Page No.
Date

Estimates of location
Mean
Weighted mean
Median
Weighted median
Trimmed mean
Robust
Outliers

Key Terms for Data Types:
Continuous:
Data that can take on any value in an interval

Discrete:
Data that can take on only integer values, such as counts.

Categorical
Data that can take on only a specific set of values representing a set of possible categories.

Binary
A special case of categorical data with just two categories of values (0/1, true/false)

ordinal
Categorical data that has an explicit ordering

Scanned with OKEN Scanner

Columns - features
rows - record

Page No.
Date

Data → Sample → Sampling distribution → hypothesis
→ Conclusion.

Rectangular Data
xcel dataset

2 chptrs → IA - I

~~Exploratority~~ Exploratory Data Analysis

Q.1 What are the elements of structured data

1. Continuous:
The data that can take any value in an interval like your weight.

2. Discrete
The data that can take only integers
Discrete data that cannot be converted into points

3. Categorical
Data that can take ~~a~~ adya specific set of values representing a set of possible categories.
College having different streams.
(I T, AIDS, comps)

4. Binary
A categorical data with just two categories of values
Ex: Switch ON/OFF, PASS/FAIL

5. ordinal.
A categorical data that has an explicit ordering like grades, rank. A group of students having roll nos. from 1 to 60.

Q3. What are the types of structured data?
~~The Nume~~ There are 2 types:
Numerical, Categorical

Types of Structured Data

- Numerical
  - Discrete: Count of occurrence of events
  - Continuous: Time duration
- Categorical: takes a fix set of values

1. For
* Purpose: The purpose of data analysis and predictive modelling; the data type is important to help determine the type of visual display, data analysis or statistical model. 3. The data type for a variable determines how software such as python/R will handle computations for that variable. 4. Knowing that data is Categorical can act as a signal telling software how statistical procedures such as producing a chart or fitting a model should behave. 5. Storage and indexing can be optimized

Q. 3 What is Rectangular data?

Ans 3) The typical frame of reference for an analysis in data science is a rectangular data object like a spreadsheet or a database table.

1. Dataframe:
Rectangular data is the basic data structure for statistical and machine learning models

2. Feature:
The column in the table is called as a feature. It can also be called as Attribute, input, predictor, and variable.

3. outcome
Ex: Mr. Narendra Modi was elected as the next Prime minister of India.
outcome can also be called as dependent variable, Response, Target or output.

4. Records
A row in a table is called as records.

Summarizing point → Rectangular data is a 2dimensional matrix with rows indicating cases and columns indicating variables

	Processor	RAM/Rom	Camera	Display	Battery	Price
Iphone 14 pro max	1.25	1.5	1.25	1.5	3.5	5
Samsung S23 Ultra	1.50	1.25	1	1	1.5	5
Pixel 7 pro	2	3.5	1.75	2.5	3	3
one plus 11	2.5	3	3	5	2.5	2.75
Asus ROG46	1	1.25	2.5	2.5	2	3
Overall	Asus	Samsung / Asus	Samsung	Samsung	Samsung	oneplus 11

Scanned with OKEN Scanner

Page No
Date

Q.4 Non Rectangular Data Structures:

1. Time Series Data
2. Spatial Data Structures
   Which are used for mapping and location analysis. the object representation. The primary focus is an object Ex. house, office, college and its spatial co-ordinates, the field view focuses on small unit of space and value of a relevant matric
3. Graph Data Structures:
   Ex: Distribution Hubs connected by roads are an example of physical network. Graphs are used for network optimization.

Q.5 What are the estimates of location:

1. mean.
2. weighted mean.
3. median
4. weighted median
5. Trimmed median
6. Robust
7. outliers.

overall

Estimates of location

weight	date	wixi
1	17	17
1	16	16
2	12	24
2	14	28
1	16	16
1	18	18
3	10	30
2	11	22
2	12	24
2	13	26
2	15	30
1	16	16
2	11	22
4	2	8
1	20	20
$\Sigma wi = 27$		$\Sigma wixi = 317$

Estimates of location:

mean = 13.53

median = 14

Trimmed mean = 14.36

outlier = 2

median: 2 10 11 11 12 12 13 14 15 16 16 16 17 18 20

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

Trimmed mean: Removing outliers

$$\frac{10+11+11+12+12+13+14+15+16+16+16+17+18+20}{14}$$

$$14.36$$

Weighted mean: $\frac{\Sigma wixi}{\Sigma wi} = \frac{317}{27} = 11.74$

Scanned with OKEN Scanner

Page No
Date

Estimates of variability:

1. Deviation
The difference between the observed value and the estimate of location.

2. Variance
The sum of squared deviations from mean divided by $n-1$ where n is the no. of data values
Also called as Mean Squared Error.

3. Standard Deviation
Square root of variance

4. Mean Absolute Deviation
The mean of the absolute values of the deviation from the mean

5. Median Absolute Deviation
The median of the absolute values of the deviation from the median.

6. Range:
The difference between the largest and the smallest value of a dataset.

7. Order Statistics
Matrix based on the data values sorted from smallest to largest

8. Percentile~~tile~~
The value such that P percent of the values take on this value or less and 100-P Percent take on this value or more.

9. Inter Quartile Range
The difference between the 75th Percentile and the 25th Percentile.

Page No.
Date

Calculate MAD, variance, standard deviation

Date :	$x - \bar{x}$	Taking mod	Taking Square
17	$17 - 15.5 = 1.5$	1.5	2.25
16	$16 - 15.5 = 0.5$	0.5	0.25
12	$12 - 15.5 = -3.5$	3.5	12.25
14	$14 - 15.5 = -1.5$	1.5	2.25
16	$16 - 15.5 = 0.5$	0.5	0.25
18	$18 - 15.5 = 2.5$	2.5	6.25
$\bar{x}$	15.5	10	23.5

$\frac{10}{6} = 1.66$ ← Mean Absolute Deviation

Standard Deviation = $\sqrt{\text{variance}}$

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

Variance = $\frac{23.5}{6-1} = \frac{23.5}{5} = 4.7.$

Standard Deviation = $\sqrt{4.7} = 2.16$

$n-1$ = Degree of Freedom

In Statistics, there is always some discussion of why we have $n-1$ in the denominator, and in the variance instead of n we use $n-1$. $n-1$ is called as Degrees of Freedom

Page No
Date

Wether to consider n or n-1 will not have a major impact by analyzing large datasets however if you want to estimate about a population based on the sample set then n or n-1 will have a greater impact. If you use the denominator as 'n' in the variance formula you will underestimate the true value of the variance and the standard deviation in the population. this is referred to as biased estimate. However if you divide n-1 instead of n, the standard deviation becomes an unbiased estimate.

Q. Fix question in IA-I
Show that why standard deviation is always greater than mean absolute deviation and why mean absolute deviation is always greater than median absolute deviation

SD > mean Absolute Deviation > median Absolute Deviation

Scanned with OKEN Scanner

Page No.
Date

Q. The runs scored cricket match by 11 players are follows : 7, 16, 121, 51, 101, 81, 1, 16, 9, 11, 6

1. Standard deviation
2. Mean absolute deviation
3. median Absolute deviation

no. of observations = 11.

