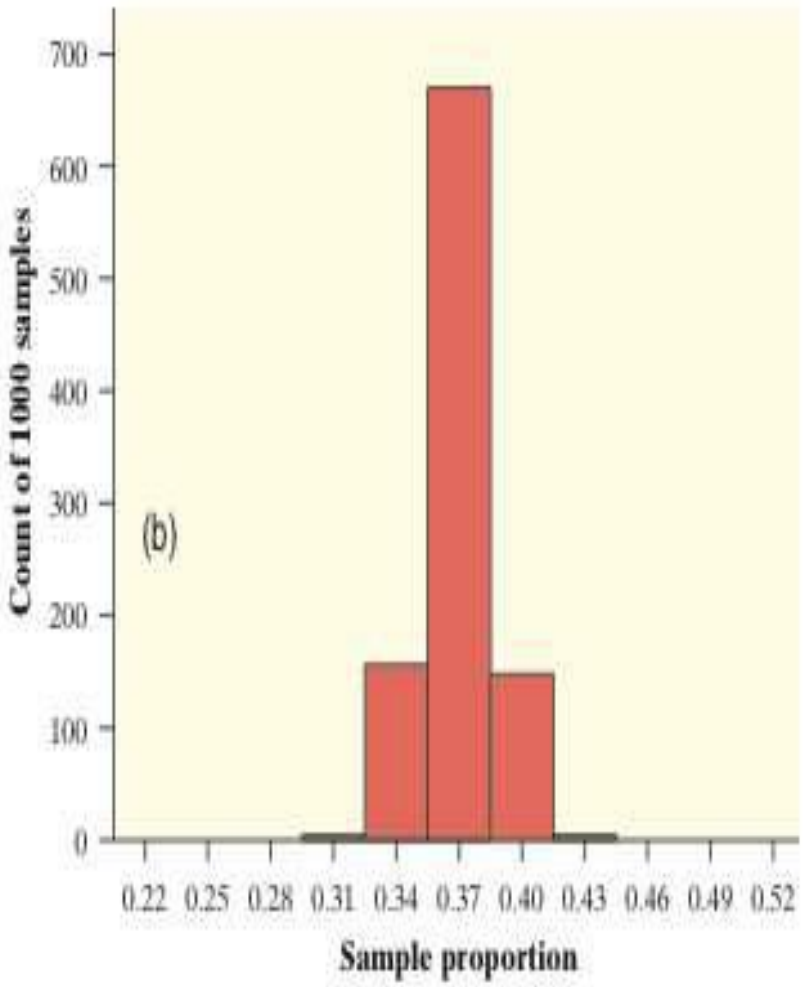


(no longer uniform)

Describing Sampling Distributions: Spread

- The **variability of a statistic** is described by the spread of its sampling distribution. This spread is determined primarily by the size of the random sample.
- Larger samples give smaller spread. The spread of the sampling distribution does not depend on the size of the population, as long as the population is at least 10 times larger than the sample.


Count of 1000 samples
200
100
0
(a)
0.22 0.25 0.28 0.31 0.34 0.37 0.40 0.43 0.46 0.49 0.52
Sample proportion


Count of 1000 samples
700
600
500
400
300
200
100
0
(b)
0.22 0.25 0.28 0.31 0.34 0.37 0.40 0.43 0.46 0.49 0.52
Sample proportion

Standard Error of the mean

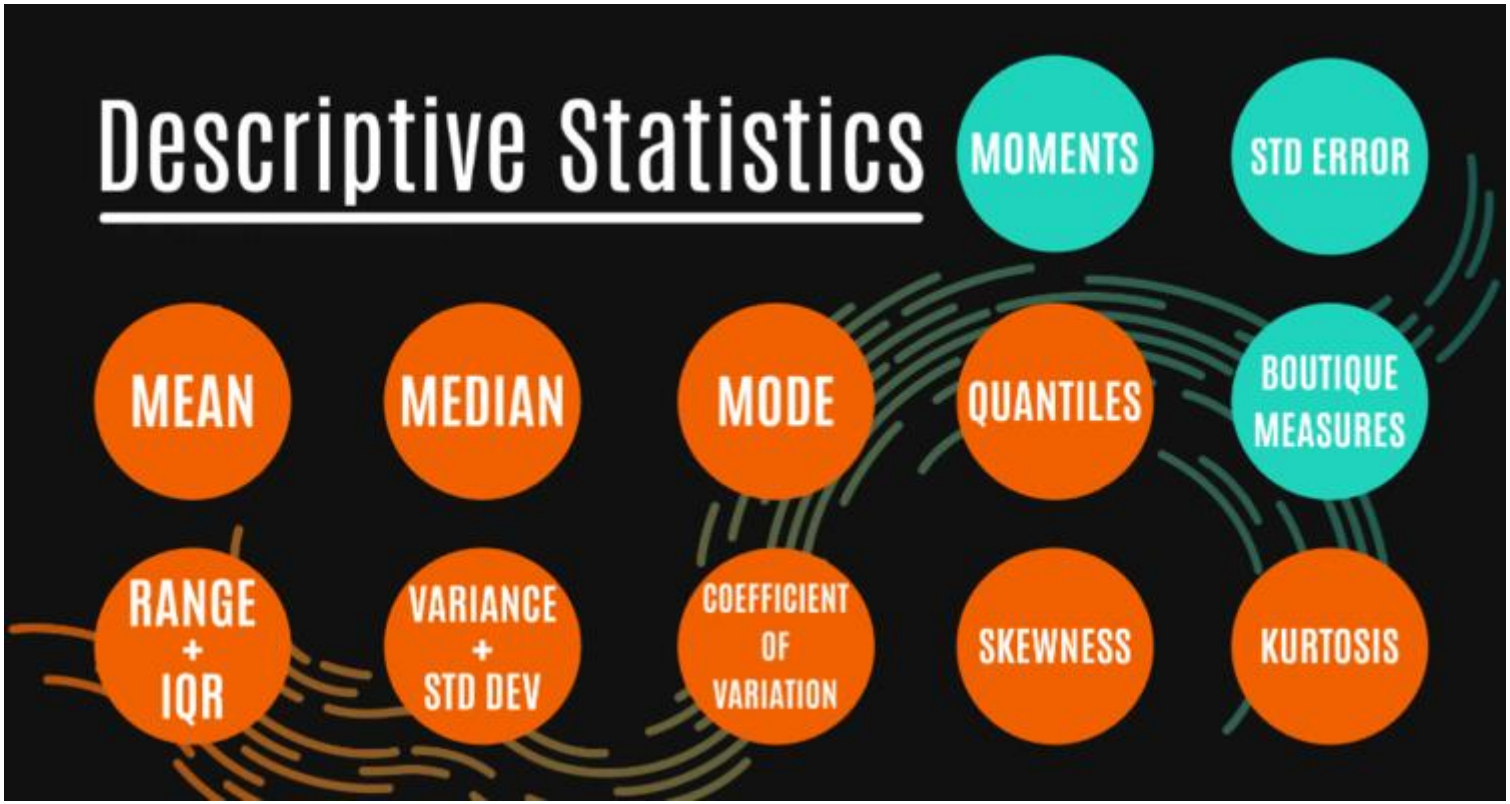
- Different samples of the same size from the same population will yield different sample means
- A measure of the variability in the mean from sample to sample is given by the Standard Error of the Mean:

Note that the standard error of the mean decreases as the sample size increases

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Descriptive Statistics

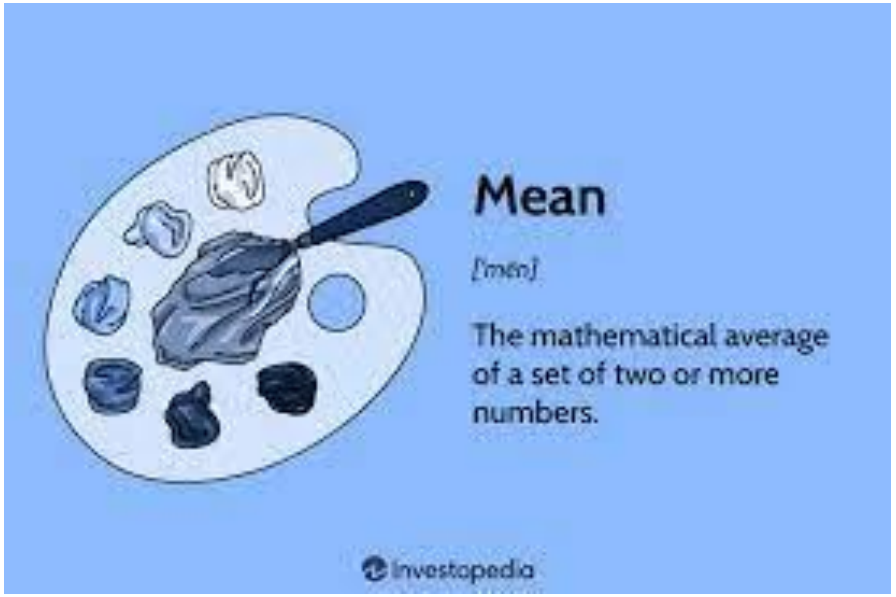
Descriptive Statistics



Understanding Descriptive Statistics

- Descriptive statistics, in short, help describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data.
- The most recognized types of descriptive statistics are measures of center: the mean, median, and mode, which are used at almost all levels of math and statistics.

Mean



$\bar{X}$ = (Sum of values ÷ Number of values)

$\bar{X} = (x_1 + x_2 + x_3 + \ldots + x_n)/n$

- **Example:**
- What is the mean of 2, 4, 6, 8 and 10?

- **Solution:**

First, add all the numbers.

2 + 4 + 6 + 8 + 10 = 30

Now divide by 5 (total number of observations).

Mean = 30/5 = 6

Weighted Mean

Definition

Weighted Mean — The mean calculated from data values with different frequencies. In the formula below the "weights," or frequencies are the coefficients of the x-terms.

$$\text{Weighted Mean} = \frac{w_1 \cdot x_1 + w_2 \cdot x_2 + \cdots + w_n \cdot x_n}{w_1 + w_2 \cdots + w_n}$$

Solved Example of Weighted Mean

Question: Suppose that a marketing firm conducts a survey of 1,000 households to determine the average number of TVs each household owns. The data show a large number of households with two or three TVs and a smaller number with one or four.
Every household in the sample has at least one TV and no household has more than four.
Find the mean number of TVs per household.

Number of TVs per Household	Number of Households
1	73
2	378
3	459
4	90

The mean number of TVs per household in this sample is 2.566.

