28-12-2020

2) Statistics & Probability basics for Data Analysis

I

Book:- statistical techniques for Business and economics McGraw Hill by D Lind

Book:- Think Stats O'reilly Allan Downey 2nd edition

Goals:

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

Statistics:-

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

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

Types of Statististics:

1) Descriptive statistics :- Method of organizing, summarizing and presenting data in an informati way.

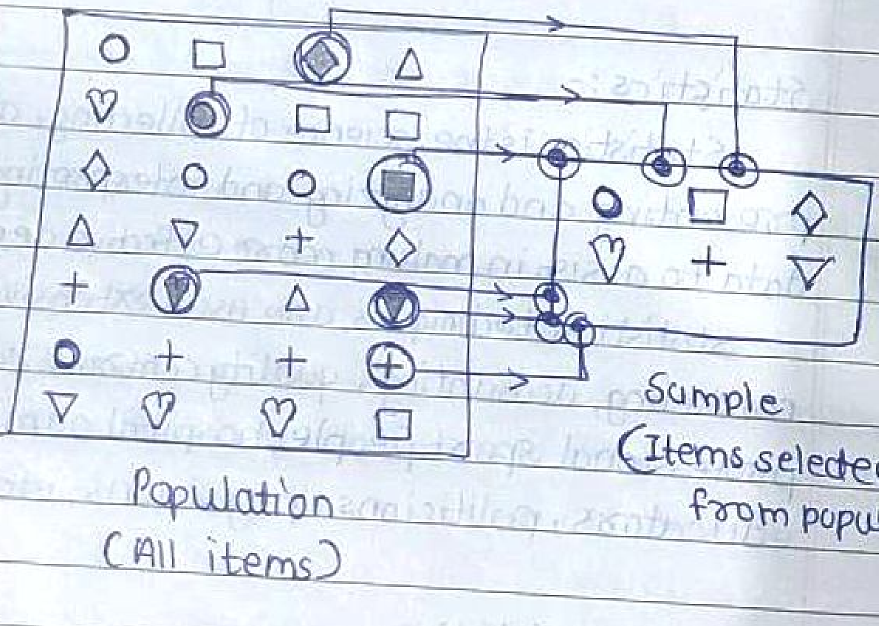
Ex:- A Gallup poll found out 49% of the people in a survey knew the name of first book of Bible. The statistic 49 describes the number out of every 100 persons ...

Ex2:- According to consumer reports, General Electric washing machine owners reported 9 problems per 100 machines during 2001. The statistic 9 describes the number of problems out of every 100 machines.

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

Population vs Sample:

- A population is a collection of all possible individuals objects or measurements of interest.
- A sample is a portion, or part, of the population of interest.



Types of variables:

1) Qualitative variable / Attribute variable - the characteristic being studied is non-numeric.

Ex:- Gender, religious affiliation, type of automobile owned, state of birth, eye color etc.

2) Quantitative variable :- information is reported numerically.

Ex :- balance in your checking account, minutes remain before class, no. of children in family, etc.

Classification of Quantative Variable :

1) Discrete variables :- can only assume certain values and there are usually gaps between values.

Ex :- no. of bedrooms in a house, no. of hammers sold at local depot (1, 2, 3, ..., etc.)

2) Continuous variables :- can assume any value within a specified range.

Ex :- The pressure in tyre, the weight of pork chop or height of students in class, etc.

Types of variables
- Qualitative
  - Brand of PC
  - Marital Status
  - Hair Color
- Quantitative
  - Discrete
    - Children in family
    - Strokes on golf course
    - TV sets owned
  - Continuous
    - Amount of tax payed
    - Weight of student
    - Yearly rainfall

Four Levels of Management:

1) Nominal Level :- data that is classified into categories and cannot be arranged in any particular order.

Ex:- eye color, gender, religious affiliation

2) Ordinal Level :- involves data arranged in some order, but the differences between data values cannot be determined or are meaningless.

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

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

Ex :- Temperature on Farenheit scale.

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

Ex:- Monthly income of surgeons, distance travelled by manufacturer's employees within a month.

Levels of management

- **Nominal**
  - Data may not be classified
    - Jersey no.s
    - Make of car
- **Ordinal**
  - Data are ranked
    - Rank in class
    - Team standings in Pac-10
- **Internal**
  - Meaningful difference between values
    - Temperature
    - Dress size
- **Ratio**
  - Meaningful 0 point and ratio between values
    - No. of patient seen
    - No. of calls made
    - Distance to class

* * * * *

II Frequency Tables:

Frequency Tables is a grouping of qualitative data into mutually exclusive classes showing the number of observation in each class.

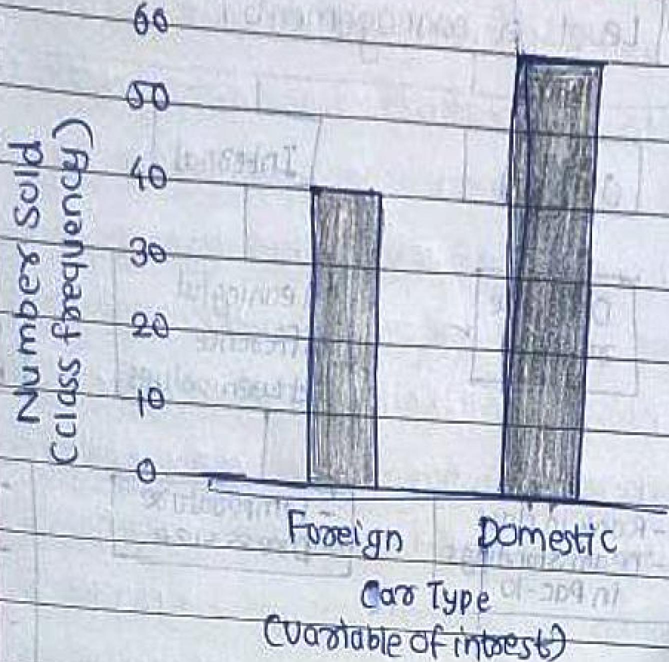
Ex:- Frequency Table for vehicles sold at Autoplex:

Car Type	No. of cars
Domestic	50
Foreign	30

$(\text{Domestic}) \cap (\text{Foreign}) = \phi$

Bar Charts:-

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



Pie Charts:-

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

Frequency Distribution:

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

Selling Prices ($ Thousands)	Frequency
15 up to 18	8
18 up to 21	23
21 upto 24	17
24 upto 27	18
27 upto 30	8
30 upto 33	4
33 upto 37	2
	80

3-12-2020

Relative Class Frequencies:

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

Vehicle Type	Number sold	Relative frequency	
Domestic	50	0.625	50/80
Foreign	30	0.375	30/80
	80	1.000	

Frequency Distribution:

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

Class Frequency :- The number of observations in each class.

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

Constructing a Frequency Table :-

⟹ Step 1 :- Deceide on number of classes.

- 2^k rule
- A useful receipe to determine the number of classes (k) is "2 to the k rule" such that $2^k > n$.

Ex :- There were 80 vehicles sold. So, n = 80. If we try k = 6, which means we could 6 classes, then $2^6 = 64$, somewhat less than 80, Hence 6 is not enough.

If we let k = 7, then $2^7 = 128$, which is greater than 80; So the recommended number of classes is 7.

⟹ Step 2: Determine the class interval or width.

