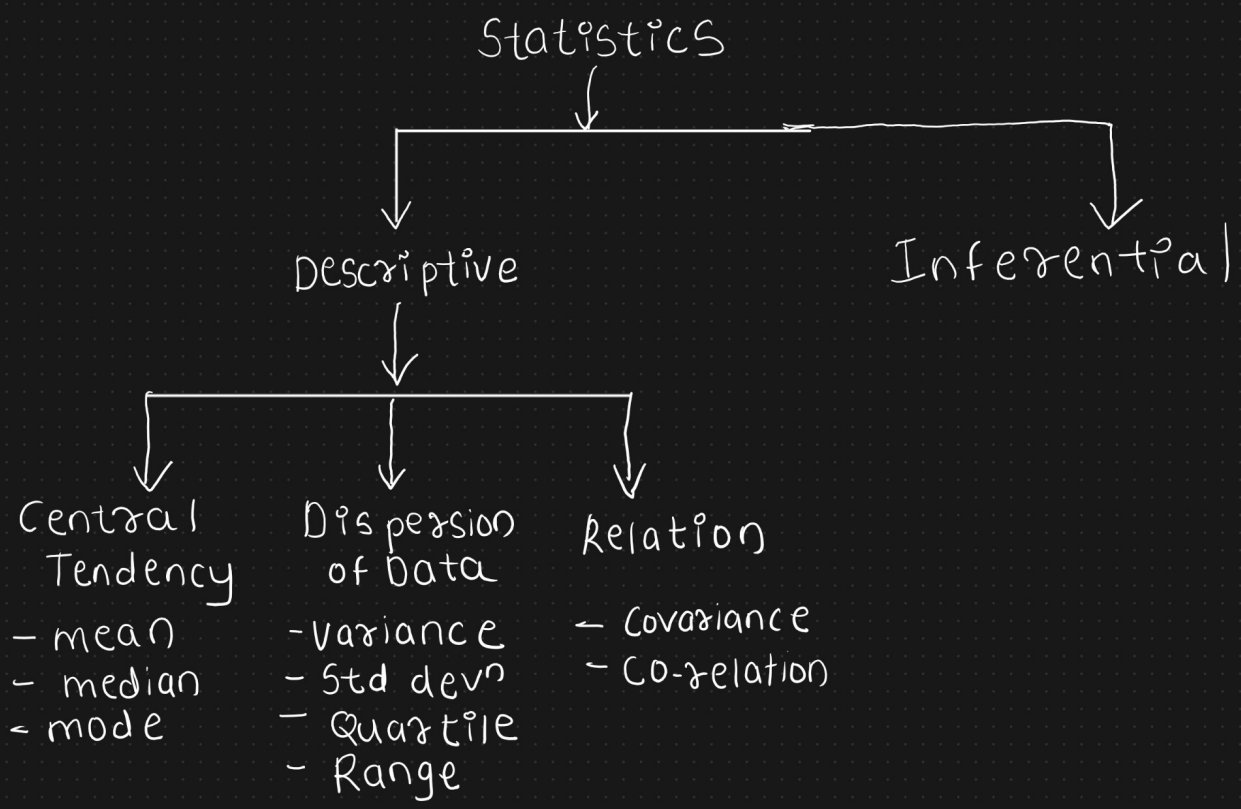
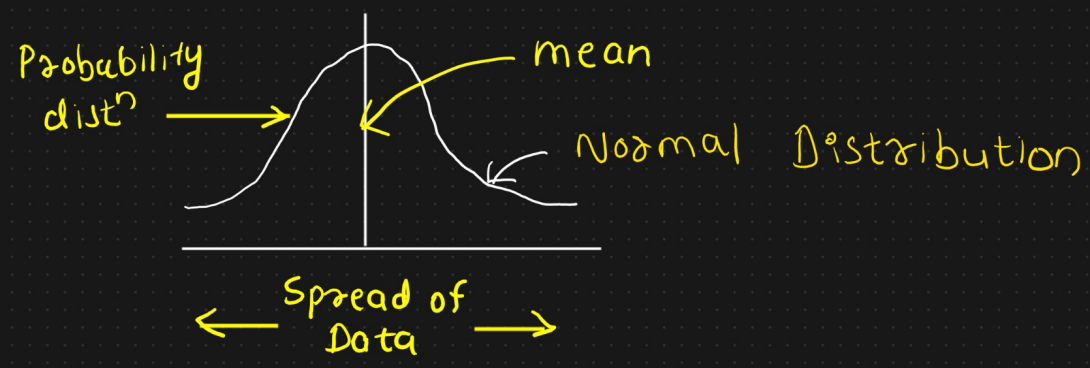
Statistics for Machine learning



* Central Tendency

(1) Mean :- Average or centremost value



② Median :-

Central value of population when population is sorted

for eg:- (i) when population size is odd

data = {1, 2, 2, 3, 4, 5, 6}

median

(ii) when population size is even

data = {1, 2, 2, 3, 4, 5, 6, 29} — outlier

$\frac{3+4}{2} = 3.5$ is mode

When we do not want our data to be influenced by outliers we use median

③ Mode

The most frequent value in a dataset is mode.