Categorical Data Set

Can you find the mean of a categorical data set

$= [M, F, F, F, M, F]$

$= [0, 1, 1, 1, 0, 1]$

$\bar{x} = \frac{0+1+1+1+0+1}{6} = \frac{4}{6} = 0.666$

$\bar{x} = 0.666$

Arithmetic Mean

Find the average height of professional basketball players, using the following random sample

$\bar{x} = \frac{\Sigma x}{n}$

player	Height (m)
x_1	1.98
x_2	2.03
x_3	2.11
x_4	2.16
x_5	1.88

$\bar{x} = \frac{1.98 + 2.03 + 2.11 + 2.16 + 1.88}{5}$

$\bar{x} = 2.032$

→ Arithmetic mean

Harmonic Mean

Harmonic Mean

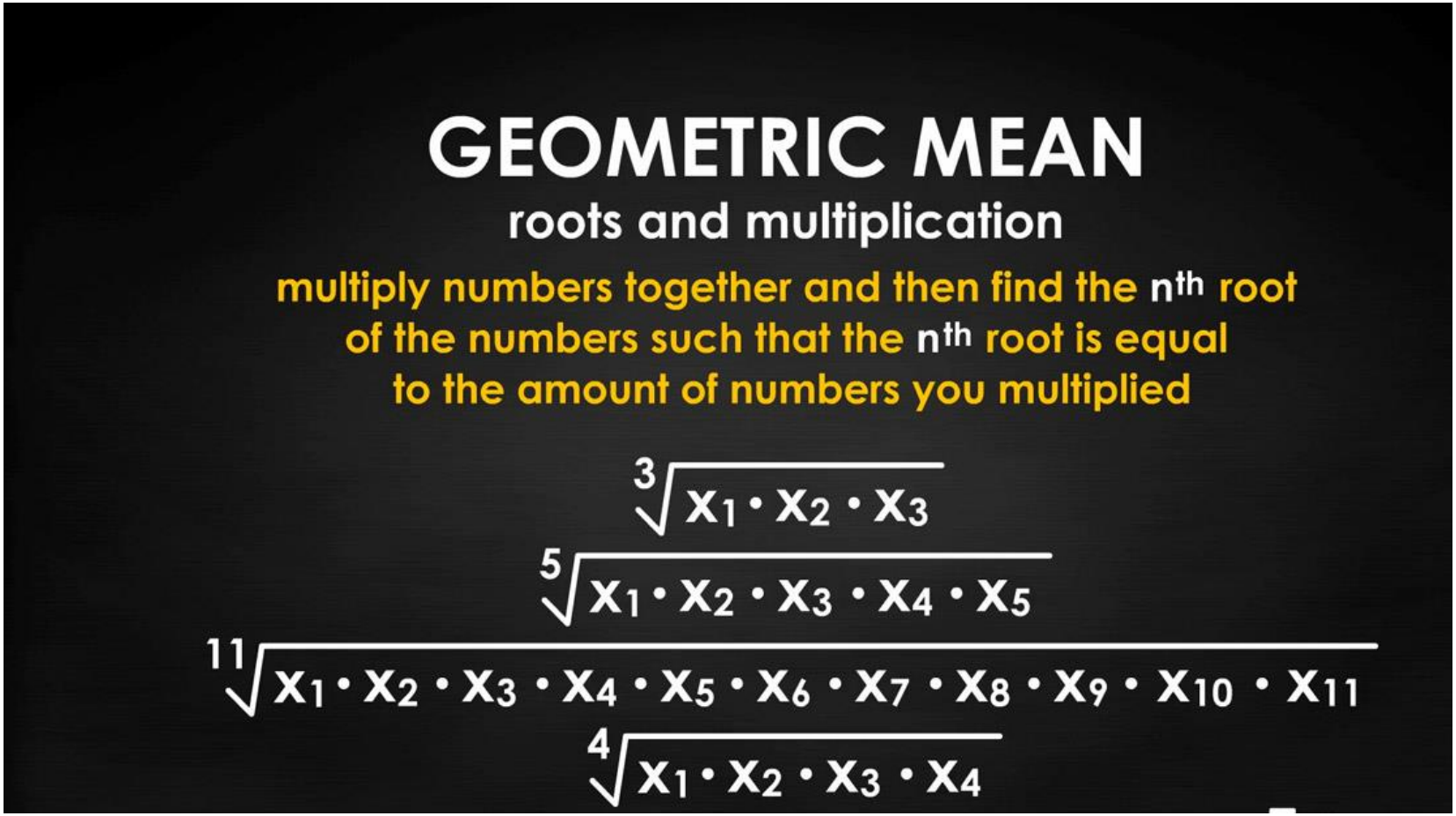
[här-'mä-nik 'mēn]

A type of numerical average calculated by dividing the number of observations by the reciprocal of each number in the series.

- **Example 1:**
  - Find the harmonic mean for data 2, 5, 7, and 9.

- **Solution:**
- Given data: 2, 5, 7, 9
- **Step 1: Finding the reciprocal of the values:**
- ½ = 0.5
- ⅕ = 0.2
- 1/7 = 0.14
- 1/9 = 0.11
- **Step 2: Calculate the average of the reciprocal values obtained from step 1.**
- Here, the total number of data values is 4.
- Average = (0.5 + 0.2 + 0.143 + 0.11)/4
- Average = 0.953/4
- **Step 3: Finally, take the reciprocal of the average value obtained from step 2.**
- Harmonic Mean = 1/ Average
- Harmonic Mean = 4/0.953
- Harmonic Mean = 4.19
- Hence, the harmonic mean for the data 2, 5, 7, 9 is 4.19.

Geometric Mean



- **Question 1: Find the G.M of the values 10, 25, 5, and 30**

- Solution : Given 10, 25, 5, 30
- We know that,

$=(10\times25\times5\times304)^{1/4}$

$=(375004)^{1/4}$

$= 13.915$

Therefore, the geometric mean = 13.915

Median

Median

['mē-dē-ən]

The middle number in a sorted, ascending or descending list of numbers. It can be more descriptive of that data set than the average.

1, 3, 3, **6**, 7, 8, 9

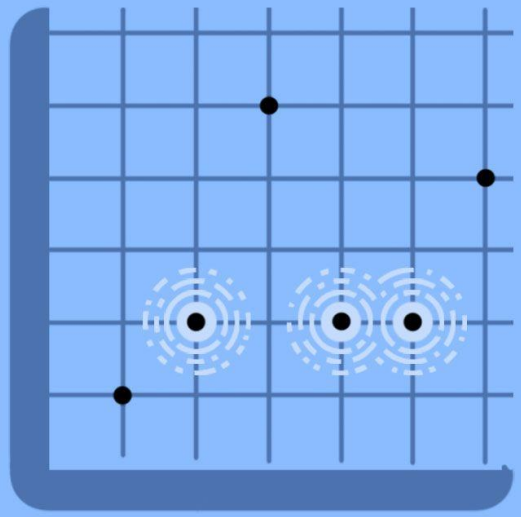
Median = **6**

1, 2, 3, **4**, **5**, 6, 8, 9

Median = (4 + 5) ÷ 2

= **4.5**

Mode



Mode

['mōd]

The value that appears most frequently in a data set. A set of data may have one mode, more than one mode, or no mode at all.

You can have more than one mode

There are two modes

3 9

Percentiles

Percentile

P_i is a measure used to indicate the value below which a given percentage of observation fall

Ex:

69	98	82	77	71
84	55	94	84	64
92	63	74	81	76

Example 1

55 63 64 69 71 74 76 77 81 82 84 84 92 94 98

55	63	64	69	71	74	76	77	81	82	84	84	92	94	98
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

$$L_p = \frac{P}{100}(n+1)$$

L_p: L = Location, p = percentile

$$L_{70} = 11.2 = 84$$

Example 2

eg 2 : 69 98 82 77 71
84 55 94 81 64
92 63 74 81 76

put in ordered array

55	63	64	69	71	74	76	77	81	82	84	84	92	94	98
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Formula : $L_p = \frac{P}{100}(n+1)$

To calculate the 50^{th} percentile = median

$L_{50} = \frac{50}{100}(15+1) = .5(16) = 8$

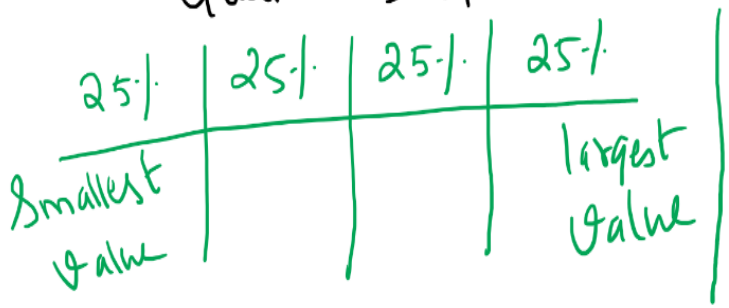
$L_{50} = 77$

Quantiles

Quantiles

The median is a quartiles because it splits the data into groups that contain the same number of data points.

Quantiles split data into fourths



Q_1 = first quantile - 25^{th} percentile

Q_2 = Second quantile - 50^{th} percentile

Q_3 = Third quantile - 75^{th} Percentile

Solution

For $Q_1 \rightarrow L_{25}$

$= \frac{25}{100}(15+1)$

$= .25(16)$

$= 4^{th}$ value

$\boxed{Q_1 = 69}$

$Q_2 \Rightarrow L_{50}$

$Q_2 = 77$

$Q_3 \rightarrow L_{75}$

$= \frac{75}{100}(15+1)$

$= .75(16)$

$= 12$

$\boxed{Q_3 = 84}$

Interquartile range

Interquartile range

- Measure of how the data is dispersed

Give us the spread of the middle 50% of the data

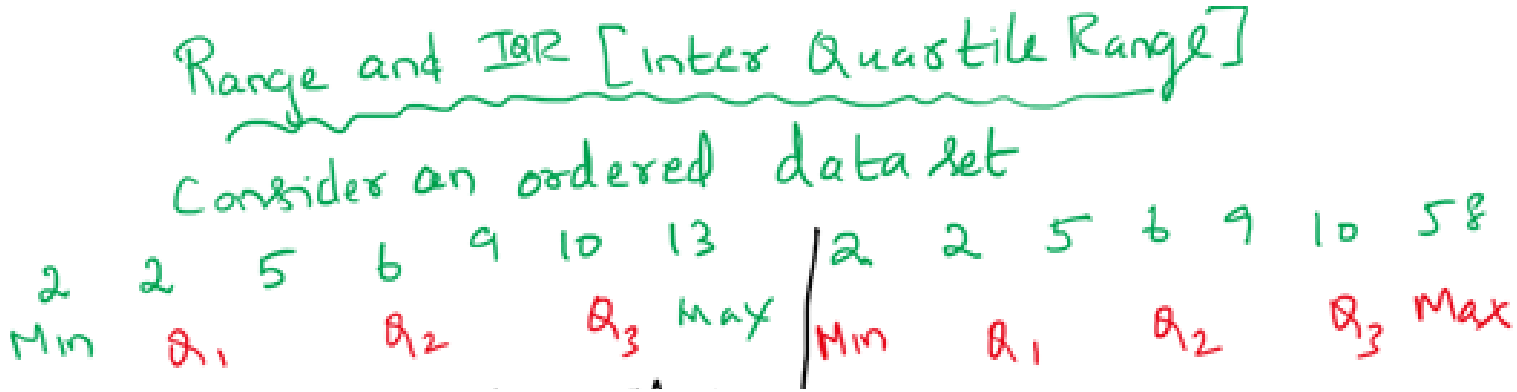
formula $IQR = Q_3 - Q_1$

$84 - 69$

$IQR = 15$

Range and IQR

The interquartile range formula is the first quartile subtracted from the third quartile: **IQR = $Q_3 - Q_1$**.