Formula : $i \geq (H-L)/K$

where, i = class interval
H = highest observed value
L = lowest observed value
K = no. of classes

Ex:- $(35,925 - 15,546)/7 = 2911$

Round up to some convenient number, such as multiple of 10 or 100. Use of class of width 3000.

⟹ Step 3: Set the individual class limits

15000 to 18000
18000 to 21000
21000 to 24000
24000 to 27000
27000 to 30000
30000 to 33000
33000 to 36000

⟹ Step 4 :- Tally all the 80 entries into the classes. |卌 ||

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

Relative Frequency Distribution:

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

Ex:-

Selling Price ($ thousands)	Frequency	Relative Frequency	Found by
15 up to 18	8	0.1000 ←	8/80
18 up to 21	23	0.2875	23/80
21 up to 24	17	0.2125	17/80
24 upto 27	18	0.2250	18/80
27 upto 30	8	0.1000	8/80
30 up to 33	4	0.0500	4/80
33 up to 36	2	0.0250	2/80
	80	1.0000	

Graphic Presentation of a Frequency Distribution:

The three commonly used graphical forms are:

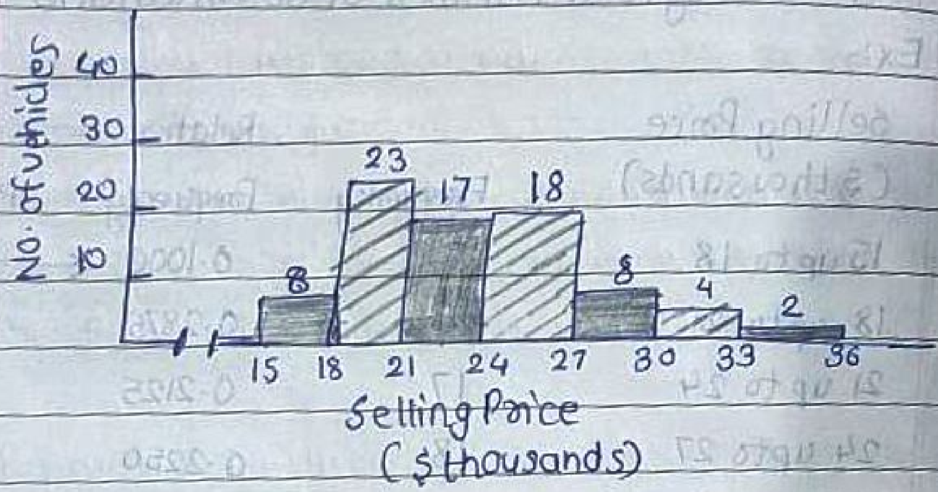
↳ Histograms

↳ Frequency Polygons

↳ Cummulative Frequency Distributions.

Histogram:

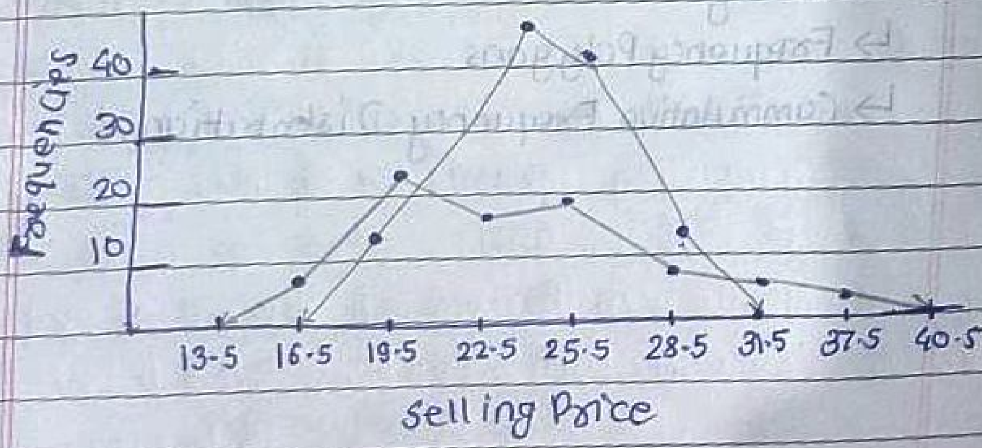
A graph in which the classes are marked on the horizontal axis and the class frequencies on the vertical axis. The class frequencies are represented by heights of the bars and the bars are drawn adjacent to each other.



Frequency Polygon:

A Frequency Polygon also shows the shape of a distribution and is similar to histogram.

It consists of line segments connecting the points formed by the intersections of the class midpoints and the class frequencies.



Histogram vs Frequency Polygon :

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

Cummulative Frequency Distribution :

Selling Price ($ thousands)	Frequency	Cummulative Frequency	Found By
15 upto 18	8	8	
18 upto 21	23	31	8+23
21 upto 24	17	48	8+23+17
24 upto 27	18	66	:
27 upto 30	8	74	:
30 upto 33	4	78	:
33 upto 36	2	80	:

*****

III

Numerical Measures:

Mean:

Characteristics of Mean:

- The Arithmetic Mean is most widely used measure of location
- Requires the interval scale.
- Major characteristics
  - ↳ All values are used.
  - ↳ It is unique.
  - ↳ The sum of deviations from mean is 0.
  - ↳ It is calculated by summing the values and dividing by the number of values.

Population Mean:

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

$$\boxed{\mu = \frac{\Sigma X}{N}}$$

where, μ = population mean. Greek lowercase for "mu"

N = number of values in population.

X = any particular value

Σ = indicates operation of adding. Greek capital letter for "sigma".

ΣX = sum of X values in population.

Ex:- Is the above information a sample or a population? What is the arithmetic mean number of patents granted?

Company	No. of patents granted	Company	No. of patents granted
General Motors	511	Mazda	210
Nissan	385	Chrysler	97
Diamler Chrysler	275	Porshe	50
Toyota	257	Mitsubishi	36
Honda	249	Volvo	23
Ford	234	BMW	13 → -182

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

Mean

Parameter vs Statistics:

- A parameter is measurable characteristic of population
- A statistic is measurable characteristic of sample.

Properties of Arithmetic Mean:

1) Every set of interval level and ratio level data has a mean.
2) All the values are included in computing the mean.
3) The mean is unique.
4) The sum of deviations of each value from the mean is zero.

Sample Mean:

For ungrouped ~~mean~~ data, the sample mean is the sum of all sample values divided by number of sample values.

Sample Mean $\bar{x} = \frac{\Sigma X}{n}$

$\bar{x}$ = Sample mean, read as "x bar"

n = no. of values in sample

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

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

What is arithmetic mean number of minutes used?

$$\text{Sample mean} = \frac{\text{Sum of all values in sample}}{\text{Number of all values in sample}}$$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{90+77+\dots+113+83}{12} = \frac{1170}{12} = 97.5$$

Weighted Mean:

The weighted mean of a set of number $x_1, x_2, \dots, x_n$, with corresponding weights $W1, W2, \dots, Wn$, is computed from the following formula.

$$\text{Weighted Mean} = \bar{x}_w = \frac{W_1x_1 + W_2x_2 + W_3x_3 + \dots + W_nx_n}{W_1 + W_2 + W_3 + \dots + W_n}$$

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

$$\bar{x}_w = \frac{14(\$16.50)+10(\$19.00)+2(\$25.00)}{14+10+2} = \frac{\$471.00}{26}$$

$$= \$18.1154$$

Median:

Median is the midpoint of the values after they have been ordered from smallest to largest, or the largest to the smallest.

