DA Assignment - 1

Name: Dheeraj Chaudhary
Roll: 17BCS009
Date of sub: 2/03/2020

Descriptive Statistics:-

Descriptive statistics are brief descriptive coefficient that summarize a given data set, which can either a representation of the entire or a sample of a population. Descriptive statistics are broken down into measures of Central Tendency and measures of variability (spread). Measures of central Tendency including the mean, median, and mode, while measures of variability include the standard deviation, variance, the minimum and maximum variables and the skewness.

e.g. Find mean median & mode of the following data.

23, 29, 20, 32, 23, 21, 33, 25

mean = $\frac{23 + \ldots + 25}{8} = 25.75$

median = 20, 21, 23, [23, 25], 29, 32, 33

$= \frac{23+25}{2} = \frac{48}{2} = 24$

mode = 23

Standard deviation $\Rightarrow \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{N}(x_i - \mu)^2}$

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

$= \frac{(23-25.75)^2 + \ldots + (25-25.75)^2}{8} = \frac{173.5}{8}$

$\sigma = \sqrt{21.68} = 4.65$

Variance, $\sigma^2 = 21.68$

Graphical Methods in Statistics :-

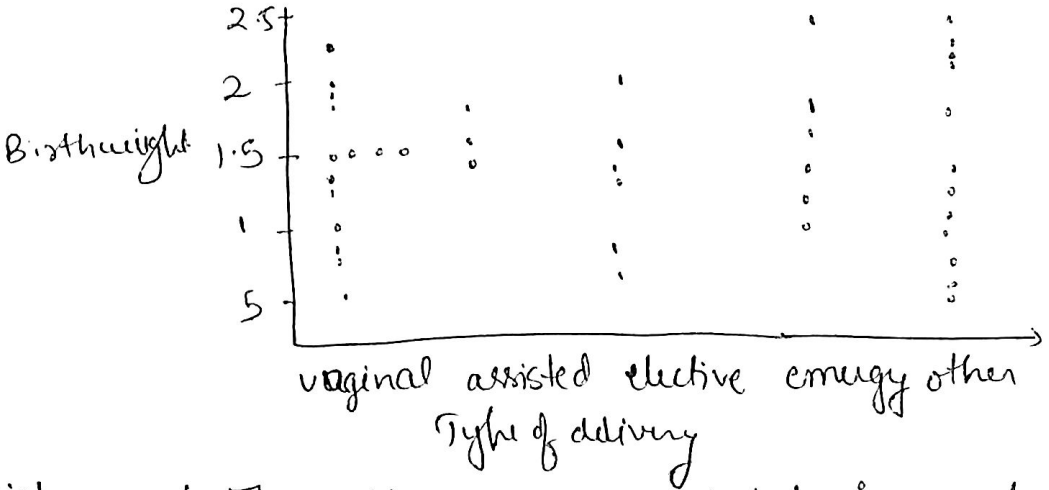
Few graphical methods are-

1⇒ **Dot Plots**:- The simplest method of conveying as much information as possible is to show all of the data.

e.g.

Data on birth weight and type of delivery.

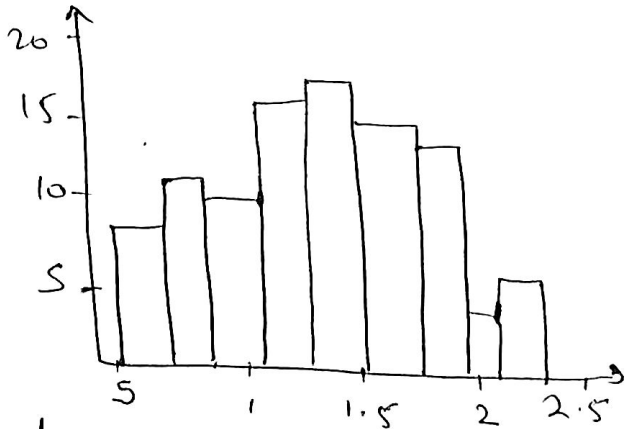
below Dot plot showing birth weight of 98 babies by type of delivery



2- **Histogram**:- The pattern may be revealed in a large dataset of a numerically continuous variable by forming a histogram.

e.g.

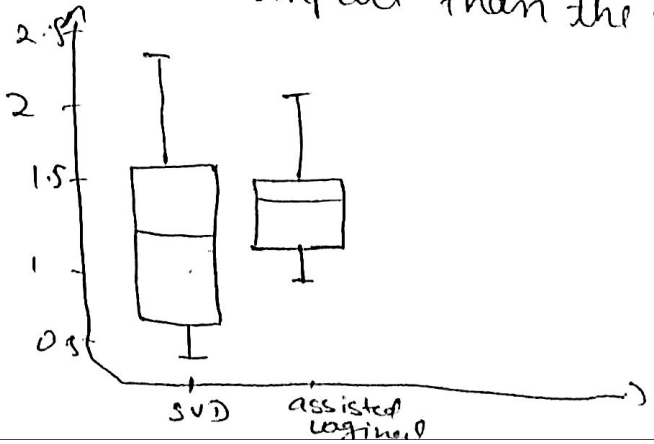
Histogram of birth weight of 98 babies.