Variance and Standard deviation

Variance and Standard deviation

$$\text{Mean } \bar{x} = \frac{\Sigma x}{n}$$

$$\text{Variance} = S^2 = \frac{\Sigma (x - \bar{x})^2}{n - 1}$$

$$\text{Std. deviation} = S = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n - 1}}$$

Problem

Week	expenditure
1	$48.50
2	$87.40
3	$19.98
4	$59.74
5	$40.87
6	$105.51
7	$40.80
8	$23.10
9	$98.10
10	$60.54
11	$64.81
12	$48.01

Solution

$$\text{Mean} = \bar{x} = \frac{\Sigma x}{n}$$

$$= \$\ \frac{(48.50 + 87.40 + 19.98 + 59.74 + 40.87 + 105.51 + 40.80 + 23.10 + 98.10 + 60.54 + 64.81 + 48.01)}{12}$$

Mean = $ 58.11

$$\text{Variance} = s^2 = \frac{\Sigma (x - \bar{x})^2}{n-1} = \$\ 748.01$$

$$= \$\ \frac{\{(48.50-58.11)^2 + (87.40-58.11)^2 + (19.98-58.11)^2 + (59.74-58.11)^2 + (40.87-58.11)^2 + (105.51-58.11)^2 + (40.80-58.11)^2 + (23.10-58.11)^2 + (98.10-58.11)^2 + (60.54-58.11)^2 + (64.81-58.11)^2 + (48.01-58.11)^2\}}{11}$$

Solution

Standard deviation

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

$$= \sqrt{₹\ 748.01}$$

$$\boxed{S = ₹\ 27.35}$$

What is the Standard Deviation?

- The **standard deviation** indicates the spread of a variable around its mean value. Thus, the standard deviation is the average distance of all measured values of a variable from the mean value of the distribution.
- The standard deviation thus indicates how much the distribution of values scatters around the mean value. If the individual values scatter strongly around the mean value, a large standard deviation of the variable results. There are two slightly different equations for the calculation. On the one hand, the entire population can be used to calculate the standard deviation. On the other hand it can also be calculated if only one sample is available. If all values of the population are available, the following results are obtained

What is the Variance?

- In statistics, variance measures variability from the mean. For the calculation of the variance, the sum of the squared variances is divided by the number of values.
- The variance thus describes the squared average distance from the mean. Because the values are squared, the result has a different unit (the unit squared) than the original values. Therefore, it is difficult to relate the results.

- The coefficient of variation is a dimensionless relative measure of dispersion that is defined as the ratio of the standard deviation to the mean.
- If there are data sets that have different units then the best way to draw a comparison between them is by using the coefficient of variation.

Co-efficient of Variation

Co-efficient of Variation

$$C.V = \frac{S}{\bar{x}} = \frac{1}{2} = 0.5$$

S → S.D

$\bar{x}$ → Mean

eg (1) $x = [1, 2, 3]$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{1+2+3}{3} = \frac{6}{3} = 2 ; \quad \boxed{\bar{x} = 2}$$

Variance

$$= \frac{\Sigma (x-\bar{x})^2}{n-1} = \frac{(1-2)^2+(2-2)^2+(3-2)^2}{3-1} = \frac{(-1)^2+(0)^2+(1)^2}{2} = \frac{2}{2} = 1$$

$\therefore$ $\boxed{=1}$

$$S_x = \sqrt{\frac{\Sigma (x-\bar{x})^2}{n-1}} \qquad S_x = \sqrt{1} \qquad \boxed{S_x = 1}$$

Example 2

eg: 2 $\quad Y = [101, 102, 103]$

$$\bar{Y} = \frac{101+102+103}{3} = \frac{306}{3} = 102$$

$$\boxed{\bar{Y} = 102}$$

$$\text{Variance} = \frac{\Sigma (Y-\bar{Y})^2}{n-1}$$

$$= \frac{(101-102)^2 + (102-102)^2 + (103-102)^2}{2}$$

$$= \frac{(-1)^2 + (0)^2 + (1)^2}{2}$$

$$= 2/2 = 1$$

$$S_Y = \sqrt{1} = 1$$

$$C.V(Y) = \frac{1}{102} \boxed{= 0.0098}$$

Example 3:HW

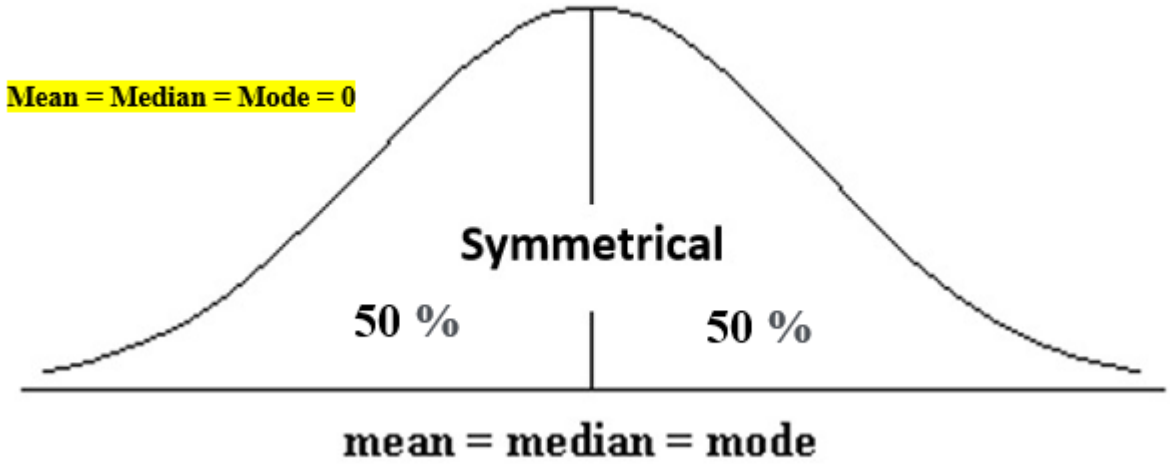
eg3 Fuel price (per gallon) were surveyed every week for 5 weeks in the US and Vietnam. which country experiences the greatest fuel price fluctuations?

USA	Vietnam
$2.70	11,612 ₫
$3.06	12,138 ₫
$2.87	12,980 ₫
$2.69	13,110 ₫
$2.71	12,084 ₫
$	

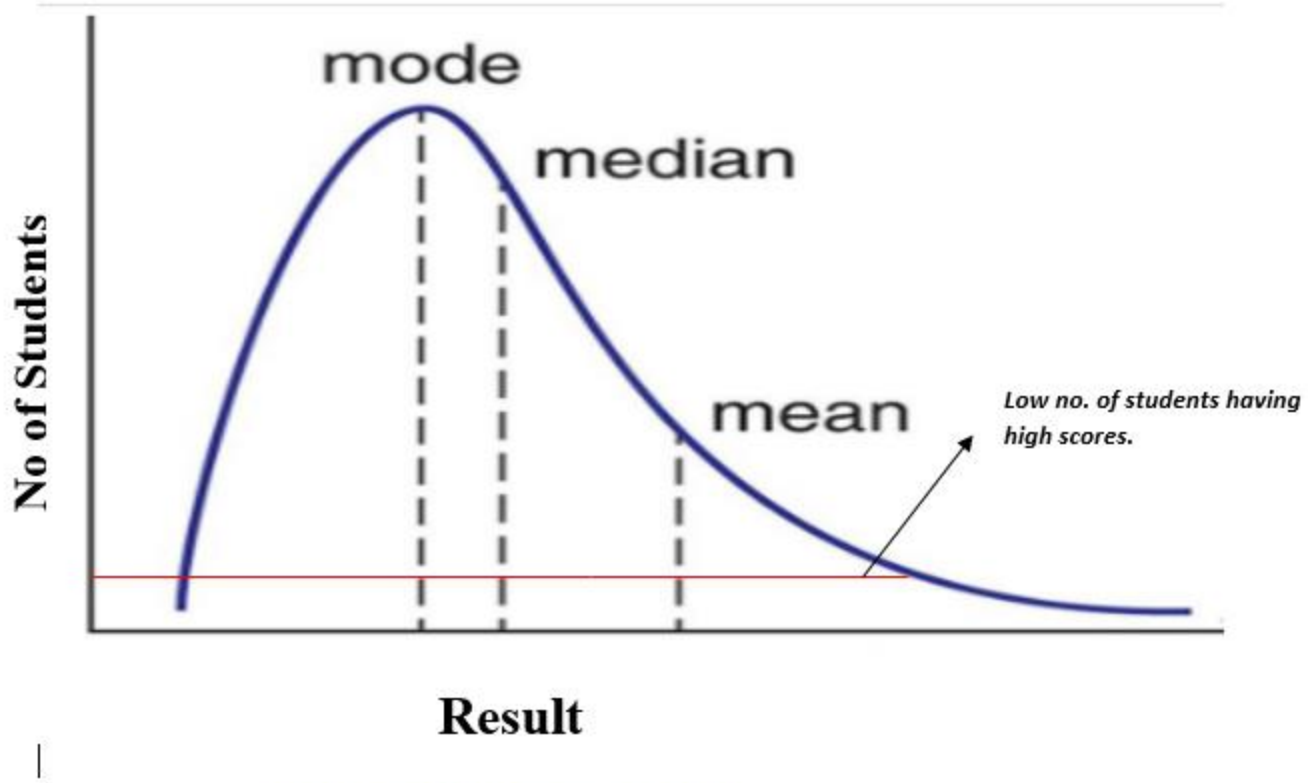
Skewness

- Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data.
- If the curve is shifted to the left or to the right, it is said to be skewed.
- Skewness can be quantified as a representation of the extent to which a given distribution varies from a normal distribution.
- A normal distribution has a skew of zero, while a lognormal distribution, for example, would exhibit some degree of right-skew.