Properties of median:

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

Ex:- The ages for a sample of five collage students are
21, 25, 19, 20, 22

Arranging the data in ascending order.
19, 20, 21, 22, 25

Thus median = 21

If n (no. of observations) is odd; $\left(\frac{n+1}{2}\right)^{th}$ observation is median.

If n is even, $\dfrac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n+1}{2}\right)^{th}}{2}$

Ex:- The heights of four basketball players in inches are:

76, 73, 80, 75

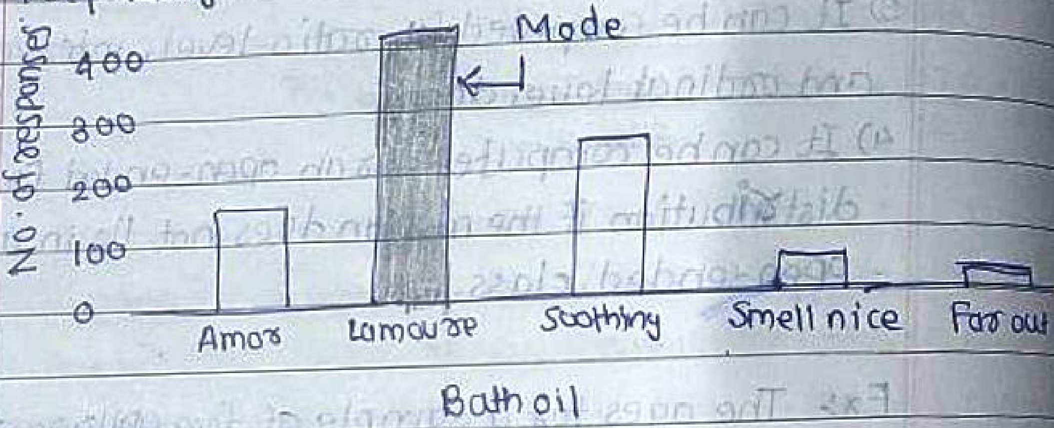
Arranging data in ascending order,

73, 75, 76, 80

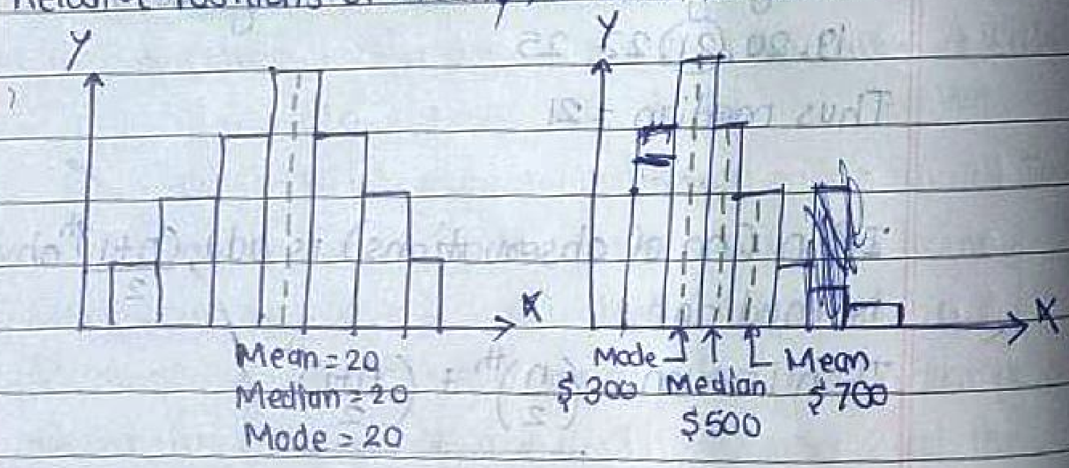
Thus median = 75.5

Mode:

Mode is the value of observation that occurs most frequently.



Relative Positions of Mean, Median, Mode

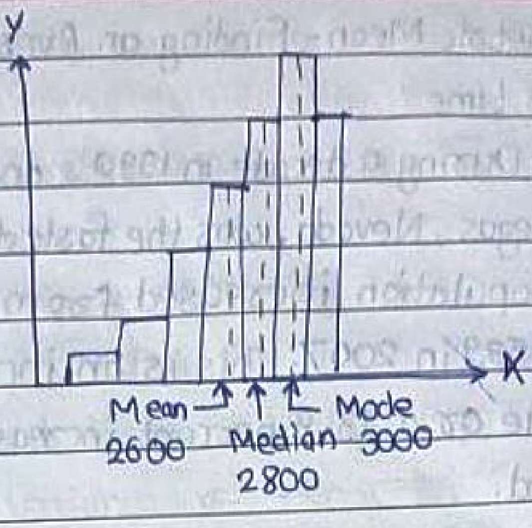


zero skewness

mode = median = mode

positive skewness

mode < median < mode



negative skewness

mode > median > mode

2-2020

The Geometric Mean :-

$$\text{Geometric Mean} = \sqrt[n]{(x_1)(x_2)\ldots(X_n)}$$

- Useful in finding the average change in percentages, ratios, indexes or growth rates over time.
- It has wide application in business and economics because we are often intrested in finding the changes in percentage in sales, salaries or economic figures. such as GDP, which compound on build on each other.

* - The geometric mean will always be less than or equal to the arithmetic mean. $GM \leq AM$

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

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

Geometric Mean- Finding an Average Percent change over time:

Ex:- During a decade in 1990's and into the 2000's, Las Vegas, Nevada, was the fastest growing city in the US. The population increased from 258,295 in 1990 to 552,539 in 2007. This is an increase of 294,244 people or a 13.9 percent increase over the 17-year period.

Average percent increase over time $GM = \sqrt[n]{\frac{\text{Value at the end of period}}{\text{Value at the start of period}}} - 1$

$$\therefore GM = \sqrt[17]{\frac{552,539}{258,295}} - 1$$

$$\therefore GM = 1.0457 - 1$$

$$\therefore GM = 0.0457 \rightarrow 4.57\% \text{ Annual increase}$$

Dispersion:

A measure of location, such as the mean or the median, only describes the center of the data. It is valueable from that standpoint, but it does not tell us anything about the spread of data.

Ex:- If your nature guide told you that river ahead averaged 3 feet in depth. Would you want to wade accross on foot without additional information?

Probably not. You would want to know something about the variation in the depth.

A second reason to learn dispersion in a set of data is to compare the spread in two or more distributions

Measures of Dispersion:

1) Range

Range = Largest value − Smallest value

2) Mean Deviation

$$\text{Mean Deviation (MD)} = \frac{\Sigma |x - \bar{x}|}{n}$$

3) Variance and Standard deviation

Population variance ~~$\sigma^2 = \Sigma (x - \mu)^2$~~

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

$$\text{Population Standard Deviation} \quad \sigma = \sqrt{\frac{\Sigma (x - \mu)^2}{N}}$$

$$\text{Standard deviation (S.D)} = \sqrt{\text{Variance}}$$

Ex:- Range

The no. of cappuchinos sold at the starbucks location in the Orange Country Airport between 4 and 7 pm for a sample of 5 days last year were 20, 40, 50, 60 and 80. Determine the range for the no. of cappuchinos sold.