3- **Box-Whisker Plot**:

If the number of point is large, a Dot plot can be replaced by a box-plot which is more compact than the corresponding histogram.

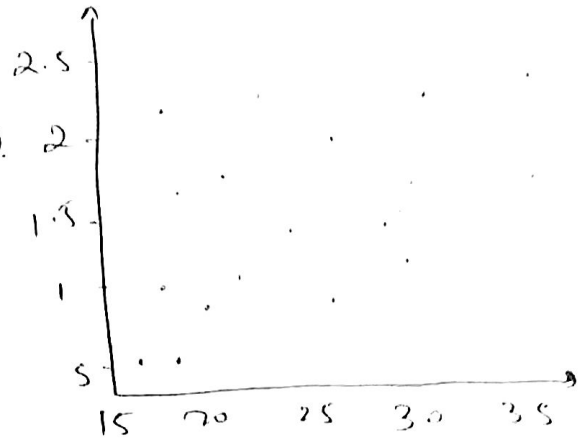
e.g. Box plot of birth weight of babies by method of delivery



3- Scatterplots:- When one wishes to show a relationship b/w two continuous variables a scatterplot can be employed.

e.g

Scatterplot of birth weight by maternal age. (Simpson 2004).

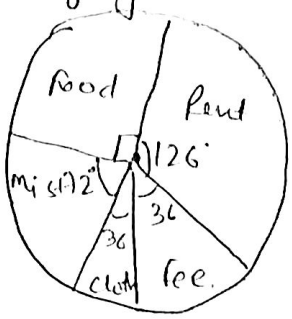


4⇒ Pie-chart

In this, a circle is divided into slices, such that each slice represent a different category and the size of each slice is proportional to relative freq. of that category.

e.g.

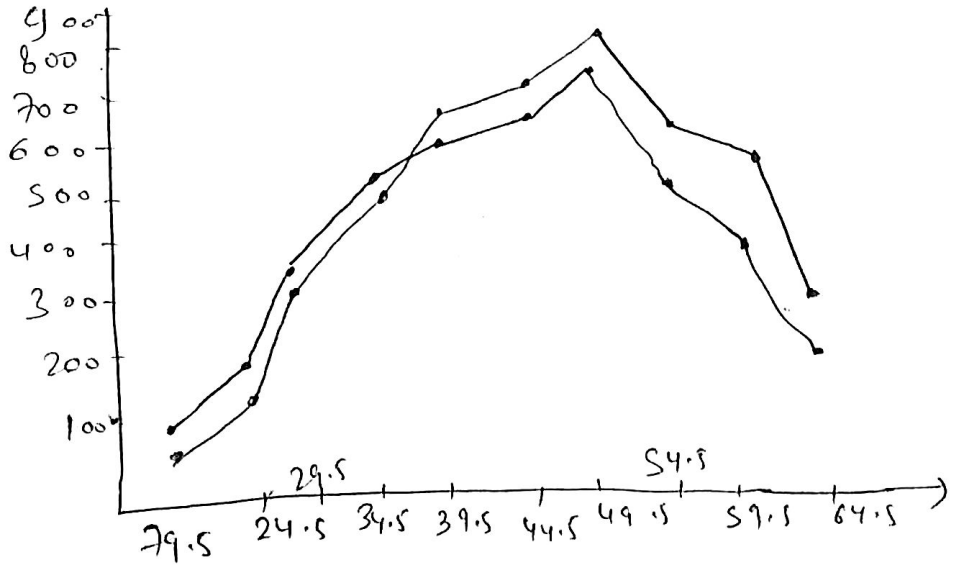
	A
Food	25%
Rent	35%
clothy	10%
Fee	10%
Misth	20%



5⇒ Freq. Polygon ⇒ It is used to join the midpoints of each interval, or bin.

e.g.

Age	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69
freq Mumbai	35	146	334	631	844	909	939	772	766	510
freq Den	48	168	358	635	803	773	845	617	587	362



Co-variance:

Co-variance is a measure of how much two random variables vary together. It is similar to variance, but where variance tells you how a single variable varies, co variance tells you how two variable vary together.

formula :-

$$Cov(X,Y) = \Sigma E((X-\mu)E(Y-\nu))/n-1, \text{ where}$$

$X \Rightarrow$ random variable.