Zero-Skewness

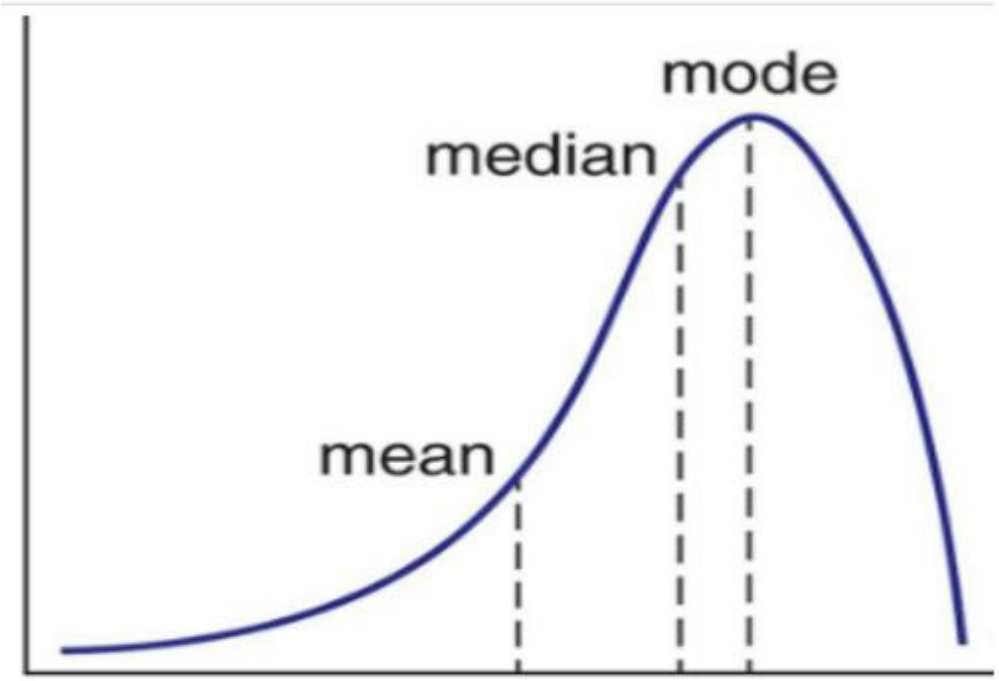


Positive-Skew



Mean > Median > Mode

Negative-Skew

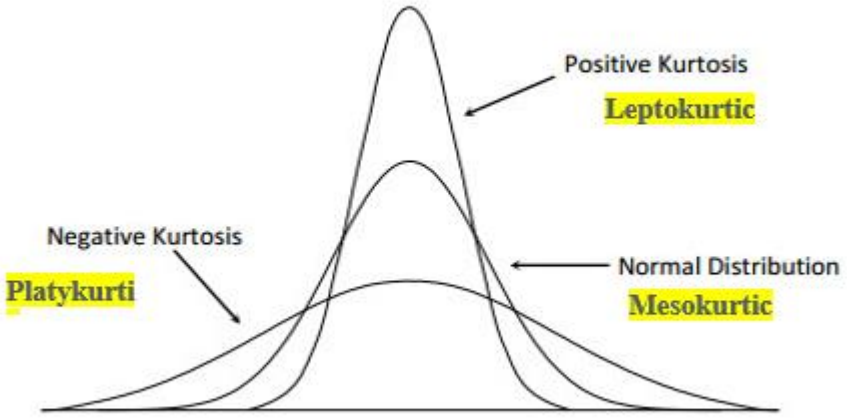


Mode > Median > mean

Kurtosis

Kurtosis refers to the degree of presence of outliers in the distribution.

Kurtosis is a statistical measure, whether the data is heavy-tailed or light-tailed in a normal distribution.



Types of excess kurtosis

- *Leptokurtic or heavy-tailed distribution (kurtosis more than normal distribution).*
- *Mesokurtic (kurtosis same as the normal distribution).*
- *Platykurtic or short-tailed distribution (kurtosis less than normal distribution).*

Standard Error

Standard Error is a measure of uncertainity in the sample mean

$$S.E\left(\bar{x}\right) = \frac{S_x}{\sqrt{n}}$$

$$\bar{x} = 560/5 = 112$$

Variance $= (127-112)^2 + (109-112)^2 + \dots$

$$\frac{(121-112)^2 + (94-112)^2 + (109-112)^2}{4}$$

$= 162$

$S_x = \sqrt{162} = 12.7$

$$S.E = \frac{S_x}{\sqrt{n}} = \frac{12.7}{\sqrt{5}} = 5.67$$

Moments

- Moments are a set of statistical parameters to measure a distribution. Four moments are commonly used:
- Mean: the average
- Variance:

Standard deviation is the square root of the variance: an indication of how closely the values are spread about the mean. A small standard deviation means the values are all similar. If the distribution is normal, 63% of the values will be within 1 standard deviation.

Moments

- Skewness: measure the asymmetry of a distribution about its peak; It is a number that describes the shape of the distribution.

It is often approximated by Skew = (Mean - Median) / (Std dev).

If skewness is positive, the mean is bigger than the median and the distribution has a large tail of high values.

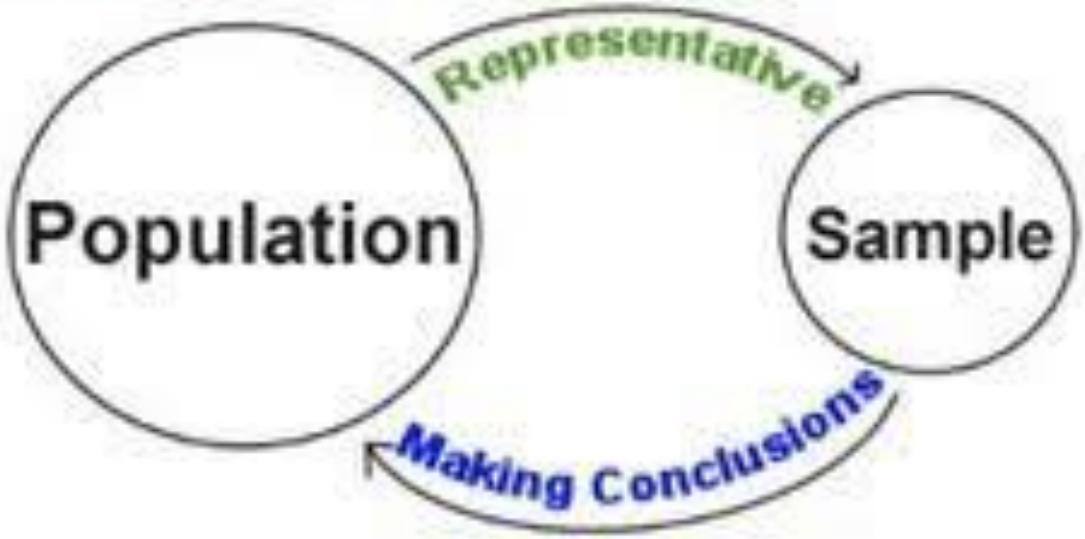
If skewness is negative, the mean is smaller than the median and the distribution has a large tail of small values.

- Kurtosis: measures the peakedness or flatness of a distribution.

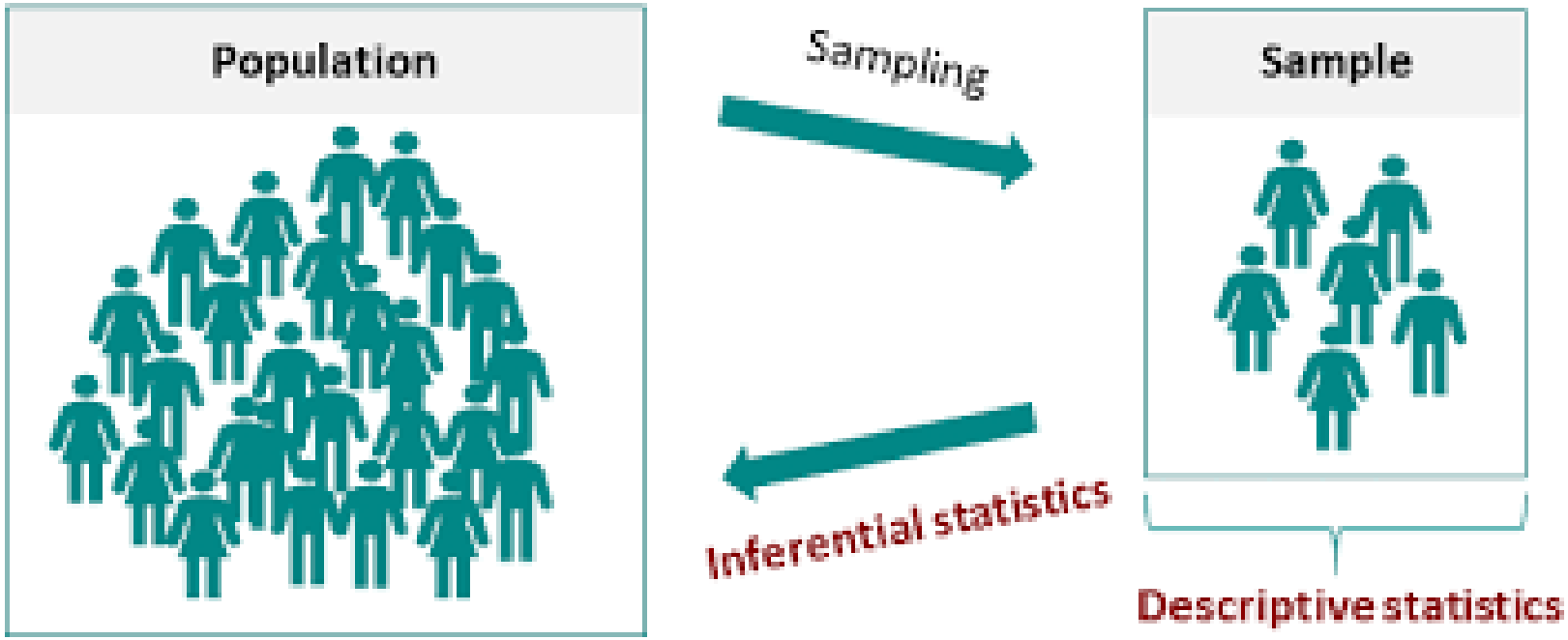
Positive kurtosis indicates a thin pointed distribution.

Negative kurtosis indicates a broad flat distribution.

Inferential Statistics



Inferential Statistics



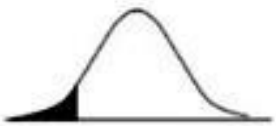
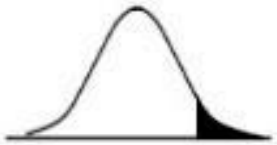
What is Inferential Statistics?

- Descriptive statistics describe the important characteristics of data by using mean, median, mode, variance etc. It summarises the data through numbers and graphs.
- In Inferential statistics, we make an inference from a sample about the population. The main aim of inferential statistics is to draw some conclusions from the sample and generalize them for the population data.

Importance of Inferential Statistics

- Making conclusions from a sample about the population
- To conclude if a sample selected is statistically significant to the whole population or not
- Comparing two models to find which one is more statistically significant as compared to the other.
- In feature selection, whether adding or removing a variable helps in improving the model or not.

Type of Hypothesis Test

	Type of Hypothesis Test	Alternate Hypotheis	Null Hypothesis
One – Tailed Tests(Values of the sample which cause rejection of H_0 fall only in one tail of the sample)	Left – tailed Test 	<	>=
	Right – tailed Test 	>	<=
	Two tailed test 	≠	=

Purpose of hypothesis testing is to get rid of randomness.