Score	$\lvert x-\bar{x} \rvert$	$(x-\bar{x})^2$
7	31.18	972.1924
16	22.18	491.9524
121	82.82	6859.1524
51	12.82	164.3524
101	62.82	3946.3524
81	42.82	1833.5524
1	37.18	1382.3524
16	22.18	491.9524
9	29.18	851.4724
11	27.18	738.7524
6	32.18	1035.5524
Summation		
420	402.54	18767.61

$$\bar{x} = \frac{420}{11} = 38.18$$

① Standard deviation = $\sqrt{\text{variance}}$

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

Scanned with OKEN Scanner

11 = no. of observations

$$= \sqrt{\frac{18767.61}{11-1}}$$

$$= \sqrt{\frac{18767.61}{10}}$$

$$= 43.32$$

(2) Mean absolute deviation $= \frac{\Sigma|(x-\bar{x})|}{n} = \frac{402.54}{11}$

$$= 36.59$$

Arranging all the scores in ascending order, we get

1, 6, 7, 9, 11, **16**, 16, 51, 81, 101, 121

median

x	$\lvert(x - \text{median})\rvert$
1	15
6	10
7	9
9	7
11	5
16	0
16	0
51	35
81	65
101	85
121	105

Scanned with OKEN Scanner

Page No.
Date

Now arranging new scores in ascending order,

0, 0, 5, 7, 9, 10, 15, 35, 65, 85, 105

Median.

Median Absolute deviation = 10

By observing all three values, we can conclude,

SD > Mean AD > Median AD

Estimates based on Percentiles.

Different approach to estimating dispersion is based on looking at the spread of the sorted statistics based on sorted rank data are referred to as ordered statistics. The most basic measure is the range. (The difference between the largest value and the smallest value)

The maximum and the minimum values themselves are useful to know and helpful in identifying outliers.

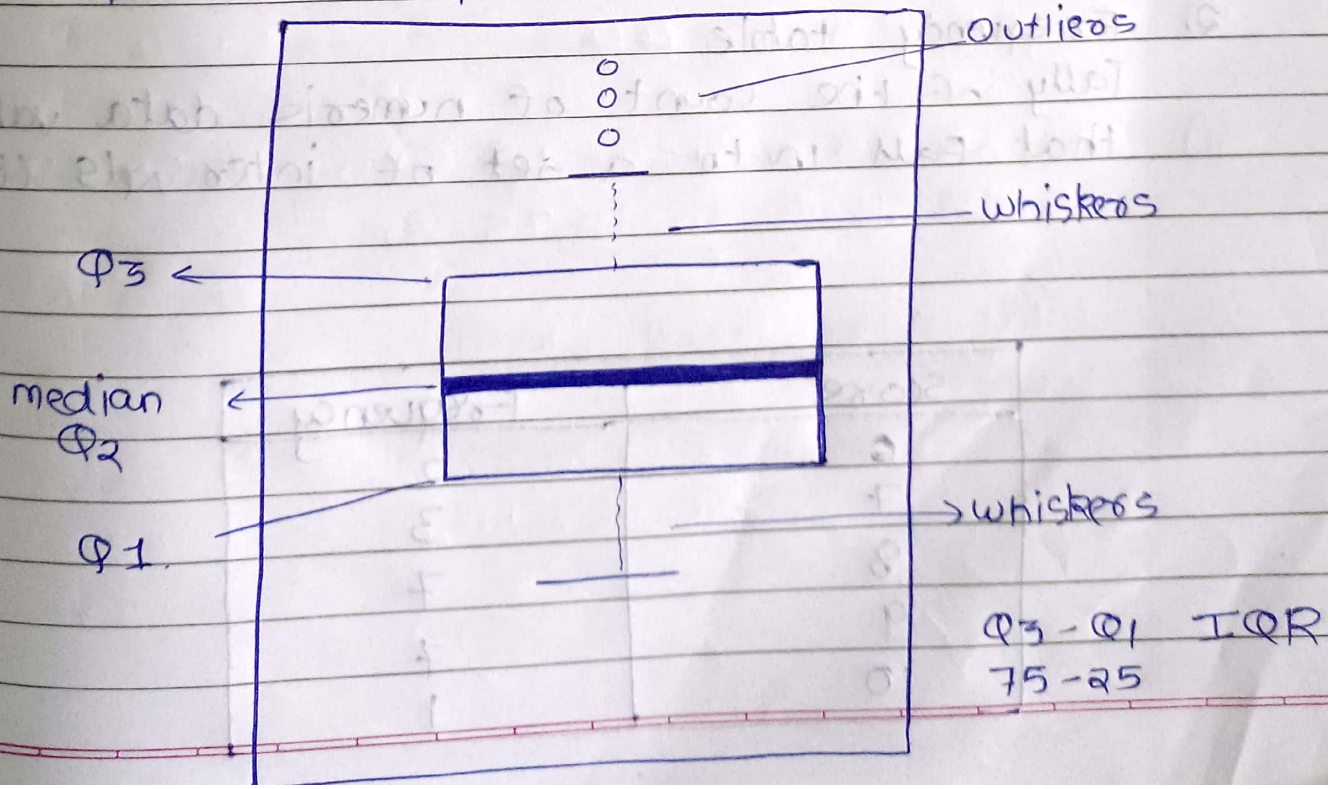
Therefore, range is most sensitive to outliers. To avoid the sensitivity to the outliers we can look at the range of the data after dropping values from each end. This types of estimates are based on differences between percentiles.

Scanned with OKEN Scanner

25^{th} percentile is called as quartile 1
50^{th} percentile is called as quartile 2 / median
75^{th} percentile is called as quartile 3
The difference between 75^{th} percentile & 25^{th} percentile is called as inter quartile range.

Q. The scores in a test are 40, 45, 49, 53, 61, 68, 71, 79, 85, 91. What will be the percentile for these scores.

1. 40
2. 45
3. 49
4. 53
5. 61
6. 68
7. 71 → 70 percentile
8. 79 → 80 percentile
9. 85 → 90 percentile
10. 91 → 100 percentile



Scanned with OKEN Scanner

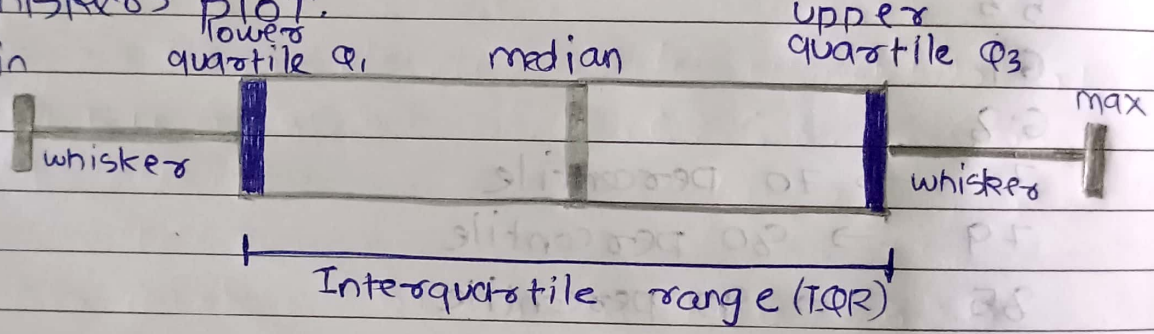
Page No.
Date

Data distribution

Each of the estimates that we have covered sums up the data in a ~~singular~~ single no. to describe the location or variability of data

It is useful to explore how the data is distributed overall.

1. Box plot:
A plot introduced by John Tukey as a quick way to visualize the distribution of data based on percentile. It is also called as box and whiskers plot.

