Session-14 Statistics

03 August 2024 02:49 PM

Agenda

- **Data**
- **Type of Data**
- **Quantitative vs Qualitative**
- **Frequency and Cumulative Frequency**
- **Measure of Frequency**
- **Measure of Central Tendency**
- **Measure of Dispersion**
- **Variance and Standard Deviation**

{mean, median} mode

spread

normal distribution

Q. What is Data?

or piece of information

Ans: Data is a collection of facts, such as numbers, words, measurements, observations or just descriptions of things.

Height: {170, 160, 155 ...}

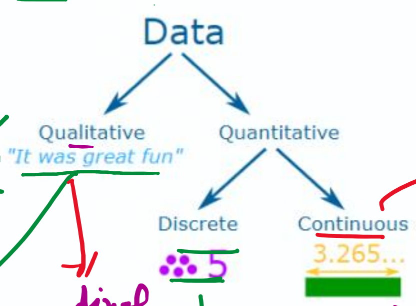
count

Qualitative vs Quantitative

Data can be qualitative or quantitative.

string, category

- **Qualitative data** is descriptive information (it *describes* something)
- **Quantitative data** is numerical information (numbers)



eg Gender

Blood group

Pincode

Nominal

ordinal

count

measure

customer feedback

Best, Good, Bad ...

Quantitative data can be Discrete or Continuous:

- **Discrete data** can only take certain values (like whole numbers)
- **Continuous data** can take any value (within a range)

Put simply: **Discrete data** is counted, **Continuous data** is measured

Numerical data

categorical data

Data

(a) Quantitative

(b) Qualitative

Discrite | Continuous

Nominal | ordinals

(4) (i) Discrete eg whole No or exact No. } count

eg (i) Bank A/c No.

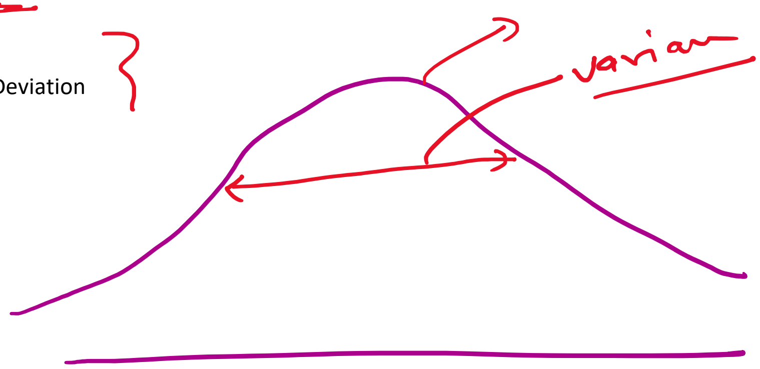
(ii) No. of children in a family.

(ii) Continous → Any value (measure)

eg → weight, height, temperature, speed.

Measures of Dispersion
(Spread Of Data)

1. Variance
2. Standard Deviation



Variance

The Variance is defined as:

> The average of the **squared** differences from the Mean.

To calculate the variance follow these steps:

- Work out the Mean (the simple average of the numbers)
- Then for each number: subtract the Mean and square the result (the *squared difference*).
- Then work out the average of those squared differences. (Why Square?)

Population variance | Sample variance

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

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

x_i = Data Point
μ = Population mean
N = Population size

x_i = Data Point
$\bar{x}$ = Sample mean
n = sample size

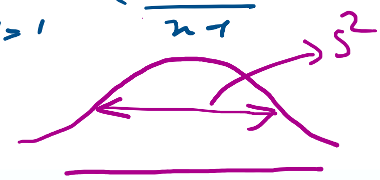
Q: why we divide sample variance by $(n-1)$?

Ans = The sample variance is devided by $(n-1)$ so that we can create an unbiased estimator of the population data

Bessel correction

eg $\{1, 2, 3, 4, 5\}$

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



x	$\bar{x}$	$(x_i - \bar{x})^2$
1	3	4
2	3	1
3	3	0
4	3	1
5	3	4
$\bar{x} = 3$		10

$s^2 = \frac{10}{5-1}$

$\boxed{s^2 = 2.5}$

$s^2 \uparrow \propto$ spreed graph

Standard Deviation

The Standard Deviation is a measure of how spread out numbers are.

Its symbol is σ (the greek letter sigma)

The formula is easy: it is the **square root** of the **Variance.** So now you ask, "What is the Variance?"