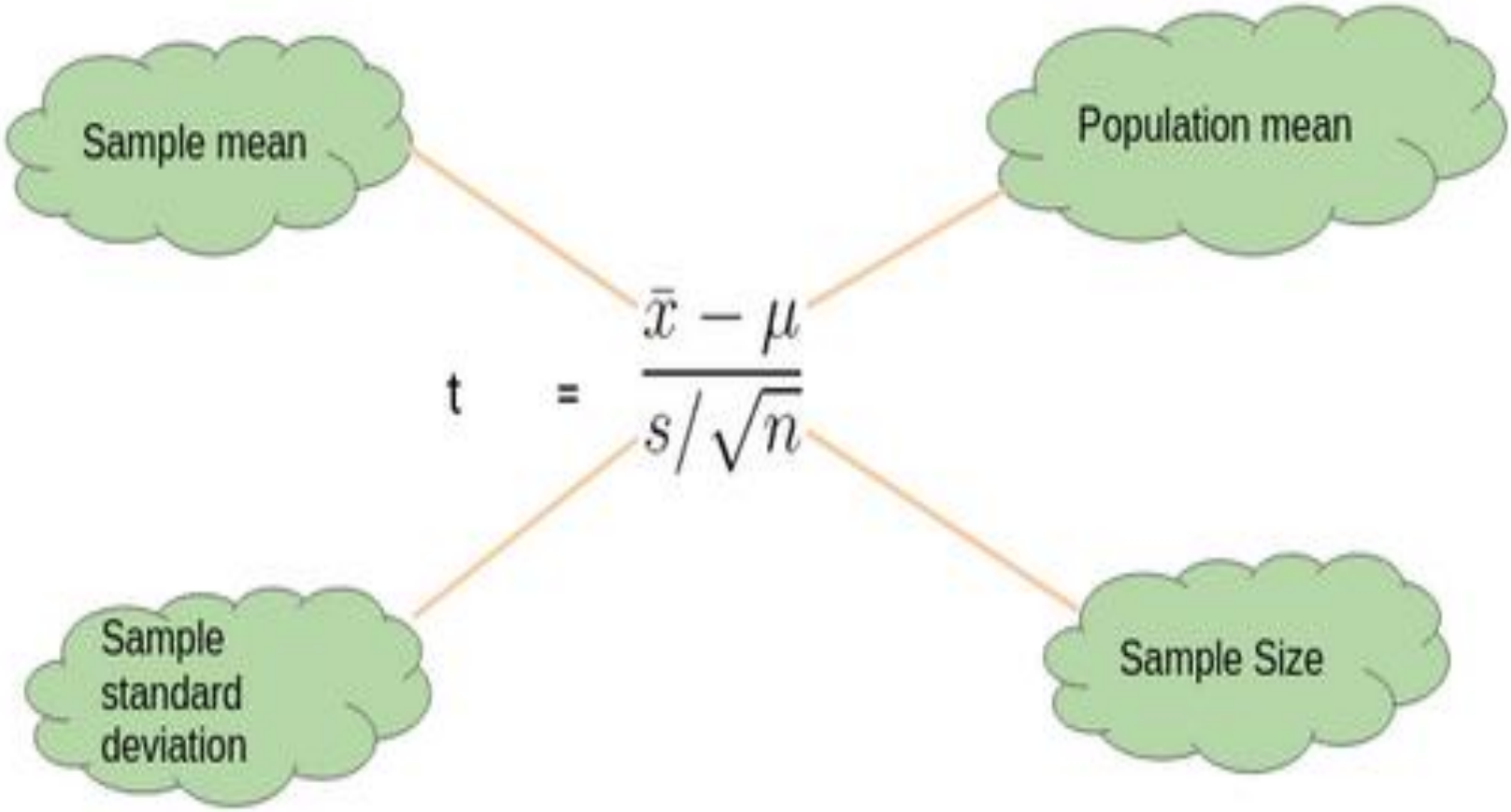
What is the t-Test?

- We do not know the population variance
- Our sample size is small, n < 30
- **One-Sample t-Test**
- We perform a One-Sample t-test when we want to **compare a sample mean with the population mean.** The difference from the Z Test is that we do **not have the information on Population Variance** here. We use the **sample standard deviation** instead of population standard deviation in this case.

Conditions for Acceptance & Rejection

- If calculated t value > Critical t then Alternate is accepted.
- If calculated t value < Critical t then Null is accepted

- If P value <= significance then Null is rejected
- If P value > significance then Null is accepted


Sample mean
Population mean
$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$
Sample standard deviation
Sample Size

Here's an Example to Understand a One Sample t-Test

- Let's say we want to determine if on average girls score more than 600 in the exam. We do not have the information related to variance (or standard deviation) for girls' scores. To a perform t-test, we randomly collect the data of 10 girls with their marks and choose our α value (significance level) to be 0.05 for Hypothesis Testing.

Girls_Score
587
602
627
610
619
622
605
608
596
592

In this example:

- Mean Score for Girls is 606.8
- The size of the sample is 10
- The population mean is 600
- Standard Deviation for the sample is 13.14

To find critical value

1. Degree of freedom(DF) = sample size -1

   = 10 -1=9

2. Significance level (SL)= 0.05

3. Refer table 2 pdf file-→ DF->9 & SL->0.05

Therefore critical value of T is 1.833

Finding P Value

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

- We know that

  T = 1.64

  DF = 9

  Significance = 0.05

Refer table 3 pdf file→ pg no:2

Therefore p values is 0.068

This p-value can also be used to assess the null hypothesis, where the null hypothesis is rejected if it is less than the value of alpha. This value can be looked up using a standard z-distribution table as found by an online search or readily available software.

Our **P-value is greater than 0.05 thus we fail to reject the null hypothesis** and don't have enough evidence to support the hypothesis that on average, girls score more than 600 in the exam.

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

$$= \frac{606.8 - 600}{13.14/\sqrt{10}}$$

$$= 1.64$$

Critical Value = 1.833

t score < Critical Value

P value = 0.0678

P value > 0.05

$H_0 : \mu \leq 600$

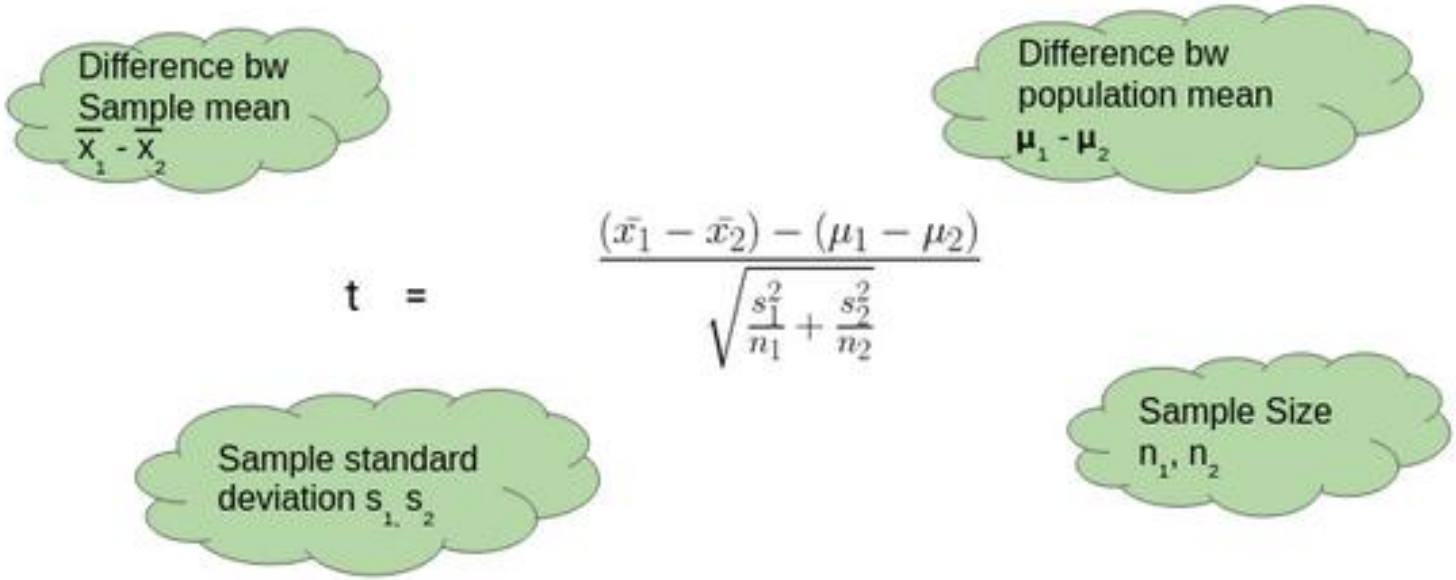
$H_1 : \mu > 600$

Conculsion

- T Score < Critical value
- P Value > Significance level
- There fore we eliminate the Alternative hypothesis
- We choose Null hypothesis

Two-Sample t-Test

- We perform a Two-Sample t-test when we want to compare the mean of two samples.



Here's an Example to Understand a Two-Sample t-Test

- Here, let's say we want to determine if on average, boys score 15 marks more than girls in the exam. We do not have the information related to variance (or standard deviation) for girls' scores or boys' scores. To perform a t-test. we randomly collect the data of 10 girls and boys with their marks. We choose our α value (significance level) to be 0.05 as the criteria for Hypothesis Testing.

- H0: Null Hypothesis: Boys will not score more than 15 marks more than girls.

- H1: Alternate Hypothesis: Boys score more than 15 marks more than girls.

Girls_Score
587
602
627
610
619
622
605
608
596
592

Boys_Score
626
643
647
634
630
649
625
623
617
607

In this example:

- Mean Score for Boys is 630.1
- Mean Score for Girls is 606.8
- Difference between Population Mean 15
- Standard Deviation for Boys' score is 13.42
- Standard Deviation for Girls' score is 13.14
- Size of sample is 20

To find critical value

1. Degree of freedom(DF) = sample size -1

   = 20 -1=19

2. Significance level (SL)= 0.05

3. Refer table 2 pdf file-→ DF->19 & SL->0.05(Two tail)

Therefore critical value of T is 2.093

Finding P Value

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- We know that

  T =1.39

  DF = 19

  Significance = 0.05

Refer table 3 pdf file→ pg no:2

Therefore p values is 0.089

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$\frac{(630.1 - 606.8) - (15)}{\sqrt{\frac{(13.42)^2}{10} + \frac{(13.14)^2}{10}}}$$

Critical Value = 2.093

t = 1.39

P value = 0.089

Conclusion

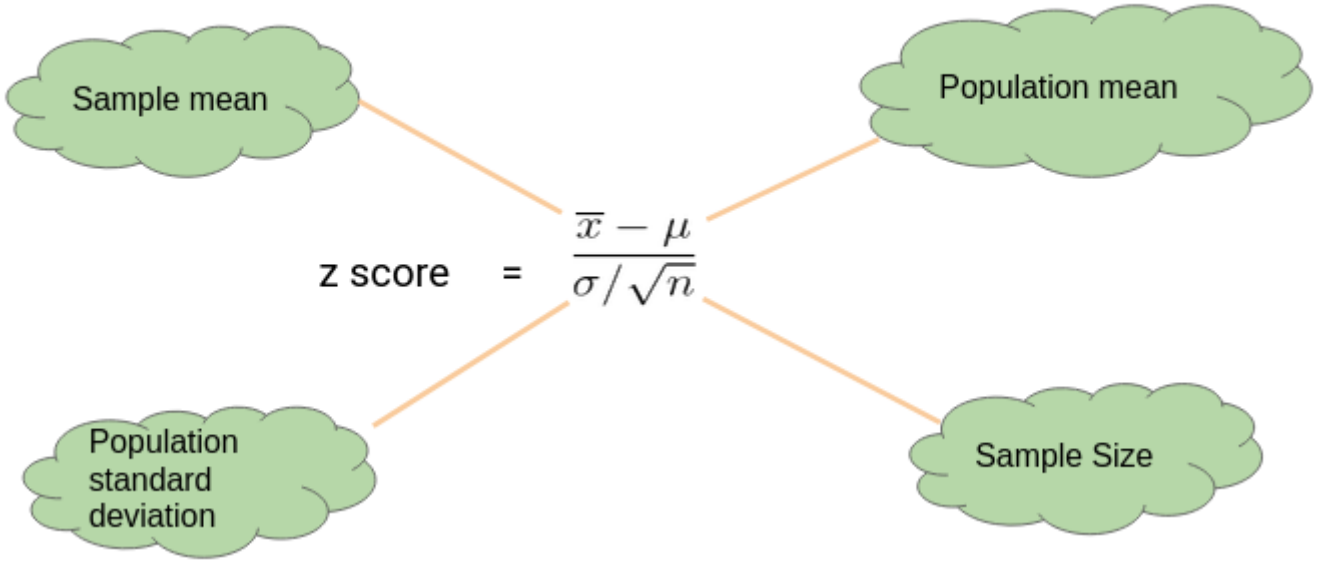
Thus, **'t' value is less than critical value and 'P'-value is greater than 0.05 so we fail to reject the null hypothesis** and conclude that on average boys do not score 15 marks more than girls in the exam.

