Revision

$\bar{x}$ (Arthemtic mean)

$\bar{x} = \frac{\sum x}{n}$ (sample mean) $\xrightarrow{\text{estimate of}}$ μ (population mean)

Median (middle number of series when ordered)

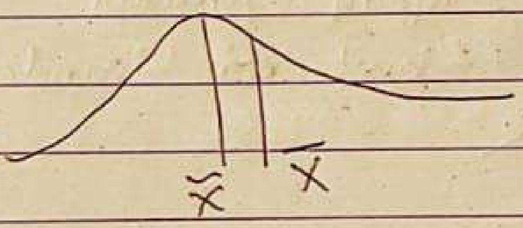
n = odd (middle no. is series)

n = even (mean of two middle number).

In symmetric distrubution the mean is roughly equal to median.

But, In assymetric distrubution mean is greater than median.

mean > median (due to outliers)



$\bar{x} > \tilde{x}$ (median)

Mode

- data having highest frequency

eg:- [28, 28, 5, 10]

mode = 28

[28, 5, 10, 15]

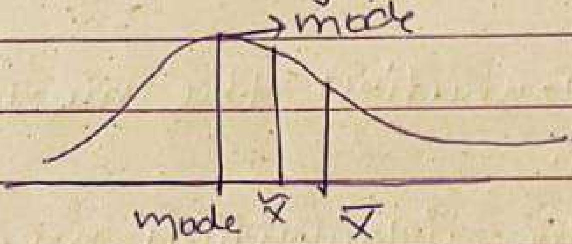
mode = ?

there is no significance of mode in small datasets.

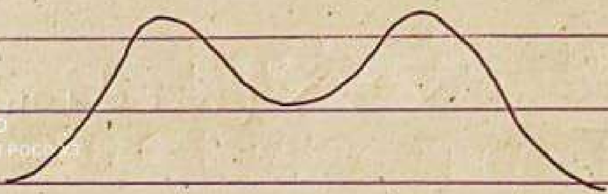
mode signifies in large dataset like census of country.

2. the three central tendancy measure, mean, median, mode which one is good representative is totally depend upon context of problem (nature of dataset and problem).

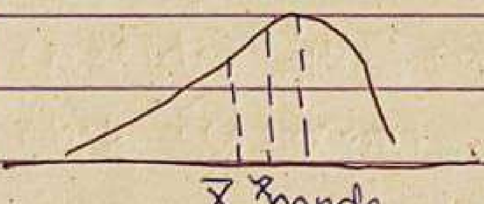
In positively skewed data.



$$\text{mode} < \underset{(\text{median})}{\tilde{x}} < \bar{x}$$



mean = median
but ~~.~~ bimodal



negatively skewed data

$\bar{x} < \tilde{x} <$ mode

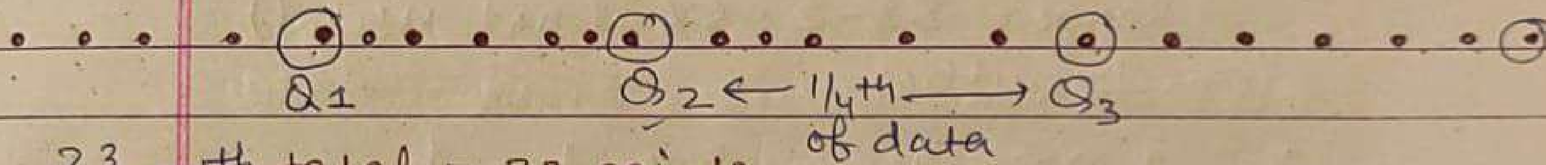
Quantiles

In a ordered dataset
How can you describe the useful locations on dataset.

[4, 7, 8, - - -, 178, 180, 189]

* the median is the observation that is half way through the ordered dataset.

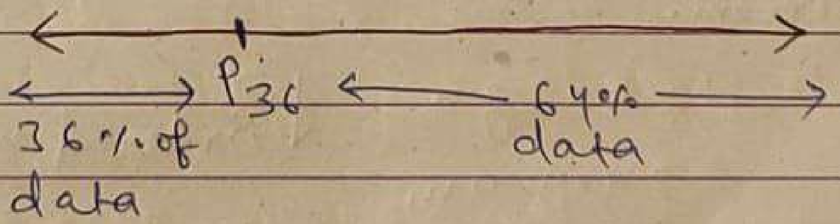
Q1 (1st Quartile) : Observation that is one through Quarter of the way through dataset.



$\frac{23}{4} \approx 6^{th}$ point
total = 23 points

percentile : One percent of the way through the ordered dataset.