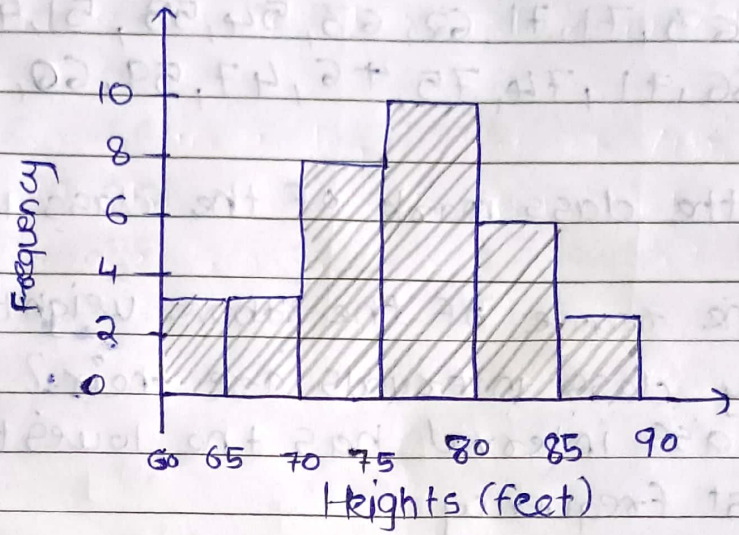


2. Frequency table.
Tally of the count of numeric data values that fall into a set of intervals (bins)

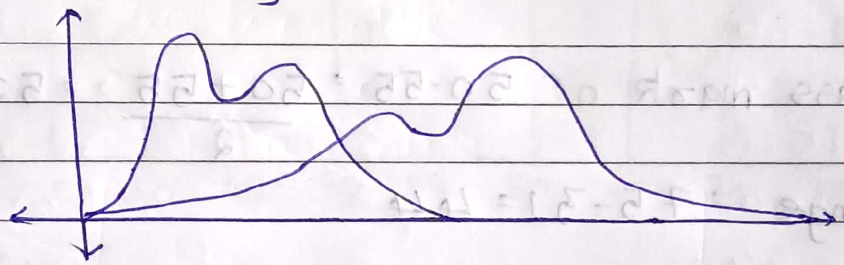
Score	Frequency
6	2
7	3
8	7
9	7
10	1

Scanned with OKEN Scanner

3. Histogram.
A plot of the frequency table with the bins on the x-axis and the count on the y-axis.



4. Density plot.
A smoothed version of the histogram, based on known density estimate.



Scanned with OKEN Scanner

Page No.
Date

Q. Construct a frequency distribution table for the following weights of 30 oranges using the equal class intervals. One of the is 40-45 (45 not included). The weights are 31, 41, 46, 33, 44, 51, 56, 63, 71, 71, 62, 63, 54, 53, 51, 43, 36, 38, 54, 56, 66, 71, 74, 75, 46, 47, 59, 60, 61, 63

What is the class mark of the class interval 50-55?
What is the range of the above weights?
How many class intervals are their?
which class interval has the lowest and the highest frequency

Soln: 31, 33, 36, 38, 41, 43, 44, 46, 46, 47, 51, 51, 53, 54, 54, 56, 56, 59, 60, 61, 62, 63, 63, 63, 66, 71, 71, 71, 74, 75

i) Class mark of $50-55 = \frac{50+55}{2} = 52.5$

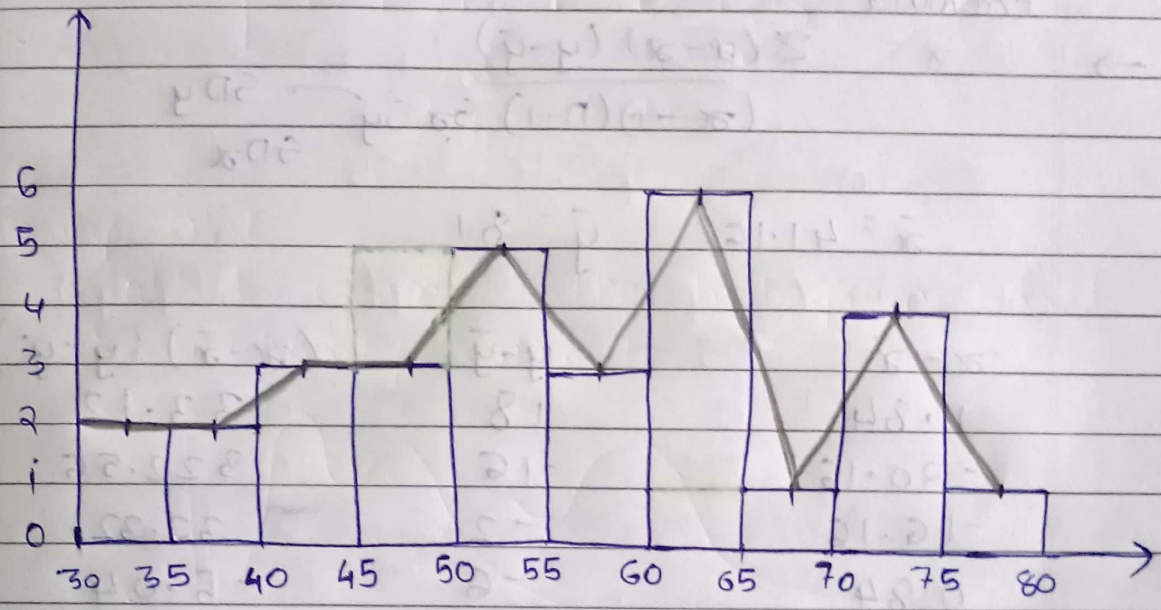
ii) Range $= 75-31 = 44$

iii) 30-35, 35-40, 40-45, 45-50, 50-55, 55-60, 60-65, 65-70, 70-75, 75-80

Total class interval $= 10$

Scanned with OKEN Scanner

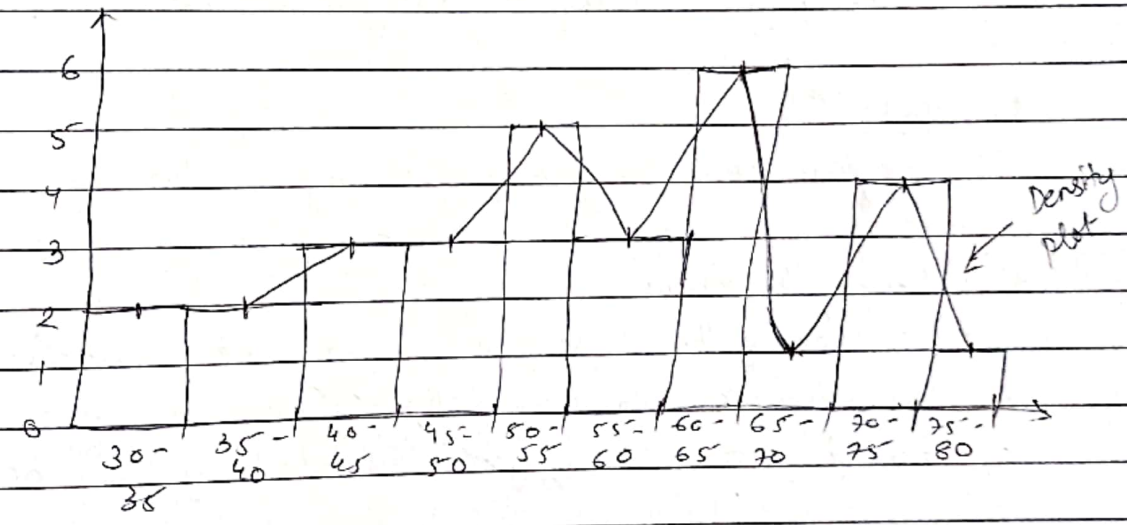
	Frequency
30-35	2
35-40	2
40-45	3
45-50	3
50-55	5
55-60	3
60-65	6 → highest.
65-70	1 → lowest
70-75	4
75-80	1 → lowest