What is the Z Test?

- z tests are a statistical way of testing a hypothesis when either:
- We know the population variance, or
- We do not know the population variance but our sample size is large n ≥ 30
- *If we have a sample size of less than 30 and do not know the population variance, then we must use a t-test.*

One-Sample Z test

We perform the One-Sample Z test when we want to compare **a sample mean with the population mean.**



Z test

Z test

$\bar{x}$ = mean score

μ = population mean

n = size of the sample

σ = Standard deviation for population

α = significance level

σ is known and $n > 30$

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

Z Test problem 1

- The population of all verbal GRE scores are known to have a standard deviation of 8.5. The UW Psychology department hopes to receive applicants with a verbal GRE scores over 210. This year, the mean verbal GRE scores for the 42 applicants was 212.79. Using a value of $\alpha = 0.05$ is this new mean significantly greater than the desired mean of 210?

Z value

$$z = \frac{(\bar{x} - \mu)}{\sigma / \sqrt{n}}$$

$$z = \frac{(212.79 - 210)}{8.5 / \sqrt{42}}$$

Given data

$\sigma = 8.5$	$\alpha = 0.05$
$n = 42$	$\mu = 210$
$\bar{x} = 212.79$	

$$= \frac{2.79}{1.31} = 2.129 = 2.13$$

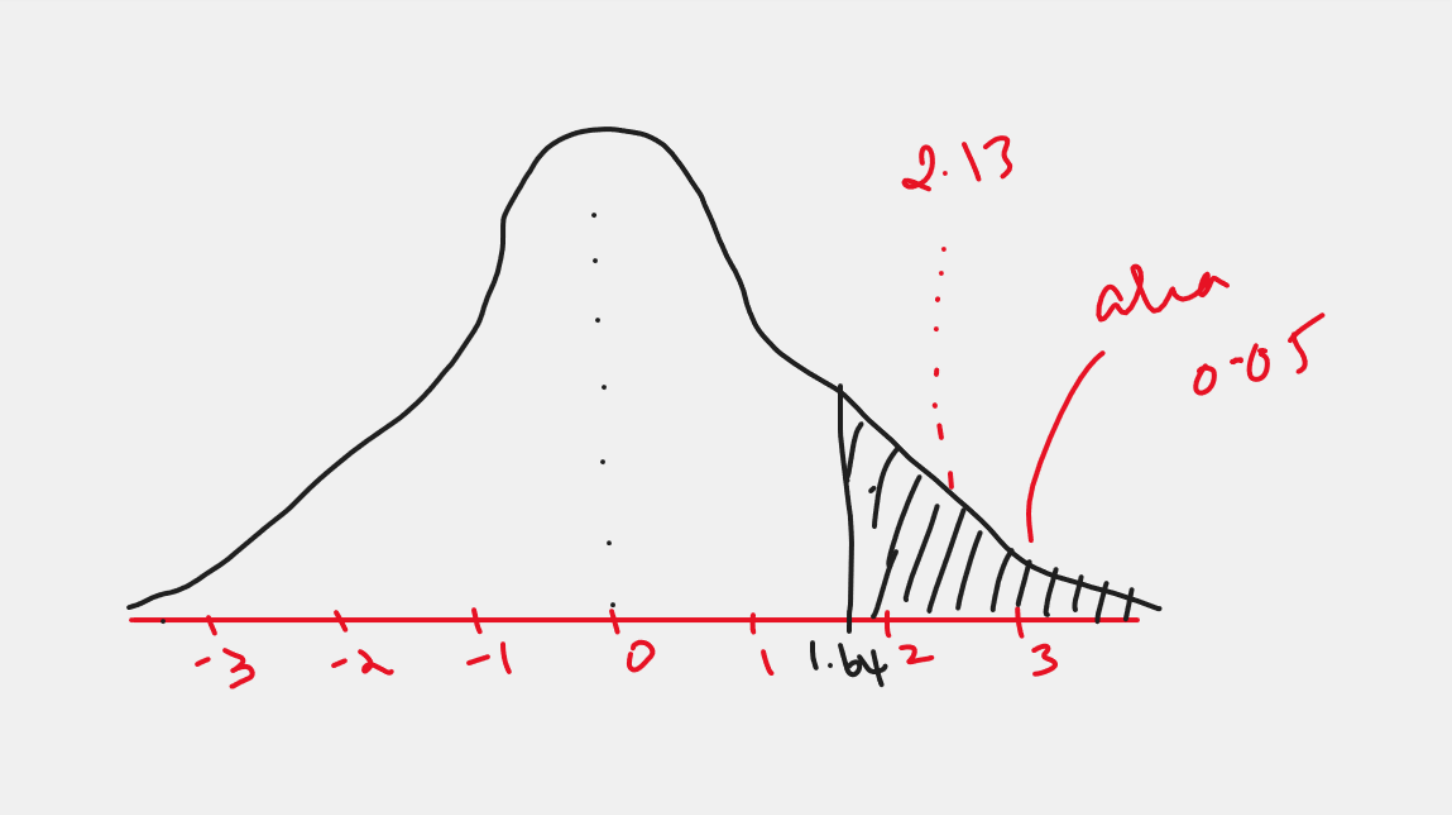
$$z = 2.13$$

You can see that the CRITICAL VALUE of Z is 1.64

Z Score > Critical Value

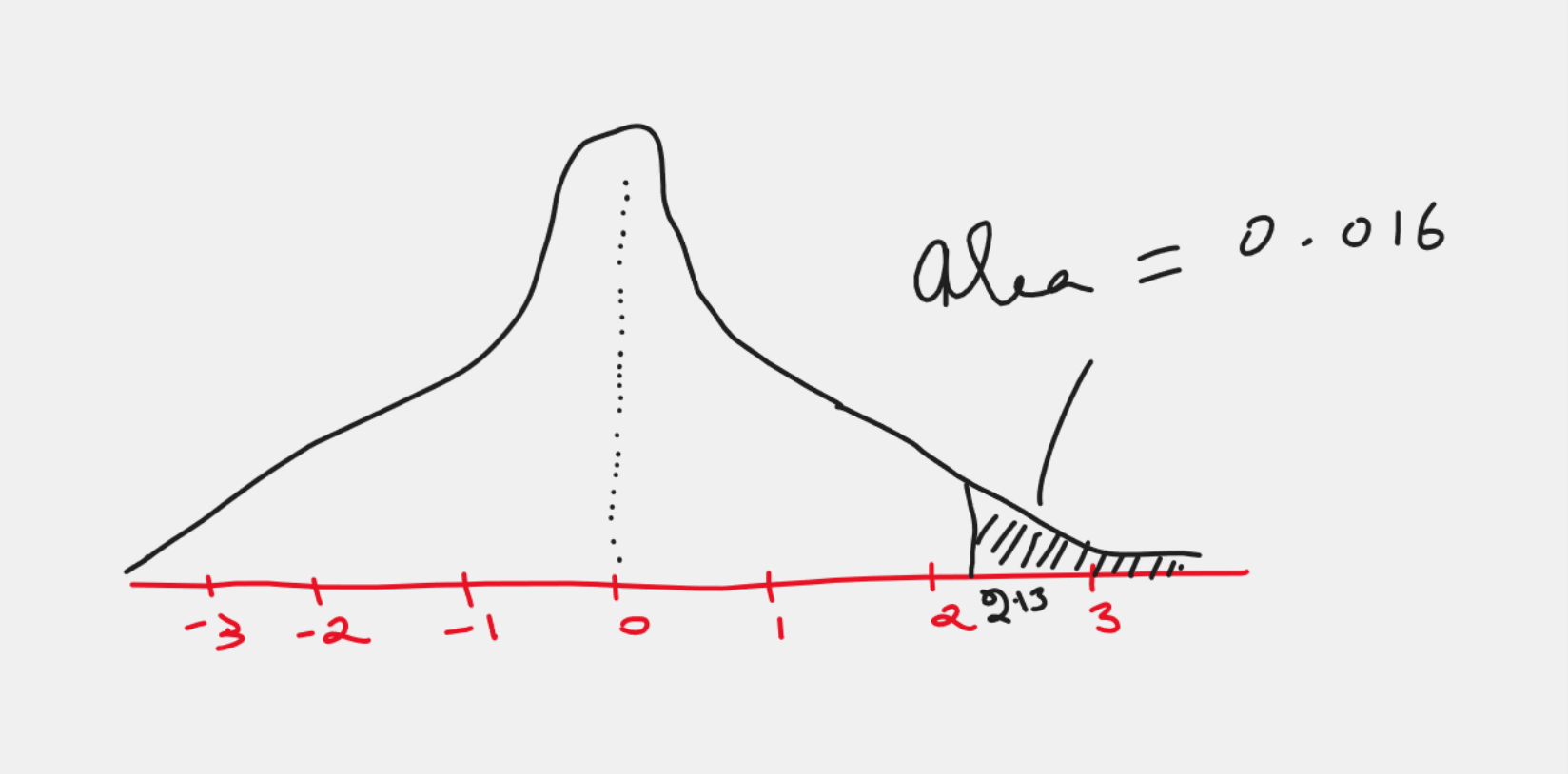
Therefore we reject Null hypothesis

Our observed value of z is 2.13 which is greater than the critical value of 1.64. We therefore reject H_0.



You can see that our p-value is p = 0.0166.