$100 - 36 = 64$



Ranges

Consider an ordered dataset

Range = (max − min)

try to give us idea on the spreadness of data.

but it is more significant when dataset not containing any outlier.

but when data containing outlier there is no significance of Range.

2 2 5 6 9 10 58

Range = 58 − 2 = 56 (not useful)

58 → outlier

∴ IQR term introduced

$IQR = Q_3 - Q_1$

IQR (inter Quaratile range)

Q. consider the sugar content in one scoop of each icecream.

A 26.1 gm ↓ ordered way

B 24.9 gm

— — —

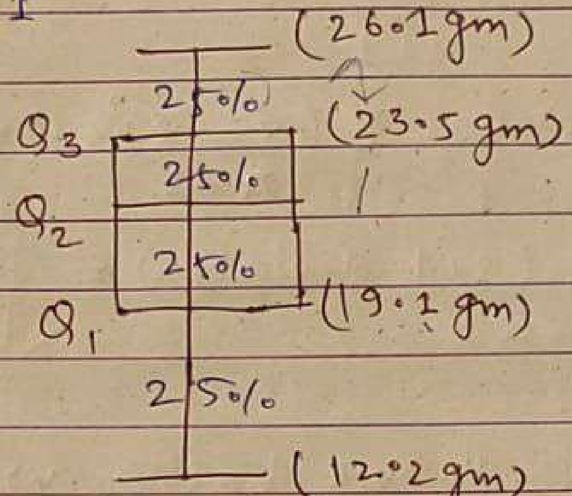
— — —

Z 12.2 gm

$$Range = 26.1 - 12.2$$
$$= 13.9 \text{ gm}$$

(max^m spreadness of data b/w point is 13.9gm)

$$IQR = Q_3 - Q_1$$
$$= 4.4 \text{ gm}$$



give much more idea about spreadness.
It gives % of data in specified range
by drawing wishker plot.

→ variance and std deviation

consider a sample of population

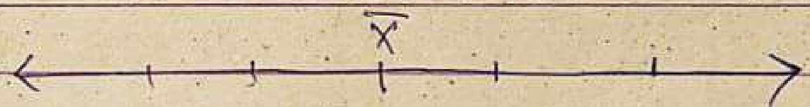
$$\bar{x} = \frac{\Sigma x}{n}$$

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

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

Q. why do we bother with variance?

objective :- describe spread of data



Let's find avg. deviation from the mean

$$\Sigma (x-\bar{x}) = 0$$

for right points from $\bar{x}$ = +ve deviation
for left point from $\bar{x}$ = −ve deviation

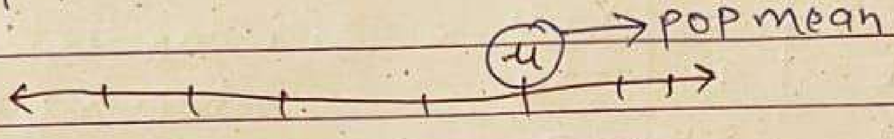
If we take $|x-\bar{x}|$ then this function is not differentiable.

∴ Average square devation from the mean.

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

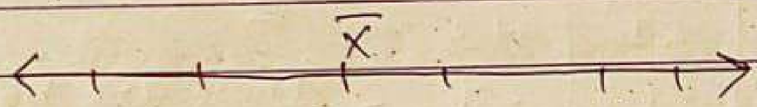
Q. why did we divide by $n-1$?

the variance is the avg. square deviation from the population mean.



$$\text{variance} = \frac{\sum (X-\mu)^2}{N} \quad N = \text{no. of data point}$$

But we don't have population mean we estimate it using sample ($\bar{x}$)



$$s^2 = \frac{\sum (x-\bar{x})^2}{??}$$

* the sample mean is one possible position for the true population mean.

* At any other position, the sum of square would be larger.

$$s^2 = \frac{\sum (x-\bar{x})^2}{n-1}$$ (smaller denominator adjust the varian upward)

* but why $n-1$ not other number this thing is empirically proven

→ Degree of freedom

$\mu = 53$

Obs	x	$x-\mu$
1	41	-12
2	59	$+6$
3	50	-3

3 indepedent observation

DOF = 3

$\bar{x} = 58$

Obs	x	$x-\bar{x}$
1	61	$+3$
2	51	-7
3	—	—

these are bound to value to specify $\frac{\sum x}{n} = \bar{x}$ eq^n.

