01.1 KHAN ACADEMY : STATISTICS AND PROBABILITY

01 SUMMARIZING QUANTATIVE DATA

~~NEGATIVE~~ 01 MEAN, MODE, MEDIAN

~~PREDICTED~~ STATISTICS

Descriptive

We have bunch of data and we have to describe that data with some number

Inferential

After we have description about the data, we started to take inference from data.

Suppose we have height → 4, 3, 1, 6, 1, 7

Average : "typical" or "middle" or "center" → Central Tendency

Arithmetic Mean → $\frac{4+3+1+6+1+7}{6} = \frac{22}{6} = 3.666$ or $3.\overline{6}$

Other method to find Average is Median - First order : 1, 1, 3, 4, 6, 7.

Middle number : $\frac{3+4}{2} = 3.5$

Another Number : 0, 7, 50, 100, 10000, Median = 50

Third method to find Central Tendency $\xrightarrow{\text{Mode}}$ Most occuring element, so in above example

Mode = (4, 3, 1, 6, 1, 7) = 1 (because 2 1's)

Eg - Find the mean, median and mode of following : 23, 29, 20, 32, 23, 21, 33, 25

Mean = $\frac{23+29+20+32+23+21+33+25}{8} = \frac{206}{8} = 25.75$

So 25.75 is the central tendency of above example.

Median = 20, 21, 23, 23, 25, 29, 32, 33 $= \frac{23+25}{2} = \frac{48}{2} = 24$

24 is another central tendency of above example.

Mode = 23 is also another central tendency of above example

$$\text{Mean} = \frac{\Sigma x_i}{n}$$

02 Inter Quartile Range –

The following data points represents the number of animal crackers in each lunch box. Find the interquartile range. – 4, 4, 10, 11, 15, 7, 14, 12, 6

Sort the data → 4, 4, 6, 7, 10, 11, 12, 14, 15.

Median

Median = $\frac{4+6}{2} = 5$ Median = $\frac{12+14}{2} = 13$.

So, IQR is = 13 – 5 = 8.

Steps 1 – Sort the data.

Steps 2 – Find the median

Step 3 – Find the median for data range less than the median and data range more than median

Step 4 – $IQR = Median_{large} - Median_{small}$ (X_{small}) (X_{large})

IQR – Interquartile Range

- Interquartile range is the amount of spread in the middle 50% of dataset
- In other word, it is the distance between first quartile (Q_1) and third quartile (Q_3)

$IQR = Q_3 - Q_1$

Step 1 – Put the data in ascending order.

Step 2 – Find the median. If the number is odd, the median is the middle data point. If the number of data point is even, the median is the average of middle two points.

Step 3 – Find the first quartile (Q_1). The first quartile is the median of the data points to the left of the median in the order list.

Step 4 – Find the third quartile (Q_3). The third quartile is the median of the data points to the right of the median in the order list.

Step 5 – Calculate the IQR by subtracting $Q_3 - Q_1$.

Example – Find the IQR of essay score in a Class – 1, 3, 3, 3, 4, 4, 4, 6, 6

Step 1 – Data is already sorted.

Step 2 – Find the median. 1, 3, 3, 3, 4, 4, 4, 6, 6.

Median

Median = $\frac{3+3}{2} = 3$ Median = $\frac{4+6}{2} = 5$

Step 3 – $IQR = Q_3 - Q = 5 - 3 = 2$.

01.3 03) Measure of Spread – Range, Variance, Standard Deviation

$-10, 0, 10, 20, 30$ $\quad$ $8, 9, 10, 11, 12$

Mean $= \frac{-10+0+10+20+30}{5}$ $\quad$ Mean $= \frac{8+9+10+11+12}{5} = 10 =$ Mean (Right)

Mean (left) $= 10$ $\quad$ So, both have exactly same mean but they have different

Left side – More dispersed $\quad$ Right side – Less dispersed

Range (left side) $= 30 - 10 = 40$ $\quad$ Range (Right Side) $= 12 - 8 = 4$.