Page No.:
Date: / /

(iv)

	Frequency
30 - 35	2
35 - 40	2
40 - 45	3
45 - 50	3
50 - 55	5
55 - 60	3
60 - 65	(6) highest.
65 - 70	1 ← lowest
70 - 75	4
75 - 80	1

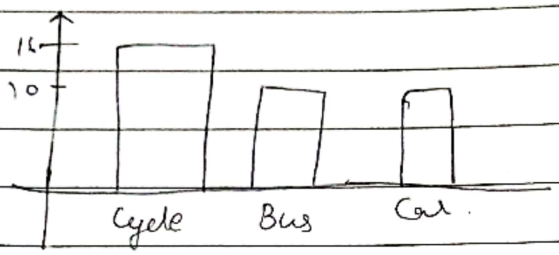


Q. A survey of 36 students of a class was done to find out the mode of transport used by them while computing to the college. The collected data is shown in the table given below. Represent the data in terms of bar graph.

Scanned with OKEN Scanner

Mode of transport	No. of student
Cycle	16
Bus	10
Car	18

→



Exploring Binary & Categorical Data

1. Mode:-
The most commonly occuring value or category.

2. Expected value :-
When the categories can be associated with a numeric value, this gives an average value based on categories probability of occurence.

3. Bar charts
The frequency for each category plotted as bars.

4. Pie charts
The frequency or proportion for each category plotted on wedges as pie.

Scanned with OKEN Scanner

Page No.:
Date: / /

Q. The company offers 3 subscriptions plans. The basic version is priced at Rs 2500. Standard Version is priced at Rs 7000. And the pro version is priced at Rs 15,000. The probability of students signing for the course are 20% for basic, 50% for standard and the remaining for pro version. Calculate the expected value

→ Expected value = 2500 × 0.2 + 7000 × 0.5 + 15000 × 0.3
= 8500

Correlation

Exploratory data analysis. (EDA) in many modeling projects (whether in data science in research) involves examining correlation among predictor vs blw prediction and a target variable. Variables X & Y are set to be +vely correlated if high values of x go with high values y, and low values of x go with low values of y.

If high values of x go with low values of y, the variables are -vely correlated

Correlation Coefficient

A metric that measures the extent to which numeric variables are associated with one another ranges from -1 to 1.

Scanned with OKEN Scanner

Page No.:
Date: / /

Correlation Matrix

A table where the variables are shown on both rows & columns, & the cell values are the correlation b/w the variables.

Scatter plot:

A plot in which the x-axis is the value of one variable and the y-axis is the value of another ~~value~~ variable.

19/8/23

(*) **Correlation Coefficient**

Mod.1 Sum16 Q. Find the value of the correlation coefficient from the given data

Subject	Age (x)	Glucose (y)	$(x-\bar{x})$	$y-\bar{y}$
1	43	99	+1.84	18
2	21	65	-20.16	-16
3	25	79	-16.16	-2
4	42	75	0.84	-6
5	57	87	15.84	+6
6	59	81	17.84	0
	41.16	81		

Sol^n:-

$$r = \frac{\Sigma (x-\bar{x})(y-\bar{y})}{(n-1) S_x \cdot S_y}$$

=

Page No.:____
Date:__/__/__

$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
33.12	1849	9801
322.56	441	4225
32.32	625	6241
-5.04	1764	
95.04	3249	
0	3481	
478		

$$r = \frac{478}{5.1}$$

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

$$= \sqrt{\frac{1240.83}{5}} = \sqrt{248.166} = 15.75$$

$(x-\bar{x})^2$	$(y-\bar{y})^2$
3.3855	324
406.4256	256
261.1456	4
0.6561	36
250.905	36
318.2656	0
1240.83	656

$$S_y = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n-1}} = \sqrt{\frac{656}{5}} = 11.45$$

Scanned with OKEN Scanner

$$r = \frac{478}{5 \times 15.75 \times 11.45}$$

$$= +0.5299 \qquad (-1.00 \text{ to } 1.00)$$

(C.S.M)

Q Calculate the correlation coefficient for the following data.

x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
4	5	-6	-5	36	25
8	10	-2	0	4	0
12	15	2	5	4	25
16	20	6	10	36	100
40	50			80	150
10	7.5				

$(x-\bar{x})(y-\bar{y})$
+30
0
-5
10
15
-45
-20

$$r = \frac{46}{3 \times 5.163 \times 5} = 0.5939$$

$$S_x = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} = \sqrt{\frac{80}{3}} = 5.163$$

$$S_y = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n-1}} = \sqrt{\frac{50}{2}} = \sqrt{25} = 5.04$$

Scanned with OKEN Scanner

$(x-\bar{x})^2$	$(y-\bar{y})^2$
36	25
4	0
4	25
36	.
80	~~50~~ 51

$SD_x = \sqrt{\text{variance}}$

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

$= \sqrt{\frac{80}{3}}$

$= 5.16$

$SD_y = \sqrt{\text{variance}}$

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

$= \sqrt{\frac{51}{3}}$

$= 4.1231$

$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
-6	-2.75	16.5	36	7.5625
-2	2.25	-4.5	4	5.0625
2	7.25	14.5	4	52.5625
6	-6.75	-40.5	36	45.5625
		-14	80	110.75

$\sqrt{\frac{110.75}{3}}$

$SD_y = 6.0759$

$SD_x = 5.16$

$r = \frac{46}{3 \times 5.16 \times 6.0759}$

$= \frac{46}{94.054932}$

$= 0.489$

x	y
4	5
8	10
12	15
16	20
40	50

$\bar{x} = \frac{40}{4} = 10$ $\bar{y} = \frac{50}{4} = 12.5$

$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$
-6	-7.5	45
-2	-2.5	5
2	2.5	5
6	7.5	45
		100

$r = \frac{100}{3 \times 5.16 \times 6.4549} = \frac{100}{99.921852} = 1.0007$

$(x-\bar{x})^2$	$(y-\bar{y})^2$
36	56.25
4	6.25
4	6.25
36	56.25
80	125

$SD_x = \sqrt{\frac{80}{3}}$ $SD_y = \sqrt{\frac{125}{3}}$

$SD_x = 5.16$ $SD_y = 6.4549$

Page No.
Date

Exploring 2 or more variables:

- Estimators like mean and variance look at variables one at a time (univarient analysis)
- Correlation analysis compares 2 variables (bivarient analysis)
- Estimators which includes more than 2 variables are called as multivarient analysis.

1. multivarient analysis

a. Contingency Table:
A tally of counts between two or more categorical variables

b. Hexagonal Binning
A plot of 2 numerical variables with the records binned into hexagons

c. Contour plots.
A plot showing the density of 2 numeric variables like a topographical map.

d. Violin plots
Similar to the box plot, but showing the density estimate.