Our p-value is less than alpha (0.05). Therefore we reject H_0



Conclusion

The verbal GRE scores of applicants ($M = 212.79$) is significantly greater than 210, $z = 2.13$ $P = 0.0166$

z score > Critical Value

2.13 > 1.64

P value < α

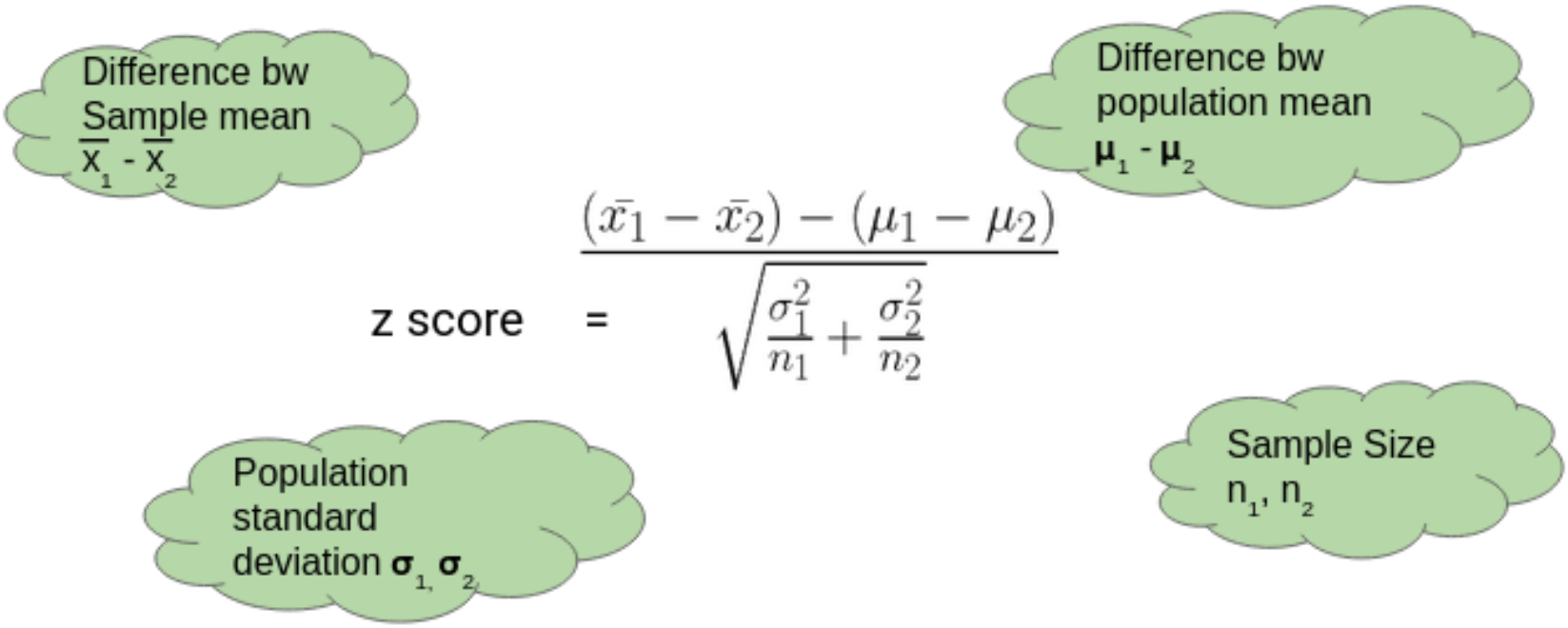
0.0166 < 0.05

$H_0 : \mu \leq 210$ ✗

$H_1 : \mu > 210$ ✓

Two Sample Z Test

We perform a Two Sample Z test when we want to compare **the mean of two samples.**



Tail

- One tailed test allow for the possibility of an effect in one direction.
- Two tailed test for the possibility of an effect in two direction.

Z Test Problem 2

- Suppose you start up a company that has developed a drug that is supposed to increase IQ. You know that the standard deviation of IQ in the general population is 15. You test your drug on 36 patients and obtain a mean IQ of 97.65. Using an alpha value of 0.05, is this IQ signicantly different than the population mean of 100?

Z value

Standard Error

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{36}} = 2.5$$

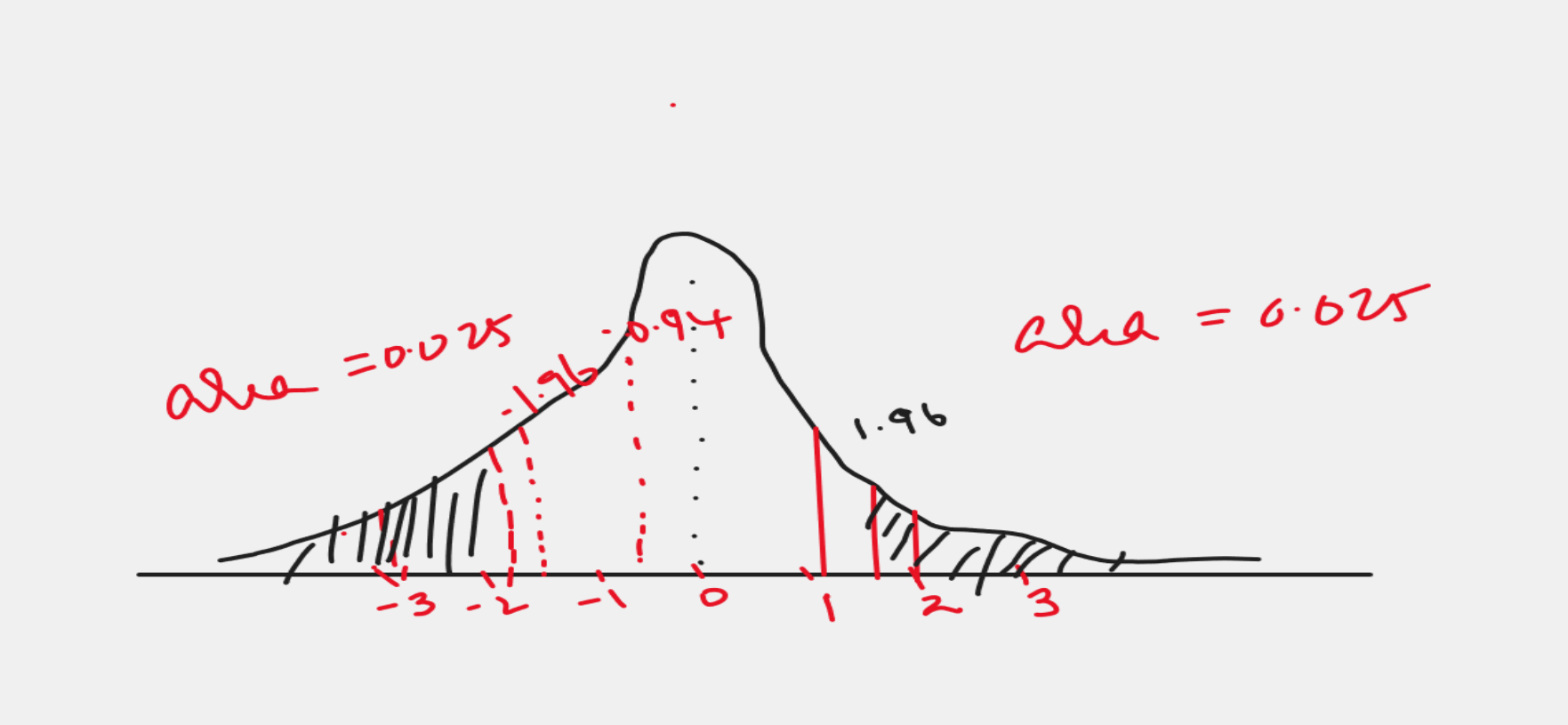
$$z = \frac{(\bar{x} - \mu)}{\sigma_{\bar{x}}} = \frac{(\bar{x} - \mu)}{\sigma_{\bar{x}}} = \frac{97.65 - 100}{2.5}$$

$$= -\frac{2.35}{2.5} = -0.94$$

- We then compare our observed value of z to the critical values of z for alpha = 0.05. We are looking for a significant difference, so this will be a two-tailed test.
- We reject the null hypothesis if our observed mean is either signicantly larger or smaller than 100.
- Our critical values of z are therefore the two values that span the middle 95% of the area under the standard normal distribution.
- This means that the areas in each of the two tails is 0.05/2 = 0.025

z	.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
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

- Which corresponds to a critical value of z = 1.96



Conclusions

Z Score < Critical value

$-0.94 < 1.96$

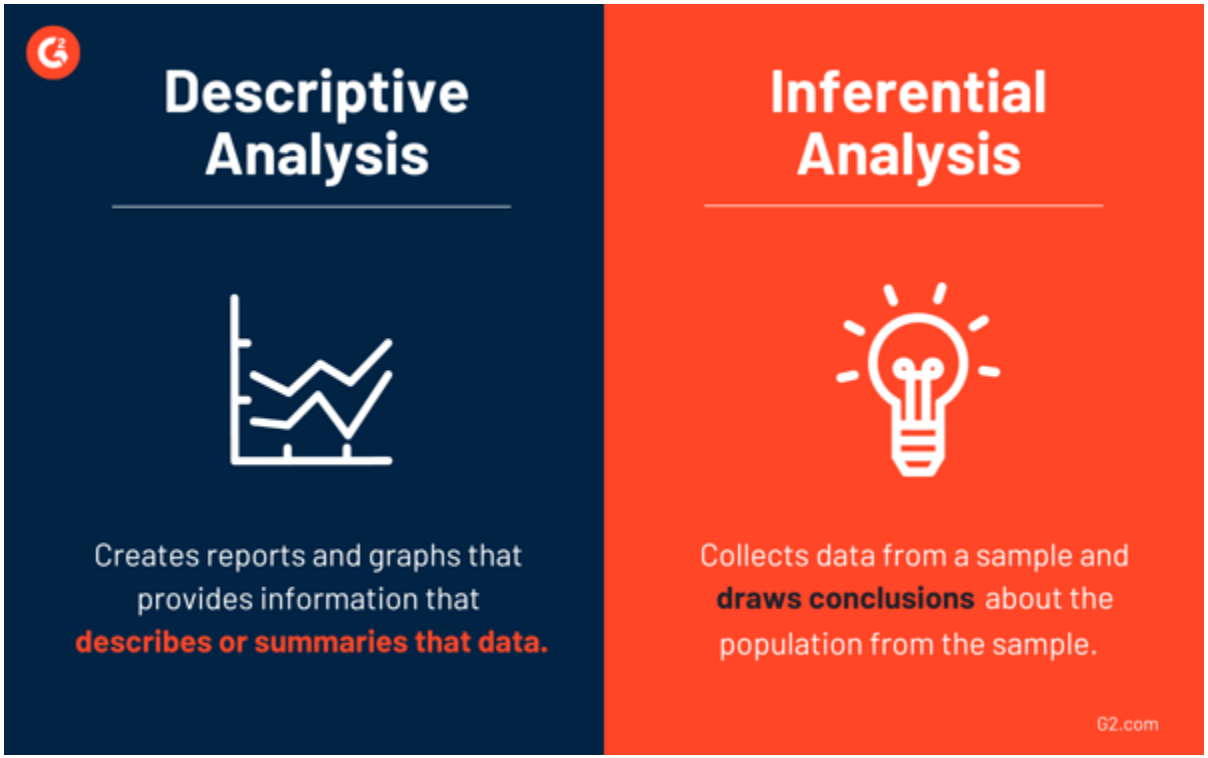
pvalue > SL

$0.8264 > 0.05$

Reject H_1

Choose H_0

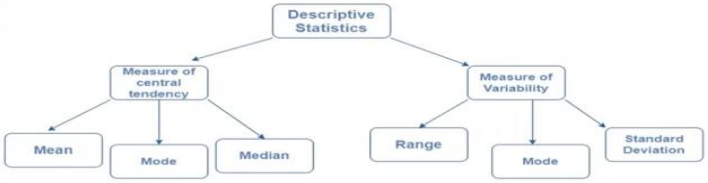
DA vs IA

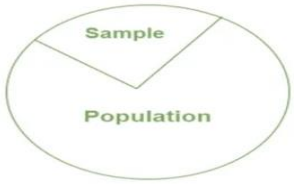

Descriptive Analysis
Creates reports and graphs that provides information that describes or summaries that data.
Inferential Analysis
Collects data from a sample and draws conclusions about the population from the sample.
G2.com