DOF = 2

$$\sigma^2 = \frac{\sum (x-\mu)^2}{N}$$
(pop variance)

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

DATE | | PAGE

→ Cofficent of variation

$$CV = \frac{s}{\bar{x}}$$

$x = [1, 2, 3]$ $\bar{x} = 2$, $s_x = 1$

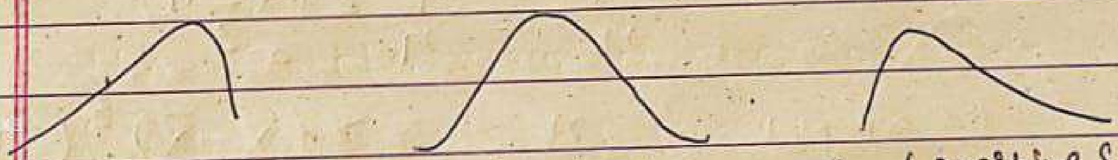
$y = [101, 102, 103]$ $\bar{y} = 102$ $s_y = 1$

$$CV(x) = \frac{1}{2} = 0.5$$

$$CV(y) = \frac{1}{102} = 0.0098$$

Scaling the variance with respect to mean(dataset)
on finding CV units don't matter for two different
datasets because it get scaled w.r.t to dataset.

→ Skewness



(negatively skew) (No skew) (positive skew)

+ve skew = mode < median < mean
-ve skew = mean < median < mode

* the greater the skew the greater distance bewteen mode, median and mean.

pearson method

$$\text{mode skewness} = \frac{\text{mean} - \text{mode}}{\text{std.dev}}$$

$$\text{median skewness} = \frac{3(\text{mean} - \text{median})}{\text{std.dev}}$$

$$\text{mode} \cong 3\,\text{median} - 2\,\text{mean}$$

Moment based calculation

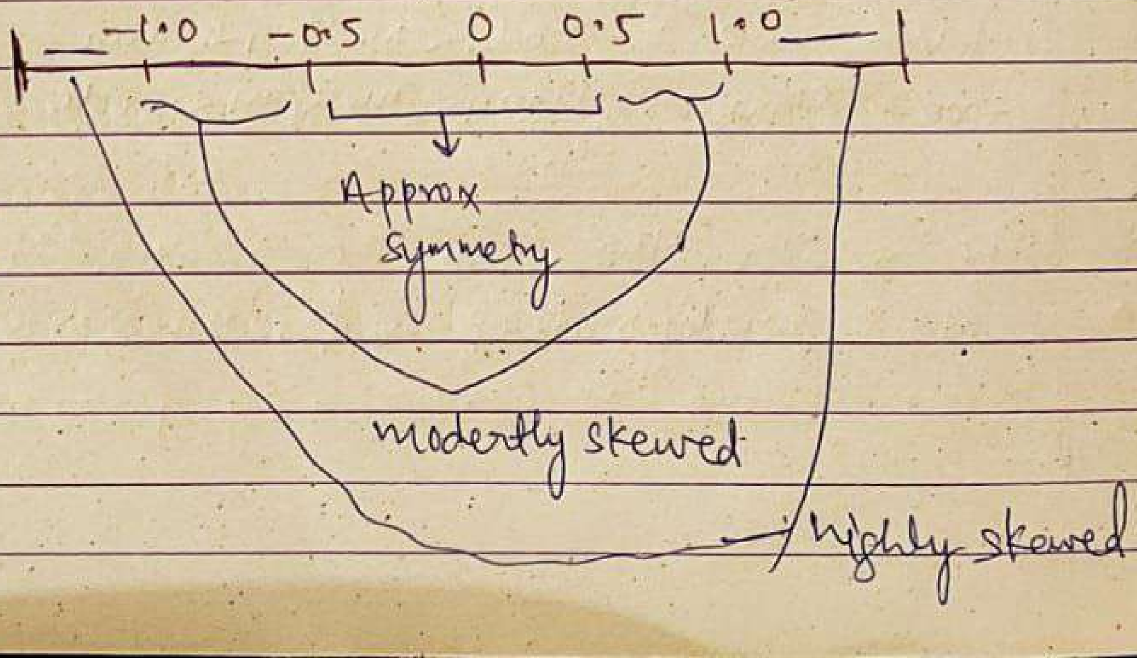
$$\text{First moment} = \frac{\Sigma x}{n}$$

$$\text{Second moment} = \frac{\Sigma x^2}{n} = \frac{\Sigma (x-\mu)^2}{n} \text{ (population)}$$