∴ Range = Largest value − Smallest value

= 80 − 20 = 60

Mean Deviation:

Mean Deviation is the arithmetic mean of the absolute values of the deviations from the arithmetic mean. It measures the mean amount by which the values in a population, or sample, vary from their mean.

Ex:- Consider the same example used for Range

Step 1 :- Compute the mean

$$\bar{x} = \frac{\Sigma x}{n} = \frac{20+40+50+60+80}{5} = 50$$

Step 2 :- Subtract the mean (50) from each of the observations, convert to positive if difference is negative

Step 3 :- Sum the absolute differences found in step 2 then divide the number by number of observations.

No. of cappuchinos sold daily	$(x-\bar{x})$	Absolute Deviation
20	$(20-50) = -30$	30
40	$(40-50) = -10$	10
50	$(50-50) = 0$	0
60	$(60-50) = 10$	10
80	$(80-50) = 30$	30
	Total:	80

$$M.D = \frac{\Sigma|x-\bar{x}|}{n} = \frac{80}{5} = 16$$

Variance and Standard Deviation

Variance is the arithmetic mean of the squared deviations from the mean

Standard deviation is the square root of mean

- The variance and standard deviations are non-negative and are zero only if all observations are the same.
- For populations whose values are near the mean, the variance and standard deviation will be small.
- For populations whose values are dispersed from the mean, the population variance and standard deviation will be large.
- The variance overcomes the weakness of the range by using all the values in the population.

Variance - Formula and Computation

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

σ^2 = variance X = value of observation
μ = arithmetic mean of population
N = number of observations in population

Step 1: Find the mean
Step 2: Find the difference between each observations and the mean, and square that difference.
Step 3: Sum all the squared differences found in step 2.
Step 4: Divide the sum of the squared differences by the number of items in the population.

Ex:- The number of traffic citations issued during the last five months in Beaufort County, South Carolina is reported below:

Months	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Citations	19	17	22	18	28	34	45	39	38	44	34	10

Step1: Find the mean.

$$\mu = \frac{\Sigma x}{N} = \frac{19+17+\ldots+34+10}{12} = \frac{348}{12} = \underline{\underline{29}}$$

Step2:

Month	Citations (X)	$X-\mu$	$(X-\mu)^2$
Jan	19	-10	100
Feb	17	-12	144
Mar	22	-7	49
Apr	18	-11	121
May	28	-1	1
Jun	34	5	25
Jul	45	16	256
Aug	39	10	100
Sep	38	9	81
Oct	44	15	225
Nov	34	5	25
Dec	10	-19	361
	348	0	1488

Step3:

$$\sigma^2 = \frac{\Sigma (X-\mu)^2}{N} = \frac{1488}{12} = 124$$

Sample Variance

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

S^2 = sample variance X = value of each observation
$\bar{X}$ = mean of the sample
n = no. of observations in sample.

Ex:- The hourly wages for a sample of part time employees at Home Depot are

\$12, \$20, \$16, \$18, \$19.

What is sample variance.

$$\bar{X} = \frac{12+20+16+18+19}{5} = 17$$

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

Hourly wage (x)	$x-\bar{x}$	$(x-\bar{x})^2$
12	-5	25
20	3	9
16	-1	1
18	1	1
19	2	4
\$85	0	40

$$S^2 = \frac{40}{5-1} = \frac{40}{4} = 10$$

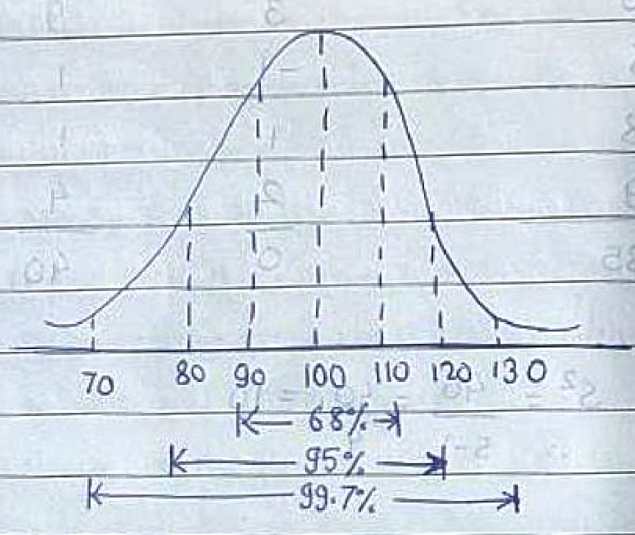
Sample Standard Deviation:

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

Chebyshev's Theorem:- For any set of observations (sample or population), the proportion of the values that lie within 'k' standard deviations of the means is at least $1 - \frac{1}{k^2}$, where k = any constant > 1.

The Empirical rule :-

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



Empirical Rule

The Arithmetic mean of grouped data :

$$\bar{X} = \frac{\Sigma fM}{n}$$

$\bar{X}$ = sample mean   M = midpoint of each class

f = frequency of each class

fM = frequency in each class times the midpoint of the class

ΣfM = sum of these products

n = total no. of frequencies.

Ex:-

Selling Price	Frequency (f)	Midpoint (M)	fM
15 up to 18	8	16.5	132.0
18 up to 21	23	19.5	448.5
21 up to 24	17	22.5	382.5
24 up to 27	18	25.5	459.0
27 upto 30	8	28.5	228.0
30 upto 33	4	31.5	126.0
33 upto 36	2	34.5	69.0
	80		1845.0

$$\bar{X} = \frac{\Sigma fM}{n} = \frac{1845}{80} = 23.1$$

Standard Deviation of Grouped Data:

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

S = sample standard deviation
M = Midpoint of the class
f = class frequency
n = no. of observations in sample
$\bar{X}$ = sample mean

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

$$= \sqrt{\frac{1531.8}{80-1}}$$

$$= 4.403$$

$(M-\bar{X})$	$(M-\bar{X})^2$	$f(M-\bar{X})^2$
−6.6	43.56	348.48
−3.6	12.96	298.08
−0.6	0.36	6.12
2.4	5.76	103.68
5.4	29.16	233.28
8.4	70.56	282.24
11.4	129.96	259.92

***
**

$\Sigma f(M-\bar{X})^2 = 1531.80$

Probability:

A probability is a measure of the likelihood that an event in future will happen. It can only assume a value between 0 and 1.

A value near zero means, the event is not likely to happen. A value near one means it is likely.

There are three ways of assigning probability:
- ↳ classical
- ↳ empirical
- ↳ subjective

Experiment:

An experiment is the observation of some activity or the act of taking some measurement.

Outcome:

An outcome is the particular result of an experiment.

Events:

An event is a collection of one or more outcomes of an experiment.

Classical Probability:

Given by, $\text{Probability of an event} = \dfrac{\text{No. of favourable outcomes}}{\text{Total no. of possible outcomes}}$

Mutually exclusive events:

Events are mutually exclusive if the occurrance of any one event means that none of the others can occur at the same time.

Events are independent if occurance of one event does not affect the occurrance of another

Independent events can occur simultaneously.

Collectively Exhaustive Events:

Events are collectively exhaustive if at least one of the events must occur when the experiment is being conducted.

Empirical Probability:

The probability of an event happening is the fraction of the time similar events happened in past.