Scanned with OKEN Scanner

Page No.
Date

IA QB

Chpt 1:

1. Difference between structured & unstructured da
2. What is the estimate of location
3. What is relation between median and robust estim
4. What are the estimates of variability / dispersio
Imp 5. Show that Standard Deviation > Mean Absolute Deviat
> Median Absolute Deviation.
6. What are the different tools used for data distribution.
7. What is expected value (sum / theory)
8. What is correlation
Why it is it called as bivarient analysis.
9. Numericals based on variance, SD, mean, mean absolute deviation, median absolute, ~~mad~~ deviation
10. Numericals based on correlation coefficient.

Chpt 2:

1. What are the types of samplings.
or. methods of samplings.
2. What is sample bias.
vImp 3. Data Quality v/s Data Quantity. [Theory Q] [5 m]
4. What is sampling procedure
5. What is selection biased
Answers includes 2 concept a. data snooping
b. vast search Effect
6. Why outliers / extreme observations tend to follow by mean / Regression to the mean
2m V. Imp 7. What is Central Limit Theorem
2m. 8. What is standard Error
V. Imp Q. Came in Sem Resampling v/s Bootstraping
10. What is Normal Distribution (Gaussian) Also called as: Pg. 70 Graph

Scanned with OKEN Scanner

Page No.
Date

11. demonstrate QQ plots for normal distribution and Long tailed Distribution
12. What is students t-distribution
13. What is binomial distribution

$$\frac{\bar{x} - \bar{x}}{\frac{s}{\sqrt{n}}}$$

Study for Their in IA:

14. What is Chi-square test
15. What is f distribution
16. What is poisson distribution
17. What is weibull Distribution

Scanned with OKEN Scanner

Page No.
Date

Chapter: 2.

Random Sampling and Sample Bias

Sample:

Definition: A subset from a larger dataset / population

$\bar{X}$ Population mean

$\bar{x}$ Sample mean

N no. of observations of Population mean

n no. of observations of sample mean

Random Sampling.

It is a process in which each available member of the population being sampled has an equal chance of being choosen for the sample at each draw.

or

The equal chance of being selected in the sample in each draw this is also called as simple Random Sampling

→ Sampling can be done with replacement or without replacement

Knowledge purpose:

Self selection Sampling bias:

(Sample)

Statistical bias: It refers to measurement or sampling errors that are systematic and produced by the measurement or sampling process.

→ An important distinction should be made b/w errors due to random chance and errors due to bias.

Scanned with OKEN Scanner

$\bar{a}$ or $\bar{y}$ $\bar{x}$

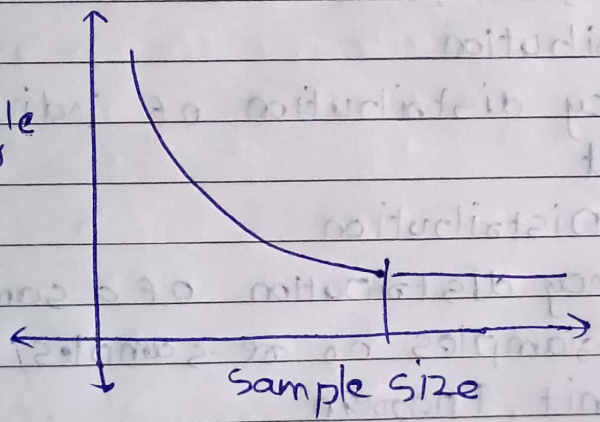
Page No
Date

Sample mean V/S Population mean

Selection bias refers to the practice of selectively choosing data conciously or unconsciously in a way that leads to a conclusion that is ~~misread~~ misleading

Data snooping
Extensively hunting to data in search of something interesting.

Vast Search Effect:
Bias or non-reproductibility resulting from ~~typi~~ repeated data modelling or modeling data with large number of predictor variables.



Regression to the mean:
It refers to a phenomenon involving successive measurements of a given variable; extreme observations tend to be followed by the more controlled ones

Central Limit Theorem

As we ↑ the sample size the data gets closer to the mean set

Page No.
Date

2 Attaching focus and meaning to the extreme value can lead to a form of selection biased

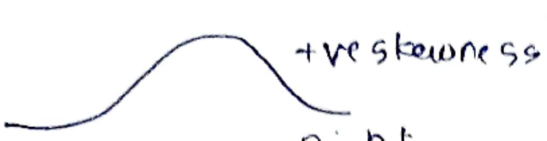
3 Regression to the mean, meaning to "go back", is distinct from the statistical modelling method of linear regression, in which a linear relationship is estimated between dependent and independent variables

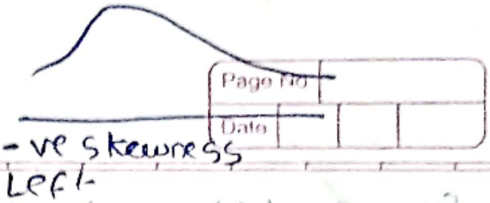
Sampling Distribution of a Statistic

It refers to a distribution of some sample statistic over many samples drawn from the same population

1. Sample statistic
A matric calculated for a sample of data drawn from a larger population
2. Data Distribution
The frequency distribution of individual values in a dataset
3. Sampling Distribution
The frequency distribution of a sample statistic over many samples or re-samples
4. Central limit Theorem
The tendency of the sampling distribution to take on a normal shape as sample size rises
5. Standard Error
The variability of sample statistic over many samples

~~Exp~~ The variability between individual values is called standard deviation.





Page No.
Date

(5m) Central Limit Theorem (take)

The means drawn from multiple samples will resemble the familier bell shaped normal curve, even if the source population is not normally distributed, provided that the sample size is large enough. The Central Limit Theorem receives a lot of attention in traditional statistics text because it underlies the machinery of hypothesis test and confidence intervals, which themselves consumes half the space in such text.

Data scientist should be aware of this rule.

Standard Error

The Standard error is a single matric that sums up the variability in the sampling distribution for a statistic.

$$SE = \frac{s}{\sqrt{n}}$$

s — Standard deviation of sample values

n — Sample size

As the sample size increases, standard error decreases.

Collecting new samples to estimate the standard ~~Samp~~ error is not ~~visi~~ feasible. Instead of drawing new samples you can use bootstrap resamples

Page No.
Date

The Bootstrap:

The effective way to estimate the sample distribution of a statistic is to draw additional samples with replacement, from the sample itself and recalculate the statistic ~~or model~~ for each resample. This procedure is called the Bootstrap

Standard Error sums.

The certain property investment company with an international presence, workers have a mean hourly wage of 12 dollars with a population standard deviation of 3 dollars. Given a sample size of 30, estimate and interprete the standard error of the sample mean

$$SE = \frac{\sigma}{\sqrt{n}} \quad \text{↑ SD of population}$$

$$= \frac{3}{\sqrt{30}}$$

$$0.5477$$

Explanation: If we ~~was~~ were to draw several samples of 30. from the workers population and construct a sampling distribution of the sample means, we would end of with the mean of 12 dollars and a standard error of $0.55

Data quality vs Data quantity.
Example

Page No
Date

Assume that we have increased the sample size to 80 what will be the standard error.