$E(X)$ & $E(Y) \Rightarrow$ expected value of mean of random variable of X & Y resp.

e.g. Calculate co-variance of

X : 2.1, 2.5, 3.6, 4.0 (mean = 3.1)

Y : 8, 10, 12, 14 (mean = 11)

Putting in formula :-

$= (2.1-3.8)(8-11)+(2.5-3.1)(10-11)\cdots /4-1)$

$= (-1)(3) + \cdots + 0.9(3)/3 \Rightarrow 3+5+0.6+2.7/3$

$= 6.8/3 = 2.267.$

The result is +ve, means variance are +vely related

Coefficient of Variance & Corelation coefficient

Coeffn. of Var :- It is also known as relative standard deviation, is a standardized measure of dispersion of a prob. distr. or freq. distribution.

co-rel. coefficient :- As co-variance only tells about the dirn which is not enough to understand the relationship completely, we divide the co-variance with a stand. dev. of x & y resp. & get co-relation coeff. which varies b/w -1 to +1.

formulas :

Coeff of Var :- $\frac{S.D}{mean} \times 100$

Corel^n coeff :- $\frac{-cov(X,Y)}{SD_x \, SD_y}$

$$r = \frac{\frac{1}{n-1}\left[\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})\right]}{\sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n}(x-\bar{x})^2\right]} \times \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n}(y-\bar{y})^2\right)}}$$

e.g. Compute the coef. of corelation & coef. of variance of X & Y.

X: 20 50 70 40 10

Y: 400 200 100 200 500

	X	Y	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
	20	400	-18	120	324	14400	-2160
	50	200	12	-80	144	6400	-960
Σ	70	100	32	-180	1024	32400	-5760
	40	200	2	-80	4	6400	-160
	10	500	-28	220	784	48400	-6160
	mean = 38				2280	108000	-15200

$$\rightarrow \text{Coef of corel}^n = \frac{\frac{1}{4}(-15200)}{\sqrt{\frac{1}{4}(2280)}\sqrt{\frac{1}{4}(108000)}} = \frac{-15200}{\sqrt{246240000}}$$

$$= \frac{-15200}{156903} \Rightarrow -0.968$$

$$\rightarrow \text{Coeff of Var. for } X = \frac{\sqrt{\frac{1}{4}2280}}{38} = 0.628$$

$$\rightarrow \text{Coef of var for } Y = \frac{\sqrt{\frac{1}{4}(108000)}}{700} = 0.234$$

Rank Co-relation

It is any of several statistics that measures an ordinal association - the relationship b/w ranking of diff ordinal variable or diff. ranking of same variable, where a ranking is the assign of the ordering labels 'first', 'second' etc to diff observation of a particular variable. It measure the degree of similarity b/w ranking, and can be used to access the significance of the relation

e.g find the ~~coef~~ Rank co-relⁿ from the following data.

X	R_1	Y	R_2	$R_1 - R_2$
48	3	15	4	-1
34	5	15	4	1
40	4	24	1	3
12	10	08	9	1
16	8	13	6	2
16	8	06	10	-2
66	1	20	2	-1
25	6	09	7.5	-1.5
16	8	09	7.5	0.5
57	2	15	4	-2
				& 27.5

$\uparrow$

$\Sigma (R_1 - R_2)^2$

$$\Rightarrow R = 1 - \frac{\left[6\left\{\Sigma[d_i^2] + \left[m\left(\frac{m^2-1}{12}\right)\right]\right]\right.}{n(n^2-1)}$$

$$\Rightarrow 1 - \frac{\left[6\left\{27.5 + 2 + 2 + \frac{1}{2}\right]\right.}{990} = 1 - \frac{192}{990} = 0.806$$

Reg. rank x on y, y on x

for y on x

$$y = a + bx$$
$$\Sigma y = na + b\Sigma x$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

for x on y

$$x = a + by$$
$$\Sigma x = na + b\Sigma y$$
$$\Sigma xy = a\Sigma y + b\Sigma y^2$$

C.g. Calculate the reg. line from the following data by the method of L. Squar.
Y on X as well as Y on X.

X	Y	XY	X^2	Y^2
9	1	9	81	1
8	2	16	64	4
10	3	30	100	9
12	4	48	144	16
11	5	55	121	25
13	6	78	169	36
14	7	98	196	49
16	8	128	256	64
15	9	135	225	81
$\Sigma X = 108$	$\Sigma Y = 45$	$\Sigma XY = 598$	$\Sigma X^2 = 1356$	$\Sigma Y^2 = 285$

Y on X from the formula.

$$45 = 9a + 108b \quad -①$$

$$598 = a(108) + 1356b \quad -②$$

Solving ① & ②

⇒ ① becomes. $5 = a + 12b$

$$\Rightarrow a = 5 - 12b \quad -③$$

Put ③ in ②

$$\Rightarrow 598 = 540 - 1296b + 1356b \quad \Rightarrow b = \frac{58}{60} = 0.96$$

Put value of b in ③

$$\Rightarrow a = -6.52$$

$$\therefore \boxed{Y = 0.96X - 6.52}$$

Similarly. X on Y, from the formula.

$$108 = 9a + 45b \quad -①$$

$$598 = 45a + b(285) \quad -②$$

from ① we get $a = 5 - 12b$ -③

Put ③ in ②

we get,

$$598 = 45(5a - 12b) + 285b$$

$$598 = 225 - 540b + 285b$$

$$598 - 225 = -255b$$

$$373 = -255b \Rightarrow b = -1.46$$

Put value of b in ③

$$a = 5 - 12(-1.46)$$

$$a = 5 + 17.53 \Rightarrow 22.553$$

$$\therefore \boxed{X = 22.55\$ - 1.464}$$