The empirical probability is based on law of large numbers. The key to establishing probability empirically is that more observations will provide a more accurate estimate of probability.

Law of Large Numbers:

Over a large number of trials, the empirical probability of an event will approach its true probability

Ex:- On Feb 1, 2003, the space shuttle Columbia exploded. This was 2nd disaster in 113 space missions for NASA. On the basis of this information, what is the probability that future mission is successfully completed?

Probability of successful flight $= \frac{\text{No. of successful flights}}{\text{Total no. of flights}}$

$= \frac{111}{113} = 0.98$

Subjective Probability:

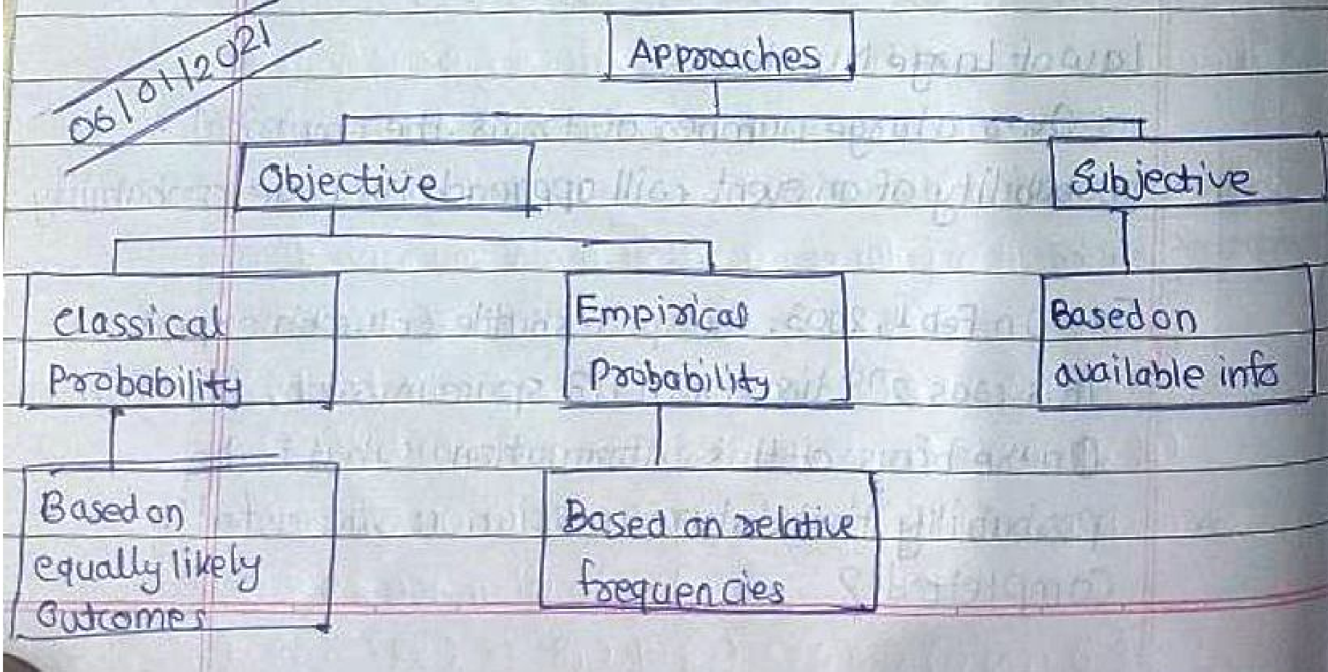
The probability of particular event happening that is assigned by an individual based on whatever information is available.

If there are little or no past experience or information on which to base a probability, it may be arrived at subjectively.

Ex:-

1) ~~Illustrating~~ Estimating the likelihood of New England Patriots will play in Super Bowl next year.
2) Estimating the likelihood you will be married before the age of 30.
3) Estimating the likelihood the U.S budget deficit will be reduced by half in the next 10 years.

06/01/2021



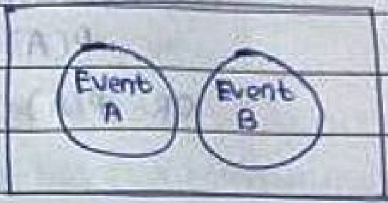
or → $\cup$
and → $\cap$

Rules for Computing Probability:

- Rules for Addition:

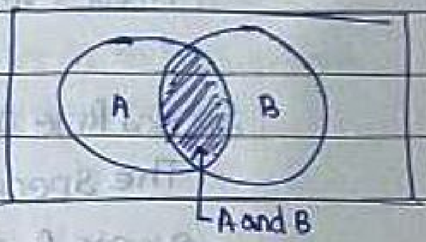
↳ Special rule of addition - if two events A and B are mutually exclusive, the probability of one or the other event's occuring equals the sum of their probabilities.

$$P(A \text{ or } B) = P(A) + P(B)$$



↳ The general rule of addition :-

If A and B are two events that are not mutually exclusive, then P (A or B) is given by the following formula



$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

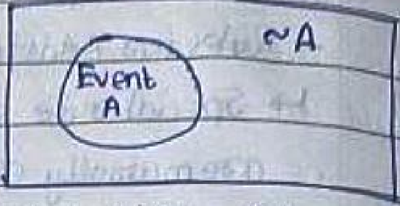
Ex:- What is the probability that a card chosen at random from a standard deck of cards will be either a king or a heart?

Card	Probability
King	$P(A) = 4/52$
Heart	$P(B) = 13/52$
King of Heart	$P(A \text{ and } B) = 1/52$

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

$$= \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52} \rightarrow 0.3077$$

The Complement Rule :



The complement rule is used to determine the probability of an event occurring by subtracting the probability of the event not occurring from 1.

$$P(A) + P(\sim A) = 1$$

$$\text{OR} \quad P(A) = 1 - P(\sim A)$$

Joint Probability - Venn Diagram

- A probability that measures the likelihood two or more events will happen concurrently.

Special Rule of Multiplication

- The special rule of multiplication requires that two events A and B are independent.
- Two events A and B are independent if the occurance of one has no effect on the probability of the occurance of the other.

$$P(A \text{ and } B) = P(A)\,P(B)$$

Ex:- A survey by the American Automobile Association revealed 60% of its members made airline reservations last year. Two members are selected at random. What is the probability both made airline reservations last year?

→ P(First member made reservation) = 0.60 $P(R_1)$

P(Second member made reservation) = 0.60 $P(R_2)$

Since, the no. of members is large, you may assume R_1 & R_2 are independent.

$P(R_1 \text{ and } R_2) = P(R_1).P(R_2) = (0.60).(0.60) = \underline{\underline{0.36}}$

Conditional Probability:

- A conditional probability is the probability of a particular event occurring, given that another event has occurred.
- The probability of the event A given that the event B has occurred is written $P(A \mid B)$

General Multiplication Rule:

- Used to find the joint probability that two events will occur.
- Use it to find the joint probability of two events when the events are not independent.
- States that for two events will happen is found by multiplying the probability that event A will happen by the conditional probability of event B occurring given that A has occurred.

$$P(A \text{ and } B) = P(A)\,P(B \mid A)$$

Ex:- A golfer has 12 golf shirts. Suppose 9 of these shirts are white and others are blue. He gets dressed in dark, so he grabs a shirt and put it on. He plays golf two days in a row and does not do laundry.