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{80}} = 0.3354 \downarrow$$

Assume that we have decrease the sample size to 15

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{15}} = 0.77 \uparrow$$

find the SE of estimate of mean ~~rate~~ weight of high school football players using the data given of weights of high school players

Players	weights (in pounds)	$x - \bar{x}$	$(x-\bar{x})^2$
player 1	150	-31.6	998.56
player 2	203	21.4	457.96
player 3	176	-5.6	31.36
player 4	190	8.4	70.56
player 5	168	-13.6	184.96
player 6	193	11.4	129.96
player 7	189	7.4	54.76
player 8	178	-3.6	12.96
player 9	197	15.4	237.16
player 10	172	-9.6	92.16

$\bar{x} = 181.6$

$\Sigma(x-\bar{x})^2 = 2270.4$

$$\sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{N}} = 15.06$$

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

$\sigma = 15.88$

Scanned with OKEN Scanner

<30 we will consider sample.

Page No.
Date

$\frac{\sigma_x}{\sqrt{n}} = \frac{15.88}{\sqrt{10}} = 5.02 \rightarrow S.E$

σ_x — sample Standard Deviation

σ — Population Standard Deviation

Q. Find the Standard Error of the estimate for the average no. of children in a household in your city by using the data collected from a sample of household in your city.

Household	no. of children	$(x-\bar{x})$	$(x-\bar{x})^2$
1	2	-0.25	0.0625
2	3	0.75	0.5625
3	1	-1.25	1.5625
4	0	2.25	5.0625
5	5	2.75	7.5625
6	2	-0.25	0.0625
7	1	-1.25	1.5625
8	4	1.75	3.0625
	2.25		$\Sigma(x-\bar{x})^2 = 19.5$

$SE = \frac{\sigma_x}{\sqrt{n}} = \frac{19.5}{\sqrt{8}}$

$\sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} \quad \sqrt{\frac{19.5}{7}}$

$\sigma_x = 1.66 \quad 1.67$

Scanned with OKEN Scanner

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{1.67}{\sqrt{8}} = 0.59$$

$$\sigma = \sqrt{0.595} \cdot 1 = 0.79$$

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{0.70}{\sqrt{2}}$$

Bootstrap

i) Bootstrap sample :
A sample taken with replacement from an observed data set

ii) Resampling
The process of taking repeated samples from observed data

Household	Sample 1	Sample 2	3	4	5
1	1	2	3	1	2
2	2	2	4	2	2
3	2	2	4	4	2
4	2	3	1	3	2
5	4	4	2	2	2

The algorithm for bootstrap resampling of a mean for a sample size of n is as follows:

Step1 Draw a sample value, recorded and then replace it
Step2 Repeat n times
Step3. Record the mean of the n resampled values
Step4 Repeat steps 1,2,3 ~~R~~ R times
(R = iteration)
Step5 : Use the R results to 1) Calculate the standard deviation

b. Produce the histogram or boxplot
c Find the confidence interval.

Resampling vs Bootstraping

1. The Bootstrap is a tool for assessing the variability of sample statistic
2. The Bootstrap can be applied in a similar fashion in wide variety of circumstances
3. Bootstrap allows us to estimate sampling distribution for statistics where no mathematical approximation has been developed.
4. When applied to ~~predictive~~ predictive models aggregating multiple bootstrap samples predictions outperform the use of single model.

Confidence Interval.

1 Confidence level
The percentage of confidence intervals, constructed in a same way from the same population that are expected to contain the statistic of interest.

2 Interval end points.
The top and bottom of the confidence interval

Note: The mean, median, variance, standard deviation for a sample is gives us or called as point estimate. Presenting an estimate not as a single number but as a range is called as confidence interval.

Scanned with OKEN Scanner

$\bar{x}$ - sample mean
μ - population mean

Page No
Date

confidence intervals always comes with a coverage levels expressed as percentages ex 90%, 95%, 99%

X is a normally distributed variable with mean of 30 ($\mu = 30$) and standard deviation as 4
Find poopability of a dataset which is
~~greate~~ $P(x < 40)$
$P(x > 21)$
$P(30 < x < 35)$

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

① $z = \frac{40 - 30}{4}$

$z = \frac{10}{4}$

$z = 2.5$

0.9938

99.38%

② $z = \frac{21 - 30}{4}$

$\frac{-9}{4}$

-2.25

~~0.12224~~ 0.01222

$x >$ ..

we will minus the answer from 1 -0.01222

0.98778

or

98.78%

Scanned with OKEN Scanner

Page No.
Date

③ $P(30 < x < 35)$

$$z = \frac{30 - 30}{4}$$

$$z = 0$$

$$z = \frac{35 - 30}{4}$$

$$z = 1.25$$

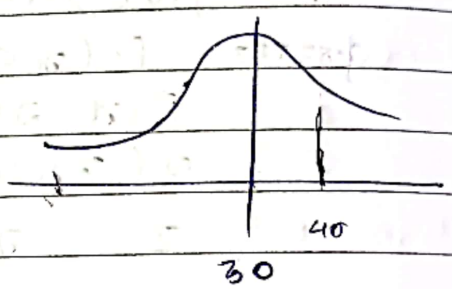
$$0 < p < 1.25$$

$$0.8943$$

$$0.8943 - 0.5000$$

$$0.3943$$

or $39.43\ \%$.



Practice Problem:

13

★ A radar unit is used to measure speed of cars on a highway. The speed are normally distributed in a mean of 90km/hr and standard deviation of 10 km/hr. what is the probability that a car picked randomly is travelling at more than 100 km/hr

$x > 100$.

$$z = \frac{x - \mu}{\sigma} = \frac{100 - 90}{10} = \frac{10}{10} = 1$$

$$\approx 0.8413$$

$$1 - 0.8413$$

$$0.1587$$

or

$$15.87$$

Scanned with OKEN Scanner

z table - Normal Distribution Table

Page No
Date

Players Question:

Confidence range:

95% $1.64 \times SE = 1.64 \times 5.022$

$= 8.23$

$173.37 < p < 189.83$

181.

Confidence interval 173.37 to 189.83

15 to 30

Children Question:

99% 2.33 → From Table.

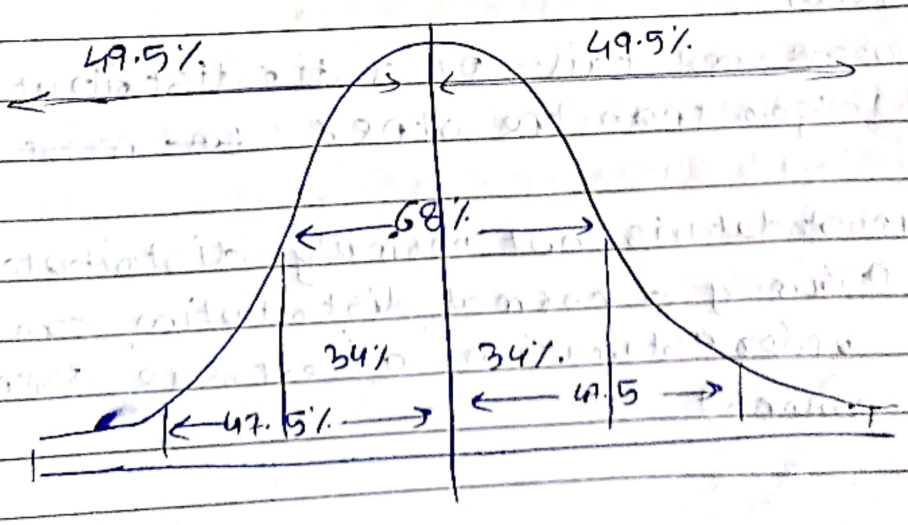
$2.33 \times SE = \cancel{2.33 \times 2.25}$ 2.33×0.59

$= \cancel{5.2425}$ 1.37.