So, Range gives a better insight than Mean in term of "Measure of Spread".

Variance → **Variance $= \sigma^2$ → population variance**

Standard Deviation → Standard Deviation $= \sigma$ → population standard deviation

$$\text{Variance (left)} = \sigma^2 = \frac{(-10-10)^2 + (0-10)^2 + (10-10)^2 + (20-10)^2 + (30-10)^2}{5}$$

$$= \frac{400+100+0+100+400}{5} = \frac{1000}{5} = 200 = \sigma^2 \text{ (left)}$$

$$\text{Variance (Right)} = \sigma^2 = \frac{(8-10)^2 + (9-10)^2 + (10-10)^2 + (11-10)^2 + (12-10)^2}{5} = \frac{10}{5} = 2 = \sigma^2 \text{ (Right)}$$

Variance (Right) $= 2$ is way less dispersed data than Variance (left) $= 200$.

Standard Deviation $= \sqrt{\text{Variance}} = \sqrt{\sigma^2} = \sigma$

SD (left) $= \sqrt{200} = 10\sqrt{2}$. $\quad$ SD (Right) $= \sqrt{2}$.

So, SD (left) is 10 times the SD (Right).

VARIANCE OF A POPULATION

Years of Experience → 1, 3, 5, 7, 14

$$\text{Mean} = \mu = \frac{\sum_{i=1}^{5} x_i}{5} = \frac{1+3+5+7+14}{5} = 6$$

$$\sigma^2 = \frac{(1-6)^2 + (3-6)^2 + (5-6)^2 + (7-6)^2 + (14-6)^2}{5} = \frac{25+9+1+1+64}{5} = 20$$

Mean Square Distance from Population mean is 20.

Population Variance $= \sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$

x_i = Each datapoint
μ = Population mean
N = Total datapoints
σ^2 = Population Variance

POPULATION STANDARD DEVIATION

Example – 4m, 4.2m, 5m, 4.3m, 5.5m

Arithmetic mean = μ = 4.6m.

Population ~~Standard Deviation~~ Variance = $\sigma^2 = \frac{(4-4.6)^2 + (4.2-4.6)^2 + (5-4.6)^2 + (4.3-4.6)^2 + (5.5-4.6)^2}{5}$

$= 0.316\ m^2$, it was in meter (originally) now m^2.

So, $\sqrt{\sigma^2} = \sqrt{\sigma} = \sqrt{0.316} \cong 0.562$ m. = Population Standard Deviation

Population Standard Deviation = σ = Measure of how much data is varying from mean.

- Higher the value, data is varying far from Mean.
- Lower the value, data is varying around the Mean.

MEAN and STANDARD DEVIATION vs MEDIAN and IQR

What is the central tendency → 35, 50, 50, 50, 56, 60, 60, 75, 250

Mean = 76.2 Median = 56

Standard Deviation ≅ 62.3 Interquartile Range ≅ 17.5

Mean = 76.2 is greater than most of the datapoints. because it is skewwed of 250

Median = 56 is much more robust in case of skewwed data.

So, we will choose Median as central tendency.

For Data Spread ? → Standard Deviation ≅ 62.3, SD is not good measure.

Firstly, $\sigma = \sqrt{\frac{\Sigma (x_i - \mu)^2}{N}}$ here μ is mean and mean is affected by outlier.

Second, all the data points are much closer to each other and only one data point is apart. but SD = 62.3

So, use IQR in case of outliers. IQR = 17.5 ~~Range will be 56 ± 17.5~~

~~Mean ± SD = 76.2 ± 62.3~~ ~~Median ± SD IQR = 56 ± 17.5~~

35, 50, 50, 50, 56, 60, 60, 75, 250

50 (Q1) Median 67.5 (Q3)