What is likelihood that both shirts selected are white?

→ Event that first shirt selected is white W_1 $\therefore P(W_1) = 9/12$

- The event that second shirt is also white W_2
- The conditional probability that the second shirt selected is white, given that 1^{st} shirt is also white $\therefore P(W_2 \mid W_1) = 8/11$
- To determine probability of two white shirts, we use

$$P(A \text{ and } B) = P(A).P(B \mid A)$$

$$P(W_1 \text{ and } W_2) = P(W_1).\,P(W_2 \mid W_1)$$

$$= \frac{9}{12} \times \frac{8}{11} = \frac{72}{132} = 0.55$$

Contingency Tables:

A contingency table is a table used to classify sample observations according to two or more identifiable characteristics.

Ex:- A survey of 150 adults classified each as to gender and the number of movies attended last month. Each respondent is classified according to two criteria - the number of movies attended and gender.

Movies Attended	Gender: Men	Gender: Women	Total
0	20	40	60
1	40	30	70
2 or more	10	10	20
Total	70	80	150

07/01/2020

Ex:-

	Loyalty	Less than 1 year B_1	1-5 years B_2	6-10 years B_3	More than 10 years B_4	Total
A_1	Would Remain	10	30	5	75	120
A_2	Would not Remain	25	15	10	30	80
		35	45	15	105	200

(Column heading: Length of Service)

What is the probability of selecting a random executive who is loyal to the company (would remain) and who has more than 10 years of service?

Event A_1, happens if selected executive is loyal

$P(A_1) = 120/200 = \underline{0.6}$

Event B_4, happens if has more 10 years of service

Thus $P(B_4|A_1)$ is required here.

$P(B_4|A_1) = 75/120$

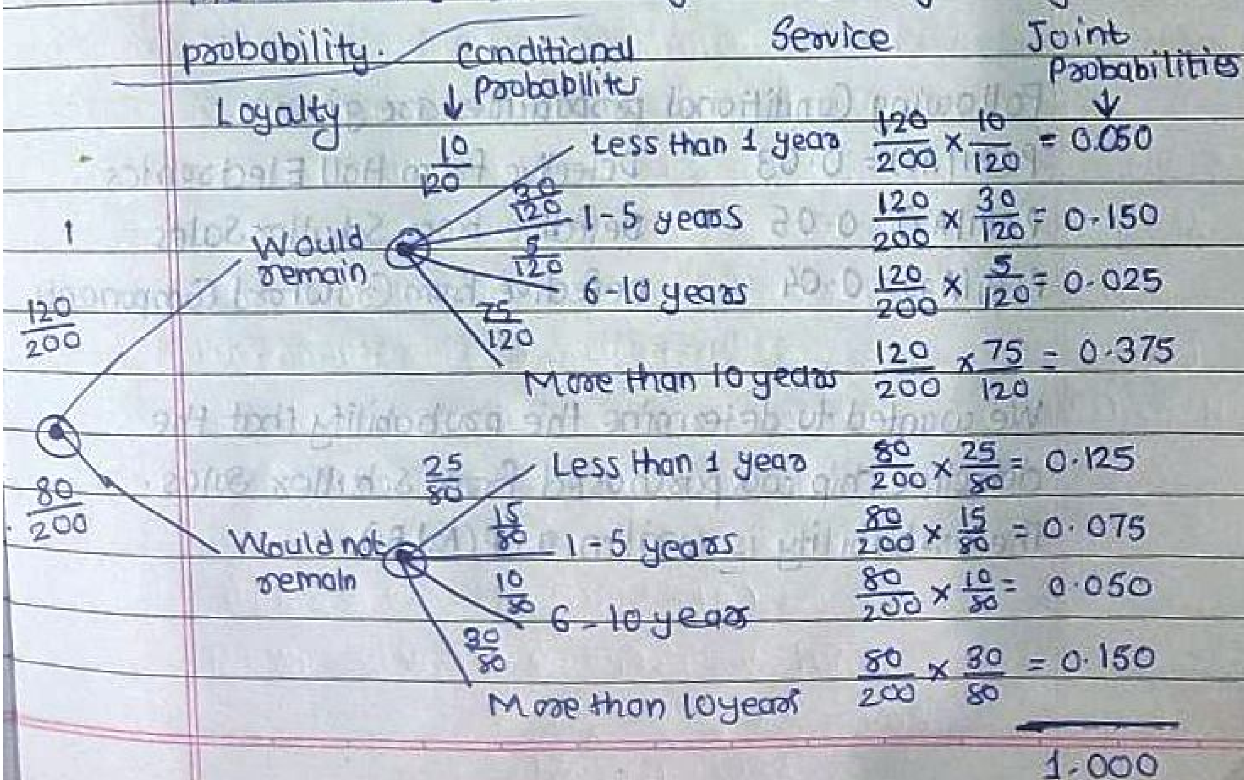
$$P(A_1 \text{ and } B_4) = P(A_1)P(B_4|A_1) = \left(\frac{120}{200}\right)\left(\frac{75}{120}\right) = \frac{9000}{24000}$$

$$= \underline{0.375}$$

Tree Diagrams:

A tree diagram is useful portraying conditional and joint probabilities. It is particularly useful for analysing business decisions involving several stages.

A tree diagram is a graph that is useful in organising calculations that involve several stages. Each segment in the tree is one stage of the problem. The branches of a tree diagram are weighted by probability.



Bayes' Theorem:

- Bayes' Theorem is a method for revising a probability given additional information.
- It is computing using the following formula:

$$P(A_i|B) = \frac{P(A_i)\,P(B|A_i)}{P(A_1)(B|A_1) + P(A_2)P(B|A_2)}$$

Ex:-

Three events (A_1, A_2, A_3) are collectively exhaustive

$A_1 \rightarrow$ LS-24 was purchased from Hall Electronics

$A_2 \rightarrow$ LS-24 was purchased from Schuller Sales

$A_3 \rightarrow$ LS-24 was purchased from CrawFord Components.

$\therefore P(A_1) = 0.30$, $P(A_2) = 0.20$, $P(A_3) = 0.50$

Additional information can be either

$B_1 \rightarrow$ The LS-24 appears defective

$B_2 \rightarrow$ The LS-24 appears not to be defective.