$\cancel{c\ p < 7.4925}$

$0.88 < p < 3.62$ (to)

The bell shaped normal distribution states that 68% of the data lies within one standard deviation of the mean and 95% lies within two standard deviation and 99% data lies within 3 standard deviation



Scanned with OKEN Scanner

Page No.
Date

Normal Distribution:

1. Errors
The difference between a data point and predicted value
2. Standardize
Subtract the mean and divide by the standard deviation
3. Z-score
The result of standardizing an individual data point
4. Standard Normal
A normal distribution with mean 0 and standard deviation 1
5. QQ Plots
A plot to visualize how close a sample distribution is to a specified distribution

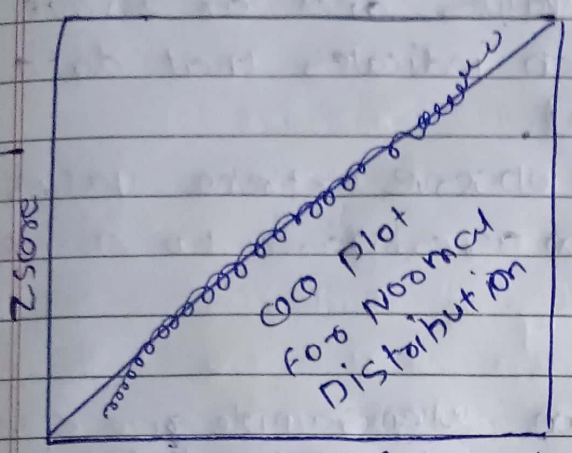
Long Tail Distribution:

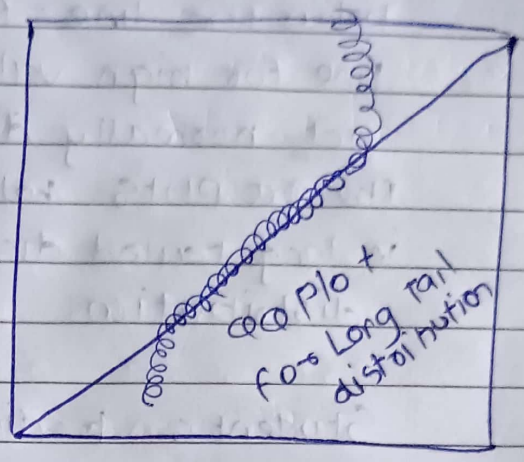
1. Tail
The long narrow portion of a frequency distribution where relatively extreme values occur at low frequency
2. Skew
where one tail of a distribution is longer than the other ~~the most data is~~

1) most data is not normally distributed
2) Assuming a normal distribution can lead to underestimation of extreme events (black swans)

Page No.
Date

Students t distribution-





Student

The t distribution is a normally shaped ~~distributed~~ distribution ~~[illegible]~~ except that it is thicker and longer on the tails.

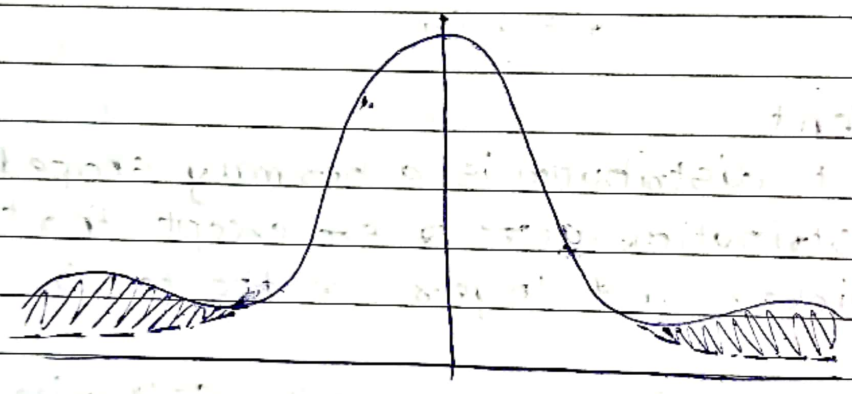
A QQ plot is used to visually determine how close a sample is to a specified distribution.

In normal distribution, QQ plot orders the z scores from low to high and plots each value's z score on the y-axis, the x-axis is the corresponding Quantile of a normal distribution for that value's rank. Since the data is normalized, the units correspond to the no. of standard deviations away from the mean.

If the points roughly fall on the diagonal line then the sample distribution can be considered closed to normal

In long tailed distribution, the points are far below the line, for low values and far above the line for high values, which indicates that data is not normally distributed.
The QQ Plots helps to observe extreme values in long tailed distribution as compared to normal distribution

Student's t distribution. when sample size ≤ 30 we use t distribution



- The t distribution is normally shaped distribution but thicker and larger on the tails
- It is used extensively in depicting distributions of sample statistics
- The larger the sample, the more normally shaped t distribution becomes

The t distribution is widely used as a reference basis for the distribution of sample means, difference b/w two sample means, regression parameters

Variance $\frac{\Sigma(x-\bar{x})^2}{N}$ ← population $\quad \frac{\Sigma(x-\bar{x})^2}{n-1}$ ← sample $\quad z = \frac{x-\mu}{\sigma}$

Page No. Date

Practice Problems

4 If die is rolled. Find variance and standard deviation

	$(x-\bar{x})$	$(x-\bar{x})^2$
1	-2.5	6.25
2	-1.5	2.25
3	-0.5	0.25
4	0.5	0.25
5	1.5	2.25
6	2.5	6.25
21		17.5

$\bar{x} = \frac{21}{6} = 3.5$

$\text{Variance} = \frac{\Sigma(x-\bar{x})^2}{N} = \frac{17.5}{6}$

Variance = 2.917

$SD = \sqrt{\text{variance}}$

$= \sqrt{2.917}$

SD = 1.707

IA-I

Sums → binomial, t, chi square distribution, normal distribution

Scanned with OKEN Scanner

Page No.
Date

sample mean / expected value / population

$$t\text{-score} = \frac{\bar{x} - \bar{u}}{\frac{s}{\sqrt{n}}}$$

t distribution sums:

If the sample mean and expected mean value of the marks obtained by 15 students in a class test is 290 and 300 resp. what is the t-score if the sd of the marks is 50?