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

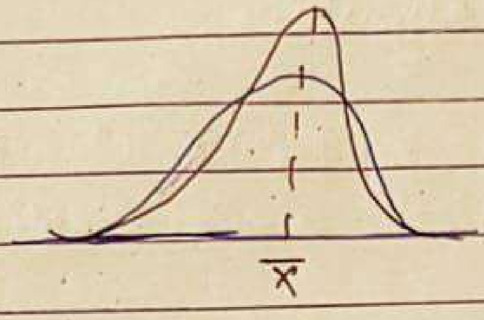
$$\text{Third moment} = \frac{\Sigma x^3}{n} = \frac{1}{n} \frac{\Sigma (x-\mu)^3}{\sigma^3}$$

$$\text{(sample)} = \frac{n}{(n-1)(n-2)} \frac{\Sigma (x-\bar{x})^3}{s^3} \text{ (skew)}$$



DATE | | PAGE

→ Kurtosis



$\sigma = 5$

skew $= 0$

$\bar{x} = 0$

the black curve is more peaked and has fatter tail.

fourth moments $= \frac{1}{n}\left(\frac{\Sigma(x-\mu)^4}{\sigma^4}\right)$

$$= \frac{n(n+1)}{(n-1)(n-2)(n-3)}\left(\frac{\Sigma(x-\bar{x})^4}{s^4}\right) - \frac{3(n-1)^2}{(n-2)(n-3)}$$

A normal distribution has kurtosis of 3 called mesokurtic

Kurtosis > 3 (leptokrutic)

Kurtosis < 3 (platykurtic)

Kurtosis ranges from (1 to ∞)

(Excess kurtosis = kurtosis $- 3$)

(normal dist)

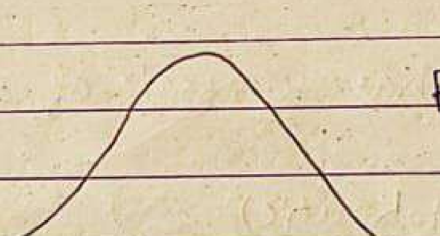
Excess kurtosis ranges from -2 to ∞ and normal distribution is 0.

* Kaplansky (1945) discovered that there were certain anomolies to the higher peak = Fatter tail relationship.

$$p(x) = \frac{1}{3\sqrt{\pi}}\left(\frac{9}{4} + x^4\right) e^{-x^3}$$ (plot the distribution)

→ "Kurtosis as peakdness" (incorrect notation)

* Kurtosis tells you virtually nothing about the shape of the peak - its only unambigious interpretation is in terms of tail extremility. i.e either existing outlier. (for the sample Kurtosis) or propensity to produce outliers for the Kurtosis of probablity distribution.



Fourth Standarised moment.

$$\frac{1}{n}\left(\Sigma \frac{(x-\mu)^4}{\sigma^4}\right)$$

$$= \frac{1}{n}\left(\frac{\Sigma(x-\mu)^4}{\Sigma(x-\mu)^2/n)}\right)$$

Outliers contribute greatly to this Summation.

"Kurtosis as tail extremity".

DATE | | PAGE

→ covariance and correlation
Describes the relationship between two numerical variable.
"No covariance No correlation"

$$cov(x, y) = \sigma_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

cov = +ve positively related
cov = -ve negatively related

x point $\bar{x}$ से positive deviation पे है। y point $\bar{y}$ से positive deviation है then
$(x - \bar{x})(y - \bar{y}) \rightarrow +ve$ both are on +ve deviation

x point $\bar{x}$ से negatively deviated है। y positively $\bar{y}$ से।
$(x - \bar{x})(y - \bar{y}) \rightarrow -ve$
x → neg y → pos
(-ve relation)

we can also see from var

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

$$= \frac{\sum (x - \bar{x})(y - \bar{y})}{n-1}$$

DATE | | PAGE

covariance doesn't describe strength of relationship

$$CORR(x,y) = \rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$(-1 \leq \rho \leq 1)$

corr(x, y) tells about strength of relationship.

→ standard error

the standard error of stastics (usually an estimate of a parameter) is the standard deviation of its sampling distrubution.
if the stastics is the sample mean, it is called standard error of the mean.

The sampling distribution of the mean is genrated by repeated sampling from the same population and recording of the sample mean obtained. this forms a distribution of different mean and this distribution has its own mean and variance.
Mathematically,
the variance of the sampling mean distribution obtained is equal to the variance of the population divided by the sample size.

∴ sample size ↑ → sample mean more around pop mean

Standard error of mean $= \frac{\sigma}{\sqrt{n}}$ σ = std. dev of population

In other word the standard error is a measure of sample means around the pop mean.