17.5

04.5 ALTERNATE VARIANCE FORMULA

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

$$= \frac{\sum_{i=1}^{N} (x_i^2 - x_i\mu - \mu x_i + \mu^2)}{N}$$

$$= \frac{\sum_{i=1}^{N} (x_i^2 - 2x_i\mu + \mu^2)}{N}$$

$$= \frac{\sum_{i=1}^{N} x_i^2 - 2\mu \sum_{i=1}^{N} x_i + \mu^2 \sum_{i=1}^{N} (1)}{N}$$

$$= \frac{\sum_{i=1}^{N} x_i^2 - 2\mu \sum_{i=1}^{N} x_i + \mu^2 N}{N}$$

$$\because \mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

$$= \frac{\sum_{i=1}^{N} x_i^2}{N} - \frac{2\mu \sum_{i=1}^{N} x_i}{N} + \frac{\mu^2 N}{N}$$

$$= \frac{\sum_{i=1}^{N} x_i^2}{N} - 2\mu(\mu) + \mu^2$$

$$= \frac{\sum_{i=1}^{N} x_i^2}{N} - 2\mu^2 + \mu^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \mu^2$$

So, alternate formula of Variance $= \boxed{\frac{\sum_{i=1}^{N} x_i^2}{N} - \mu^2 \text{ or } \frac{\sum_{i=1}^{N} x_i^2}{N} - \frac{\left[\sum_{i=1}^{N} x_i\right]^2}{N^2}}$

Example - Find the Variance of $-10, 0, 10, 20, 30$

Method 1 → Mean $= \frac{-10+0+10+20+30}{5} = 10$.

Variance $= \frac{(-10-10)^2 + (0-10)^2 + (10-10)^2 + (20-10)^2 + (30-10)^2}{5} = \frac{400+100+0+100+400}{5} = 200$ //

Method 2 → $\frac{(-10)^2 + 0 + 10^2 + 20^2 + 30^2}{5} - (10)^2 = \frac{100+100+400+900}{5} - 100 = \frac{1500}{5} - 100$

$= 300 - 100$.

$= 200$ //.

SAMPLE VARIANCE

- Suppose who many people watch TV averagely per day?
If we want to take survey, we need to survey 300 million people.

We cannot survey 300 million people, because some people may have died till the survey is done / some people or born after we cross that area / cost of covering the whole population etc etc

So, if we find average mean of total population, then that mean is population mean

Suppose we took sample = 6, watching TV per hour - 1.5, 4, 1, 2.5, 2, 1.

$$\text{Sample Mean} = \bar{x} = \frac{1.5+4+1+2.5+2+1}{6} = \frac{12}{6} = 2.$$

So, sample mean is 2, now we find Variance. Not possible for population.

So find ~~population~~ sample variance $= S_n^2 = \frac{(1.5-2)^2+(4-2)^2+(1-2)^2+(2.5-2)^2+(2-2)^2+(1-2)^2}{6} = 1.08$

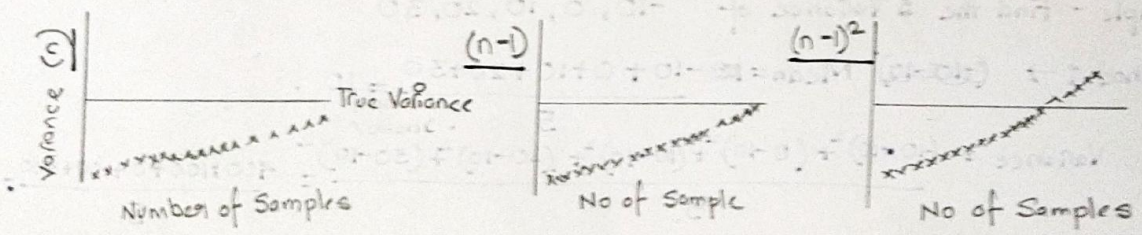
Sample variance $= S_6^2 = 1.08$

If for sample we divide by $n-1$ $= S_{n-1}^2 = \frac{\sum_{i=1}^{n} s^2}{n-1} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$ $n \rightarrow$ no of sample, $N \rightarrow$ no of population

WHY $(n-1)$ for Samples ?