Following Conditional probabilities are given

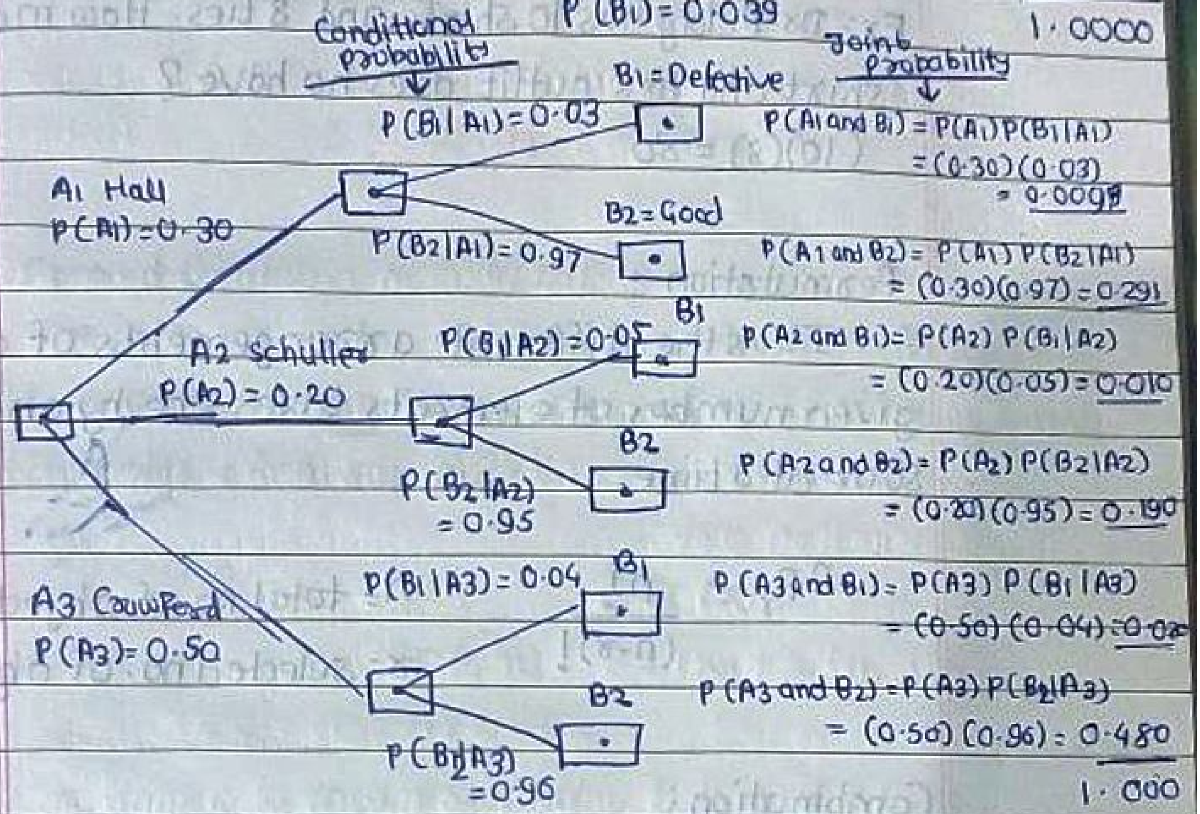
$P(B_1|A_1) = 0.03$ Defective from Hall Electronics

$P(B_1|A_2) = 0.05$ Defective from Schuller Sales

$P(B_1|A_3) = 0.04$ Defective from CrawFord Components

We wanted to determine the probability that the defective chip was purchased from Schuller Sales. The probability is written as $P(A_2|B_1)$

Event (A_i)	Prior Probability $P(A_i)$	Conditional Probability $P(B_1 \mid A_i)$	Joint Probability $P(A_i \text{ and } B_1)$	Posterior Probability $P(A_i \mid B_1)$
Hall	0.30	0.03	0.009	(0.009/0.039) = 0.2308
Schuller	0.20	0.05	0.010	(0.010/0.039) = 0.2564
Crawford	0.50	0.04	0.020	(0.020/0.039) = 0.5128
			$P(B_1) = 0.039$	1.0000



The probability that the defective LS-24 chip came from Schuller sales can be formally found by using Bayes' Theorem. We compute $P(A_2 \mid B_1)$ where,

$A_2 \rightarrow$ Schuller sales $\quad B_1 \rightarrow$ Defective

$$P(A_2 \mid B_1) = \frac{P(A_2)\,P(B_1 \mid A_2)}{P(A_1)\,P(B_1 \mid A_1) + P(A_2)\,P(B_1 \mid A_2) + P(A_3)(B_1 \mid A_3)}$$

$$= \frac{(0.20)(0.05)}{(0.30)(0.03) + (0.20)(0.05) + (0.50)(0.04)}$$

$$= \frac{0.010}{0.039} = \underline{0.2564}$$

8-01-2020

Count Rules - Multiplication

The multiplication formula indicates that if there are m ways of doing one thing and n ways of doing another thing, there are m×n ways of doing both.

Ex: Dr. Delong has 10 shirts and 8 ties. How many shirt and ties outfit does he have?

$(10)(8) = 80$

Permutation:

It is the different arrangements of a given number of elements taken one by one, or some at a time.

$$^{n}P_{r} = \frac{n!}{(n-r)!}$$

n = total no. of objects
r = selected no. of objects.

Combination:

It is the number of ways to choose 'r' objects from a group of 'n' objects without regard to order.

$$^{n}C_{r} = \frac{n!}{r!(n-r)!}$$

n = total no. of objects
r = selected no. of objects

Ex:- Combination.

There are 12 players on the Carolina Forest High Forest School. Coach must pick 5 players among them to compromise the starting line up. How many different groups are possible?

$$^{12}C_5 = \frac{12!}{5!\,(12-5)!} = \underline{\underline{792}}$$

Permutation:

Suppose that in addition, to selecting a group of 5, he must also rank each player in the starting lineup according to their ability.

$$^{12}P_5 = \frac{12!}{(12-5)!} = \underline{95040}$$

* * * * *

Discrete Probability Distributions

- A Random Variable is a numerical description of the outcome of an experiment.
- A Discrete Random Variable may assume either a finite number of values or an infinite sequence of values. Ex:- Family Size
- A Continuous Random Variable may assume any numerical value in an interval or collection of intervals. Ex:- Distance from Home to store

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

Probability Distribution

- Discrete Probability Distribution
- Continuous Probabity Distribution

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

The required conditions for a discrete probability function are:

i) $f(x) \geq 0$

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

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

$$E(x) = \mu = \sum x f(x)$$

The Variance, summarizes the variability in the values of a random variable.

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

The standard deviation, σ, is defined as the positive square root on variance.

Binomial Distribution:

① The experiment consists of a sequence of n identical trails.

② Two outcomes, sucess and faliure are possible on each trial.

③ The probability of sucess, denoted by p, does not change from trail to trial

④ The trials are independent.

29-01-2020

Our intrest is in the number of sucesses occurring in n trials.

We let x denote the number of sucesses occurring in the n trials.

Binomial Probability Function:

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

$\rightarrow p \times p \times p \ldots \times p$ (x times)

$f(x)$ = probability of x sucesses in 'n' trials.

n = the number of trials

p = the probability of success on any one trial

$1-p$ = the probability of faliure

Ex:- In recent years, management has seen a turnover of 10% of the hourly employees annually. Thus, for any hourly employee chosen at random, management estimates a probability of 0.1 that the person will not be with the company next year.

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

Let $p = 0.10$, $n = 3$, $x = 1$

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

$$f(1) = \frac{3!}{1!\,(3-1)!} (0.1)^{1} (0.9)^{2}$$

$$= 3\,(0.1)(0.81)$$

$$= \underline{\underline{0.243}}$$

— Expected value

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

$= 3\,(0.1) = \underline{\underline{0.3}}$ employees out of 3

— Variance

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

$$= 3(0.1)(0.9) = \underline{0.27}$$

— Standard Deviation

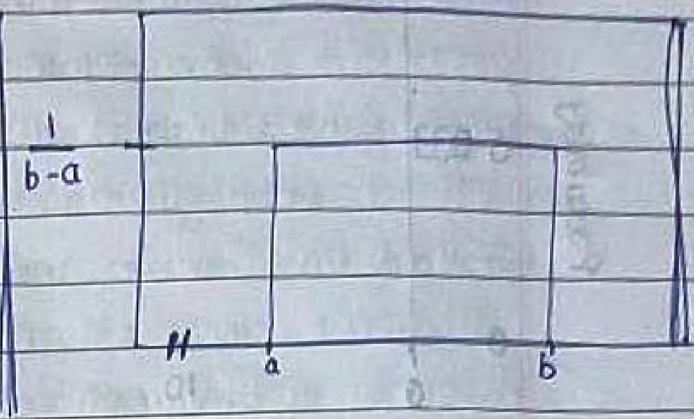
$$\sigma = \sqrt{np\,(1-p)}$$

$$= \sqrt{3\,(0.1)(0.9)} = \underline{0.52}$$

* * * * *

I] The Uniform Distribution:

- The uniform distribution is perhaps the simplest distribution for a continuous random variable.