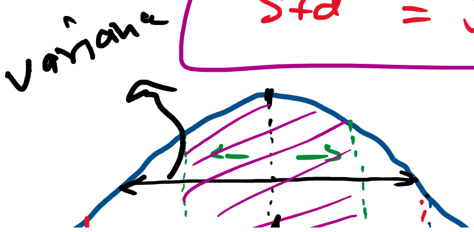
Population std

$$\boxed{\sigma = \sqrt{\text{variance}}}$$

Sample std

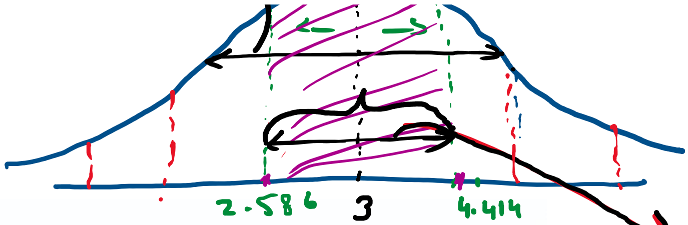
$$\boxed{\text{std} = \sqrt{s^2}}$$

$\bar{x} = 3$
$\sigma = 1.44$



3 ± 1.44

$\sigma = 1.44$



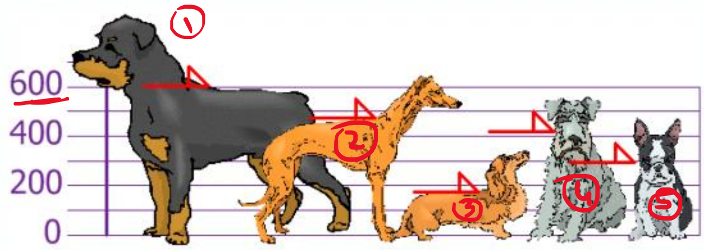
3 + 1.44

3 − 1.44

1st std ≈ 68%

Example

You and your friends have just measured the heights of your dogs (in millimeters):



The heights (at the shoulders) are: 600mm, 470mm, 170mm, 430mm and 300mm.

Find out the Mean, the Variance, and the Standard Deviation.

Your first step is to find the Mean:

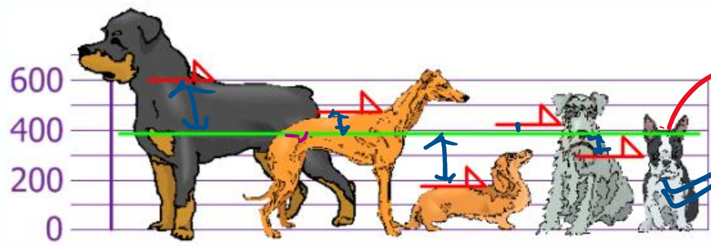
Answer:

$$\text{Mean} = \frac{600 + 470 + 170 + 430 + 300}{5}$$

$$= \frac{1970}{5}$$

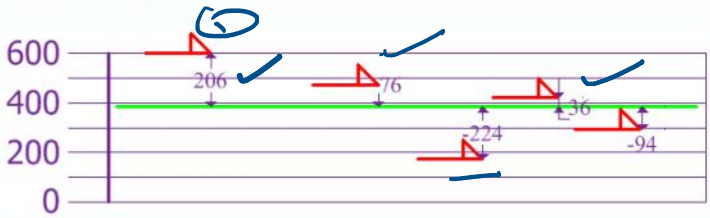
$$= 394$$

so the mean (average) height is 394 mm. Let's plot this on the chart:



394 mean (Avg height)

Now we calculate each dog's difference from the Mean:



To calculate the Variance, take each difference, square it, and then average the result:

Variance

$$\sigma^2 = \frac{206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5}$$

$$= \frac{42436 + 5776 + 50176 + 1296 + 8836}{5}$$

$$= \frac{108520}{5}$$

$$= \frac{42436 + 5776 + 50176 + 1296 + 8836}{5}$$

$$= \frac{108520}{5}$$

$$= \boxed{21704}$$

So the Variance is **21,704**

And the Standard Deviation is just the square root of Variance, so:

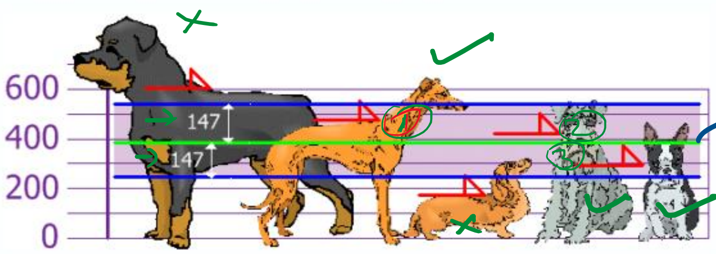
Standard Deviation

$$\sigma = \sqrt{21704}$$

$$= 147.32...$$

$$= \mathbf{147} \textit{ (to the nearest mm)}$$

And the good thing about the Standard Deviation is that it is useful. Now we can show which heights are within one Standard Deviation (147mm) of the Mean:



mean

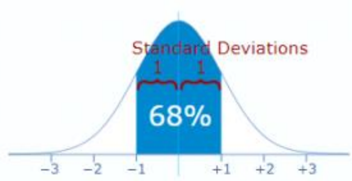
68%

So, using the Standard Deviation we have a "standard" way of knowing what is normal, and what is extra large or extra small.

Rottweilers **are** tall dogs. And Dachshunds **are** a bit short, right?

Using

We can expect about 68% of values to be within plus-or-minus 1 standard deviation.



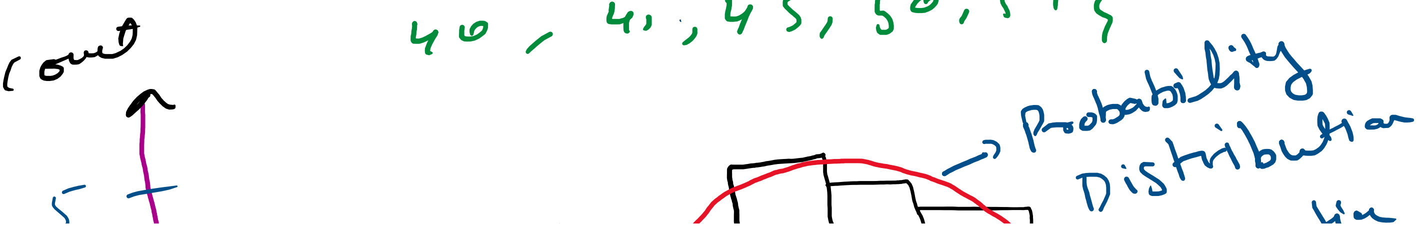
Read Standard Normal Distribution to learn more.

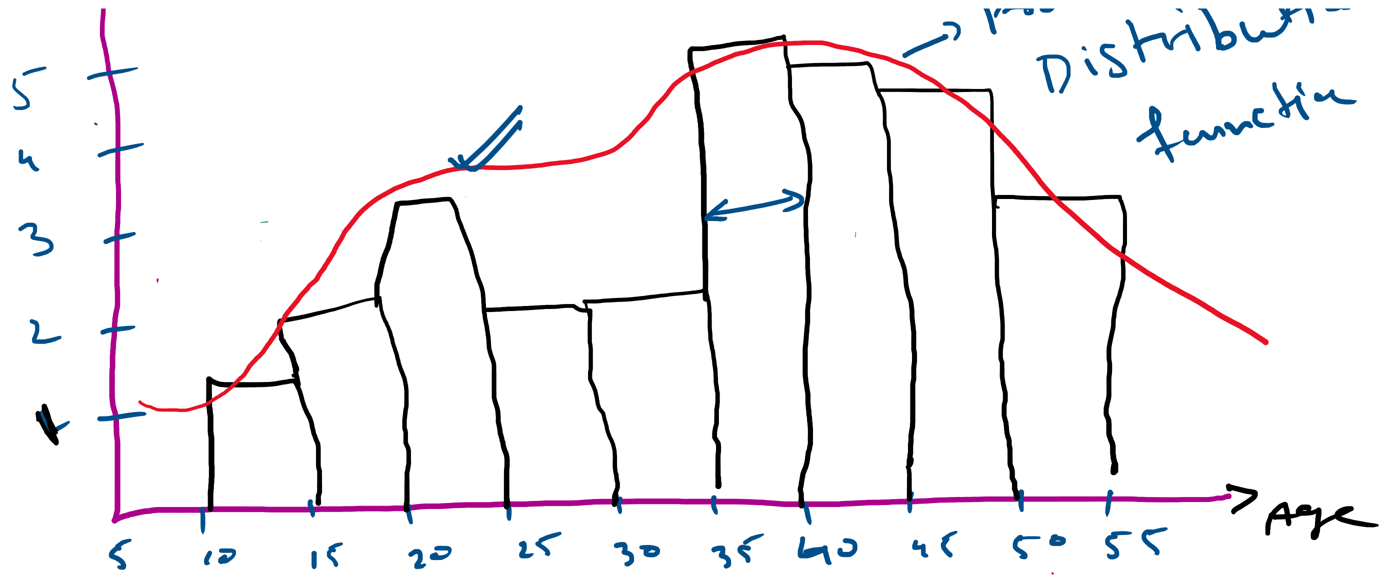
Also try the Standard Deviation Calculator.