~~How many seeds are there in Watermelon?~~ As we are estimating the population from variance, instead of using 'n' use 'n-1' because it will increase the estimate toward the population. So,

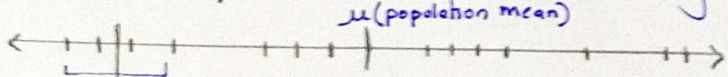
$$\text{Sample Variance} = S_{n-1}^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$



So as the sample increased, sample ~~mean~~ variance converged to true variance.

Over the period $(n-1)$ converge more to True Variance.

If we divide by n, we are underestimating the True Variance.



Suppose we take this sample

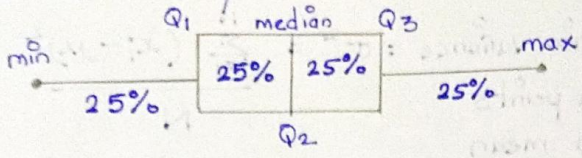
this is sample mean

If we divide by 'n', which is far from population mean

So, if we divide by 'n-1' it will be closer to population mean.

04 Box Plot / Box Whisker Plot

- A box and whisker plot also called box plot.
- It displays the five number summary of a set of data.
- Five number summary is the minimum, first quartile, median, third quartile, max



Eg - Finding the five-number summary

A sample of 10 boxes of raisins has weights: 25, 28, 29, 29, 30, 34, 35, 35, 37, 38

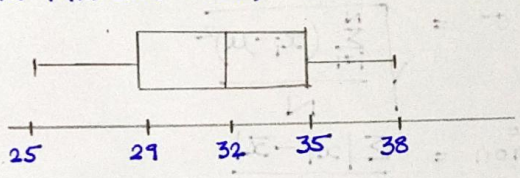
Step 1: Order the dataset, our data is already in order.

Step 2: Find the median, $\frac{30+34}{2} = 32$

Step 3: Find the quartiles, 25, 28, 29, 29, 30 $Q_1 = 29$ ← left side

Right side, 34, 35, 35, 37, 38 $Q_3 = 35$

Step 4: Find the Min and Max, Minimum = 25, Maximum = 38



05 Mean Absolute Deviation (MAD)

Measure of Center → Mean Measure of Variability → Mean absolute deviation

On average how far each datapoint are from Mean.

2, 2, 4, 4.

$\text{Mean} = \frac{2+2+4+4}{4} = \frac{12}{4} = 3.$

1, 1, 6, 4.

$\text{Mean} = \frac{1+1+6+4}{4} = \frac{12}{4} = 3.$

Even if both the distribution is different their mean is same.

$\text{Mean Absolute Deviation (Left)} = \frac{|2-3| + |2-3| + |4-3| + |4-3|}{4} = \frac{4}{4} = 1.$

So, averagly mean distance from each data point is 1 from Mean

$\text{Mean Absolute Deviation (Right)} = \frac{|3-1| + |3-1| + |6-3| + |4-3|}{4} = \frac{8}{4} = 2.$

So, mean absolute deviation = 2

So, even if Mean is equal there Mean Absolute Deviation is different. And therefore, right one (2) is more spread out than left (1)

$$\text{Mean Absolute Deviation} = \frac{\sum |x_i - \bar{x}|}{n}$$

SUMMARIZE QUANTITATIVE DATA FORMULA REVISION

Mean = $\frac{\sum_{i=1}^{n} x_i}{n}$

Inter Quartile Range (IQR) = $Q_3 - Q_1$

Variance = σ^2, Population variance = $\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$

$x_i \rightarrow$ Each data points
$\mu \rightarrow$ Population mean
$N \rightarrow$ Total datapoints
$\sigma^2 \rightarrow$ Population Variance

Other formula, Variance = $\frac{\sum_{i=1}^{N} x_i^2}{N} - \mu^2$ or $\frac{\sum_{i=1}^{N} x_i^2}{N} - \frac{\left[\sum_{i=1}^{N} x_i\right]^2}{N^2}$

Sample Variance, $S_{n-1}^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{n-1}$

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

Mean Absolute Deviation = $\frac{\Sigma |x_i - \bar{x}|}{N}$