DS vs IS

DESCRIPTIVE STATISTICS Vs INFERENTIAL STATISTICS

DESCRIPTIVE	INFERENTIAL
It is the analysis of data that helps to describe, show and summarize data under study	It is the analysis of random sample of data taken from a population to describe and make inference about the population
Organize, analyze and present data in a meaningful way	Compares, test and predicts data
It is used to describe a situation	It is used to explain the chance of occurrence of an event
It explain already known data and limited to a sample or population having small size	It attempts to reach the conclusion about the population
Types: Measure of central tendency & Measure of variability	Types: Estimation of parameters & Testing of hypothesis
Results are shown with help of charts, graphs, tables etc.	Results are shown with help of probability scores





Probability Uses In Business and Calculating Probability from a Contingency Tables.

- A **contingency table** provides a way of portraying data that can facilitate calculating probabilities. The table helps in determining conditional probabilities quite easily.
- The table displays sample values in relation to two different variables that may be dependent or contingent on one another.

Example 1

Suppose a study of speeding violations and drivers who use cell phones produced the following fictional data:

	Speeding violation in the last year	No speeding violation in the last year	Total
Cell phone user	25	280	305
Not a cell phone user	45	405	450
Total	70	685	755

The total number of people in the sample is 755. The row totals are 305 and 450. The column totals are 70 and 685. Notice that 305 + 450 = 755 and 70 + 685 = 755.

Calculate the following probabilities using the table.

- Find P(Person is a car phone user)

$$\frac{Number\ of\ \text{cell}\ phone\ users = 305}{total\ numbers\ in\ study = 755}$$

- Find P(person had no violation in the last year)

$$\frac{\text{number that had no violation} = 685}{\text{total number in study} = 755}$$

Calculate the following probabilities using the table.

- Find P(Person had no violation in the last year AND was a phone user).

$$\frac{280}{755}$$

- Find P(Person is a phone user OR person had no violation in the last year).

$$\left(\frac{305}{755}+\frac{685}{755}\right)-\frac{280}{755}=\frac{710}{755}$$

Calculate the following probabilities using the table.

- Find P(Person is a phone user GIVEN person had a violation in the last year).

$\frac{25}{70}$(The sample space is reduced to the number of persons who had a violation)

- Find P(Person had no violation last year GIVEN person was not a phone user)

$\frac{405}{450}$(The sample space is reduced to the number of persons who were not phone users)

Try it

This table shows the number of athletes who stretch before exercising and how many had injuries within the past year.

	Injury in last year	No injury in last year	Total
Stretches	55	295	350
Does not stretch	231	219	450
Total	286	514	800

1. What is *P*(athlete stretches before exercising)?

   Show Answer

   P(athlete stretches before exercising) = $\frac{350}{800}$ = 0.4375

2. What is *P*(athlete stretches before exercising|no injury in the last year)?

   Show Answer

   P(athlete stretches before exercising|no injury in the last year) = $\frac{295}{514}$ = 0.5739

Example 2:

- This table shows a random sample of 100 hikers and the areas of hiking they prefer. Hiking Area Preference

Sex	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16	___	45
Male	___	___	14	55
Total	___	41	___	___

Solution:

1. Complete the table.

Show Answer

Hiking Area Preference

Sex	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16	11	45
Male	**16**	**25**	14	55
Total	**34**	41	**25**	**100**

2. Are the events "being female" and "preferring the coastline" independent events?

Hint:

Let F = being female and let C = preferring the coastline.

Check if $P(F \text{ AND } C) = P(F) * P(C)$.

If $P(F \text{ AND } C) = P(F) * P(C)$, then F and C are independent.
If $P(F \text{ AND } C) \neq P(F) * P(C)$, then F and C are not independent.

Show Answer

$$P(F \text{ AND } C) = \frac{18}{100} = 0.18$$

$$P(F) * P(C) = \left(\frac{45}{100}\right)\left(\frac{34}{100}\right) = (0.45)(0.34) = 0.153$$

$P(F \text{ AND } C) \neq P(F) * P(C)$, so the events F and C are not independent.

3. Find the probability that a person is male given that the person prefers hiking near lakes and streams.

Hint:

Let M = being male, and let L = prefers hiking near lakes and streams.

1. What word tells you this is a conditional?
2. Fill in the blanks and calculate the probability: $P(\_\_\_|\_\_\_) = \_\_\_$.
3. Is the sample space for this problem all 100 hikers? If not, what is it?

Show Answer

The word given tells you that this is a conditional.

$$P(M|L) = \frac{25}{41}$$

No, the sample space for this problem is the 41 hikers who prefer lakes and streams.

Find the probability that a person is female or prefers hiking on mountain peaks.

Hint:

Let F = being female, and let P = prefers mountain peaks.

1. Find $P(F)$.
2. Find $P(P)$.
3. Find $P(F \text{ AND } P)$.
4. Find $P(F \text{ OR } P)$.

Show Answer

The probability that a person is female or prefers hiking on mountain peaks = $\frac{59}{100}$

- $P(F) = \frac{45}{100}$
- $P(P) = \frac{25}{100}$
- $P(F \text{ AND } P) = \frac{11}{100}$
- $P(F \text{ OR } P) = \frac{45}{100} + \frac{25}{100} - \frac{11}{100} = \frac{59}{100}$

TRY IT

This table shows a random sample of 200 cyclists and the routes they prefer. Let M = males and H = hilly path.

Gender	Lake Path	Hilly Path	Wooded Path	Total
Female	45	38	27	110
Male	26	52	12	90
Total	71	90	39	200

1. Out of the males, what is the probability that the cyclist prefers a hilly path?

   Show Answer

   $P(H|M) = \frac{52}{90} = 0.5778$

2. Are the events "being male" and "preferring the hilly path" independent events?

   Show Answer

   For M and H to be independent, show $P(H|M) = P(H)$

   $P(H|M) = 0.5778$, $P(H) = \frac{90}{200} = 0.45$

   $P(H|M) \neq P(H)$, so M and H are not independent.

Example 3

Muddy Mouse lives in a cage with three doors.

If Muddy goes out the first door, the probability that he gets caught by Alissa the cat is $\frac{1}{5}$ and the probability he is not caught is $\frac{4}{5}$.

If he goes out the second door, the probability he gets caught by Alissa is $\frac{1}{4}$ and the probability he is not caught is $\frac{3}{4}$.

The probability that Alissa catches Muddy coming out of the third door is $\frac{1}{2}$ and the probability she does not catch Muddy is $\frac{1}{2}$.

It is equally likely that Muddy will choose any of the three doors so the probability of choosing each door is $\frac{1}{3}$.

Door Choice

Caught or Not	Door One	Door Two	Door Three	Total
Caught	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	____
Not Caught	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	____
Total	____	____	____	1

- The first entry $\frac{1}{15} = (\frac{1}{5})(\frac{1}{3})$ is *P*(Door One AND Caught)
- The entry $\frac{4}{15} = (\frac{4}{5})(\frac{1}{3})$ is *P*(Door One AND Not Caught)

Verify the remaining entries.

1. Complete the probability contingency table. Calculate the entries for the totals. Verify that the lower-right corner entry is 1.

Show Answer

Door Choice

Caught or Not	Door One	Door Two	Door Three	Total
Caught	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{19}{60}$
Not Caught	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	$\frac{41}{60}$
Total	$\frac{5}{15}$	$\frac{4}{12}$	$\frac{2}{16}$	1

2. What is the probability that Alissa does not catch Muddy?

Show Answer

$\frac{41}{60}$

3. What is the probability that Muddy chooses Door One OR Door Two given that Muddy is caught by Alissa?

Show Answer

$\frac{9}{19}$

Example 4

This table contains the number of crimes per 100,000 inhabitants from 2008 to 2011 in the U.S.

United States Crime Index Rates Per 100,000 Inhabitants 2008–2011

Year	Robbery	Burglary	Rape	Vehicle	Total
2008	145.7	732.1	29.7	314.7	
2009	133.1	717.7	29.1	259.2	
2010	119.3	701	27.7	239.1	
2011	113.7	702.2	26.8	229.6	
Total					

TOTAL each column and each row. Total data = 4,520.7

1. Find $P(2009 \text{ AND Robbery})$.
   Show Answer

   0.0294

2. Find $P(2010 \text{ AND Burglary})$.
   Show Answer

   0.1551

3. Find $P(2010 \text{ OR Burglary})$.
   Show Answer

   0.7165

4. Find $P(2011|\text{Rape})$.
   Show Answer

   0.2365

5. Find $P(\text{Vehicle}|2008)$.
   Show Answer

   0.2575

Try it

Try It

This table relates the weights and heights of a group of individuals participating in an observational study.

Weight/Height	Tall	Medium	Short	Totals
Obese	18	28	14	
Normal	20	51	28	
Underweight	12	25	9	
Totals				

1. Find the total for each row and column.

Show Answer

Weight/Height	Tall	Medium	Short	Totals
Obese	18	28	14	60
Normal	20	51	28	99
Underweight	12	25	9	46
Totals	50	104	51	205

2. Find the probability that a randomly chosen individual from this group is Tall.

Show Answer

$$P(\text{Tall}) = \frac{50}{205} = 0.244$$

3. Find the probability that a randomly chosen individual from this group is Obese and Tall.
   Show Answer

   P(Obese AND Tall) = $\frac{18}{205}$ = 0.088

4. Find the probability that a randomly chosen individual from this group is Tall given that the idividual is Obese.
   Show Answer

   P(Tall|Obese) = $\frac{18}{60}$ = 0.3

5. Find the probability that a randomly chosen individual from this group is Obese given that the individual is Tall.
   Show Answer

   P(Obese|Tall) = $\frac{18}{50}$ = 0.36

6. Find the probability a randomly chosen individual from this group is Tall and Underweight.
   Show Answer

   P(Tall AND Underweight = $\frac{12}{205}$ = 0.0585

7. Are the events Obese and Tall independent?
   Show Answer

   No. P(Tall) $\neq$ (Tall|Obese).


Thank you!