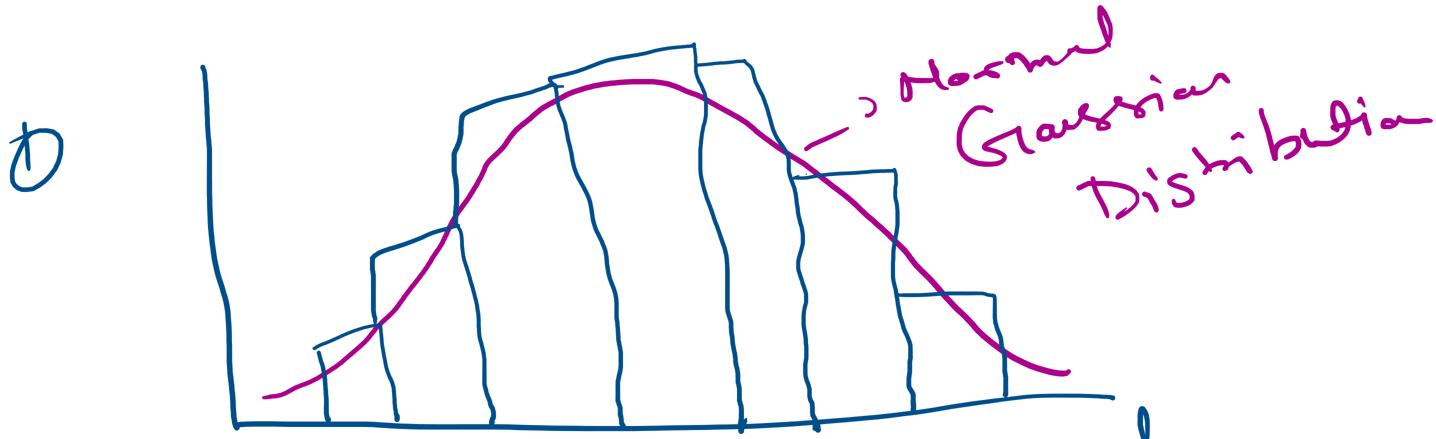
Histogram and Skewness

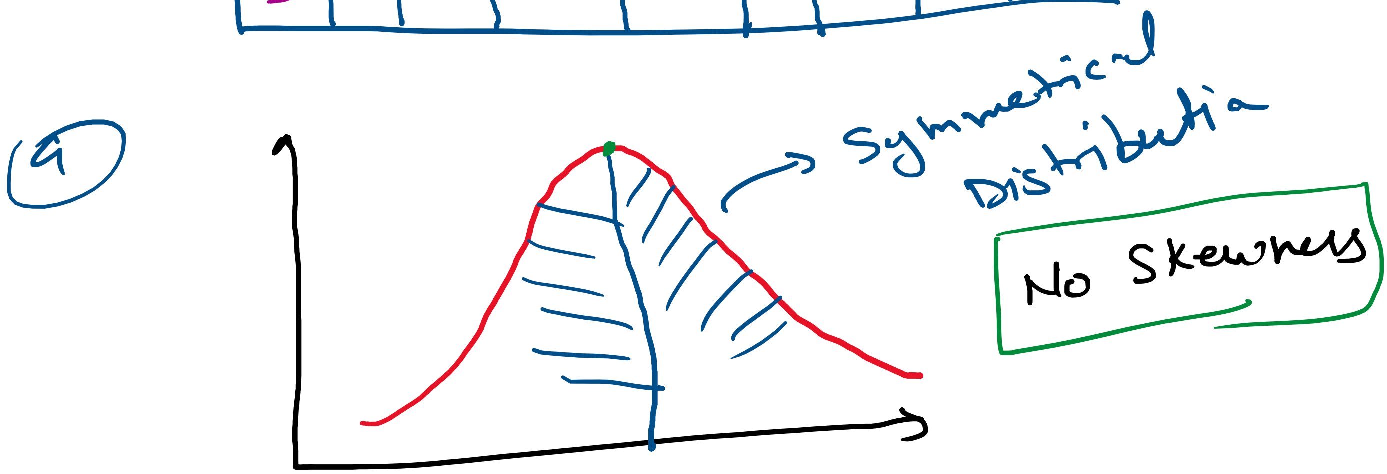
[Frequency]