- The distribution is rectangular in shape and is defined by minimum and maximum values.

Mean of Uniform Distribution : $\mu = \frac{a+b}{2}$

S.D of Uniform Distribution : $\sigma = \sqrt{\frac{(b-a)^2}{12}}$

Uniform Distribution : $P(x) = \frac{1}{b-a}$

if $a \le x \le b$ and 0 elsewhere

Ex:- The time that a student waits is uniformly distributed from 0 to 30 minutes.

① Draw a graph of this distribution.

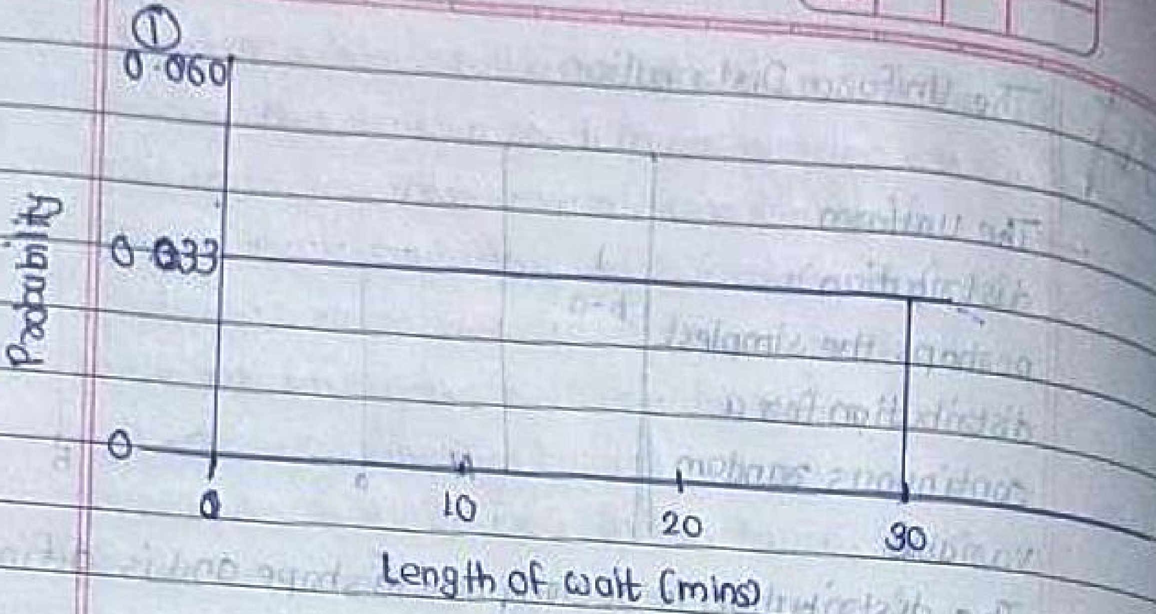
② Show that the area of this uniform distribution is 1.00.

③ What is the mean of waiting time? What is S.D of waiting time?

④ What is the probability that student will wait more than 25 minutes?

⑤ What is the probability that student will wait between 10 and 20 minutes?

①



②

$a = 0 \quad b = 30$

$$\text{Area} = (\text{height})(\text{base}) = \frac{1}{(30-0)}(30-0) = \underline{\underline{1.00}}$$

③

$$\mu = \frac{a+b}{2} = \frac{0+30}{2} = \underline{\underline{15}}$$

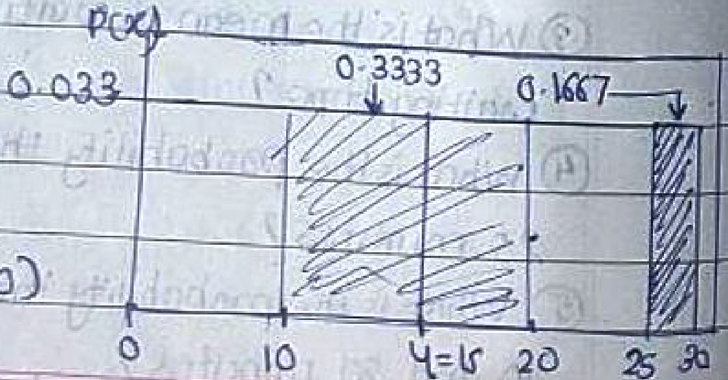
$$\sigma = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{(30-0)^2}{12}} = \underline{\underline{8.66}}$$

④

$P(25 < \text{Wait Time} < 30) = (\text{height})(\text{base})$

$= \frac{1}{(30-5)}(5)$

$= 0.1667$



⑤

$P(10 < \text{wait Time} < 20)$

$= (\text{height})(\text{base})$

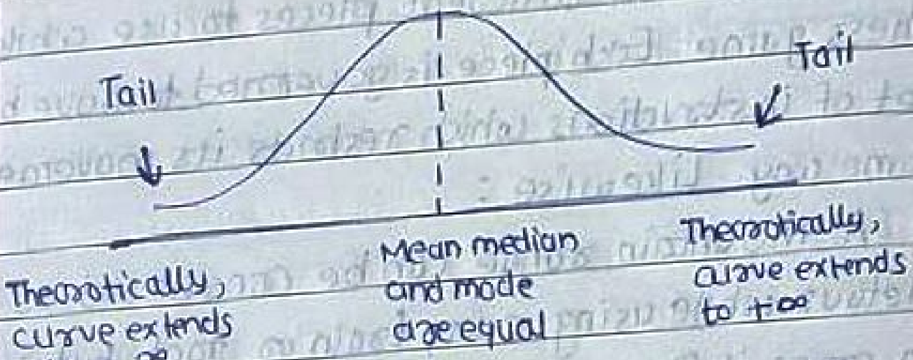
$= \frac{1}{(30-0)}(10) = 0.3333$

Characteristics of a Normal Probability Distribution:

Characteristics of a Normal Probability Distribution:
① It is bell-shaped and has a single peak.
② It is symmetrical about the mean.
③ It is asymptotic: The curve gets closer and closer to the x-axis but never actually touches it.
④ The arithmetic mean, median, mode are equal.
⑤ The total area under the curve is 1.00.
⑥ The area to the left of the mean

= The area to the right of the mean

= 0.5



The Standard Normal Probability Distribution:
- Any normal distribution can be standardized.
- The standard normal distribution is a normal distribution with a mean of 0 and S.D of 1.
- Also called z-distribution.

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

$X \rightarrow$ signed distance between selected value

$\mu \rightarrow$ Population mean

$\sigma \rightarrow$ Population standard deviation