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

$$t = \frac{290 - 300}{\frac{50}{\sqrt{15}}}$$

$$t = \frac{-10}{12.909}$$

$$t = -0.7745$$

3. If the sample mean and expected mean value of ~~marks~~ height obtained by 16 students in a class test is ~~290~~ 170 and ~~300~~ 165 resp. What is the t-score if the sd of the heights is 21.05

$$t = \frac{\bar{x} - \bar{u}}{\frac{s}{\sqrt{n}}} = \frac{170 - 165}{\frac{21.05}{\sqrt{16}}} \quad \frac{5}{5.2625}$$

$$t = 0.950$$

Scanned with OKEN Scanner

Page No
Date

Binomial Distribution

Properties:

1. The experiment consist of a sequence of n identical trials
2. Two outcomes are possible on each trial. we can refer to one outcome as a success and the other outcome as the failure.
3. The probability of a success is denoted by p and does not change from trial to trial. Consequently the probability of a failure is denoted by (1-p) does not change from trial to trial.
4. The trials are independent

Example:

Consider the experiment of tossing a coin 5 times and on each toss observing whether the coin lands with a head or a tail on its uppered face Suppose we have to count the no. of heads appearing over the 5 tosses. Does this experiment show the properties of binomial experiment

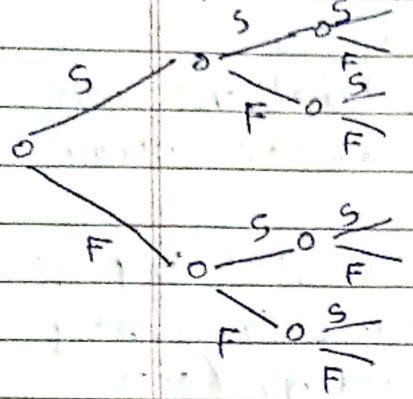
Ans: 1. The experiment consists of 5 identical trials i.e each trial involves the tossing of one coin

2. Two outcomes are possible for a trial head or a tail. head is can be success and tail can be failure.
3. The probability of head and the probability of tail is same for each trial with p as 0.5 (50%)
   $1-p = 0.5$
4. The Trials tossed for tosses are independent bcoz the outcome of any one trial is not affected by the outcome of other trials.

Scanned with OKEN Scanner

Binomial Probability Function

Rushikesh Shubham Nairuty a



S S S	3/3	$\binom{3}{3} = {}^3C_3 = 1$
S S F	2/3	
S F S	2/3	
S F F	1/3	$\binom{n}{x} = \frac{n!}{x!(n-x)!}$
F S S	2/3	
F S F	1/3	
F F S	1/3	$\binom{3}{2} = \frac{3!}{2!(3-2)!} = 3$
F F F	0/3	

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

x is no. of successes

p is the probability of success of one trial

$1-p$ is the probability of failure

n is the no. of trials

$f(x)$ is the probability of x success in n trials

x values

$$f(0) = \binom{3}{0} p^0 (1-p)^{(3-0)}$$

$$\frac{3!}{0!(3-0)!} (0.3)^0 (0.7)^3$$

$$\frac{3!}{3!} \cdot 1 (0.7)^3$$

$$f(0) = 0.343$$

Scanned with OKEN Scanner

$$1 = \binom{3}{1} (p)^1 (1-p)^{3-1}$$

$$= \frac{3!}{1!(3-1)!} (0.3)^1 (0.7)^2$$

$$\frac{3!}{2!} (0.3)(0.7)^2$$

$$\frac{3 \times 2!}{2!} (0.3)(0.7)^2$$

$$3 \; 0.441$$

$$2 = \binom{3}{2} (p)^2 (1-p)^{3-2}$$

$$\frac{3!}{2!(3-2)!} (0.3)^2 (0.7)^1$$

$$\frac{3!}{2!1!} (0.3)^2 (0.7)^1$$

$$3 \; (0.3)^2 (0.7)$$

$$0.189$$

$$3 = \binom{3}{3} (p)^3 (1-p)^{3-3}$$

$$\frac{3!}{3!(3-3)!} (0.3)^3 (0.7)^0$$

$$1 \; (0.3)^3 \; 1$$

$$3 = 0.027$$

Adding $f(0) + f(1) + f(2) + f(3)$

$0.343 + 0.441 + 0.189 + 0.027$

$= 1$

Scanned with OKEN Scanner

Page No.
Date

Suppose for binomial experiment $n = 10$ $x = 4$
Probability of success is 30%

$$f(4) = \binom{10}{4} (0.3)^4 (0.7)^6 = 0.2001$$

Therefore the probability of making exactly 4 sales to 10 customers entering the store is 20%.

An agent sales life insurance to 5 equally aged, healthy people according to the recent data the probability of a person living in this conditions for 30 years or more is 2/3 (success rate). Calculate the probability that after 30 years if

i) all 5 people are still living
ii) atleast 3 people are still living
iii) exactly 2 people are still living

$p = 0.667$
$1 - p = 1 - 0.667 = 0.333$
$n = 5$

i) $x = 5$

$$f(5) = \binom{5}{5} (p)^5 (1-p)^{5-5}$$

$$\frac{5!}{5!(5-5)!} (0.667)^5 (0.333)^0$$

$$\frac{5!}{5!} (0.667)^5 (0.333)^0$$

$$1 \; (0.667)^5 \; (1)$$

$$f(5) = 0.132$$

atmost ($F \leq 3$)

atleast

ii) $F(x \geq 3) = F(3) + F(4) + F(5)$

$F(3) = \binom{5}{3} (0.667)^3 (0.333)^2$

$5! \quad (0.667)^3 (0.333)^2$

$\overline{(5-3)!}\,3!$

$5! \quad (0.667)^3 (0.333)^2$

$\overline{2!\,3!}$

$\frac{5 \times 4 \times 3!}{2!\,3!} (0.667)^3 (0.333)^2$

$\frac{20}{2} (0.667)^3 (0.333)^2$

$10 \; (0.667)^3 (0.333)^2$

10×0.032

0.32

$F(4) = \binom{5}{4} (0.667)^4 (0.333)^{5-4}$

$5! \quad (0.667)^4 (0.333)^1$

$\overline{(5-4)!}\,4!$

$5! \quad (0.667)^4 (0.333)^1$

$\overline{1!\,4!}$

$\frac{5 \times 4!}{4!} (0.667)^4 (0.333)^1$

$5 \; (0.667) (0.333)$

0.32

$F(5) = \binom{5}{5} (0.6667)^5 (0.333)^{5-5}$

Scanned with OKEN Scanner

Page No.
Date

Q In 800 families with 4 children each, classify according to given criteria how many families would you expect to have

i) 2 boys and 2 girls
ii) atleast 1 boy
iii) no girl

Chi Square Test.

The chi square distribution is typically concerned with counts of subject or items falling into categories. The chi square distribution is the probability distribution used in statistics for hypothesis testing and comparing observed and expected values, For example you might want to test wether one variable (representating gender) is independent of another (boss promoted in job) and stat. The statistics that measure the extent to which result departs from the null expectation is the chi square statistic. The

divided by the square root of expected value

Notin IP-I

F distribution

The f statistic measures the means which are greater than the expectation under normal random variation. It is the ratio of variability among the grouped means to the variability within each group.

Scanned with OKEN Scanner

Page No.
Date

tr
IA-1

Poisson distribution:
The poisson distribution tells us the distribution of events per unit of time or space when we sample many such units
Ex: It is useful when addressing queuing questions like (how much capacity do we need to be 95% sure of fully processing the internet traffic that arrives a server in any 5 sec period)
$P(X=x) = \frac{e^{-m} m^x}{x!}$

Exponential distribution:
Using the same parameter λ that we use in the poisson distribution, we can also model the distribution of time between events i.e
Ex1: Time between visits to a website
Ex2: Cars arriving at a toll plaza
Timebtw

Weibull distribution:
The weibull distribution is an extension of the exponential distribution in which the event ~~rate~~ rate is allowed to change as specified by a shaped parameter. (β) If $\beta > 1$ the probability of an event increases over time. If $\beta < 1$, the probability decreases over time. Because the weibull distribution is used with time to failure analysis instead of event ~~rate~~ rate the second parameter η (eta) Greek letter) is expressed in terms of characteristics life instead of rate of events per interval.