age = { 10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 43, 50, 51 }









No Skewness

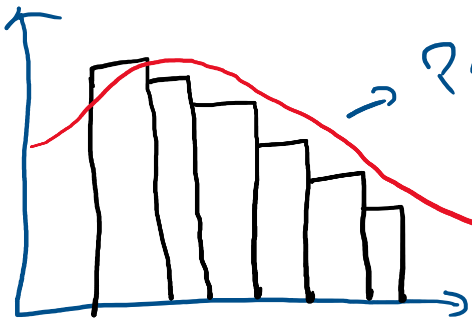
* The mean, median & mode all are Perfectly at the Centre.

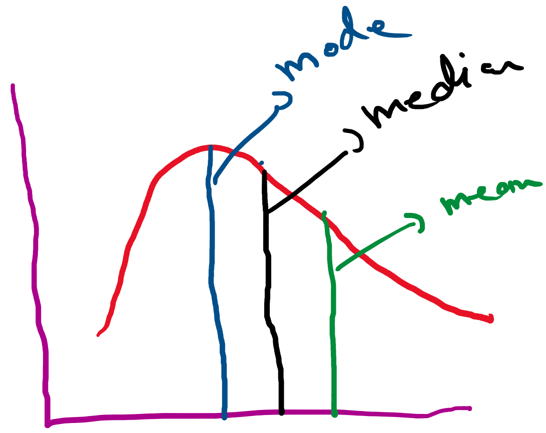
$\boxed{\text{mean} = \text{median} = \text{mode}}$

↳ No Skewness

↳ No Skewness

(b) Right Skewed

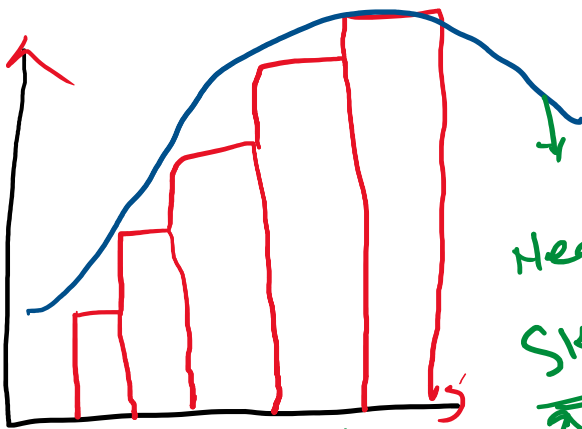




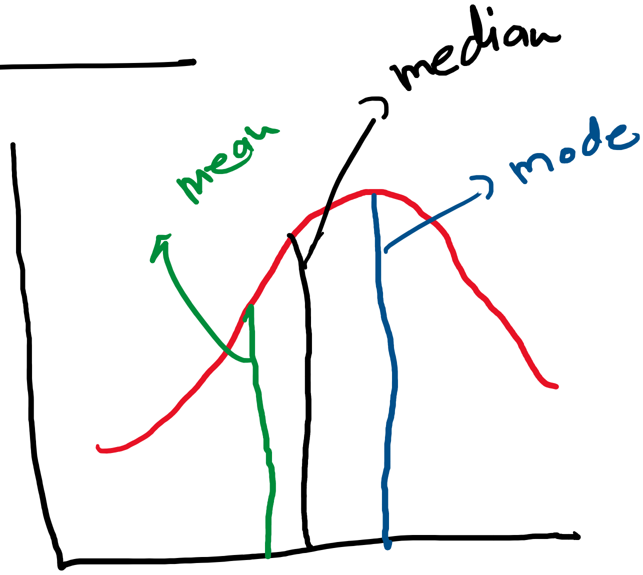
+ve skewed

$\boxed{\text{Mean} \geq \text{median} \geq \text{mode}}$

(c) Left Skewed Distributia



Negative Skewed



mean $\leq$ median $\leq$ mode