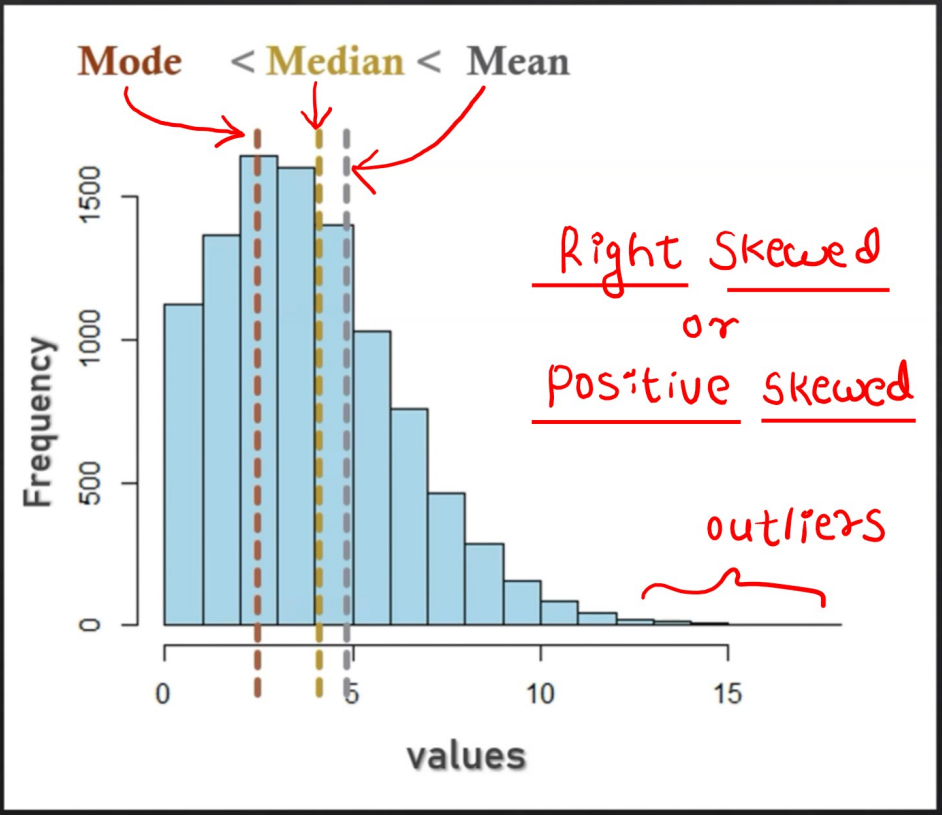
data = {1, 2, 2, 3, 4}

2 is mode

We use mode when dealing with categorical dataset to replace null values sometimes.

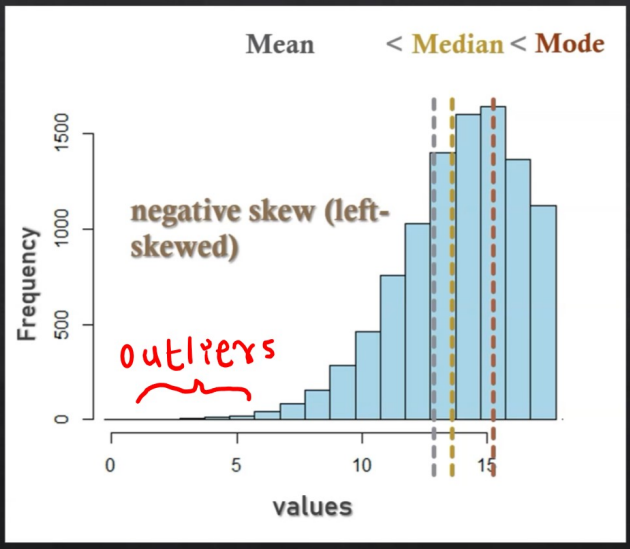
* Skewness

① Right Skewed



Here most outliers lie on right side

② Left skewed



Here outliers lie on the left side

* Calculation of skewness

Calculation of the Skewness:

$$\gamma_m = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\sigma^3}$$

mean

n = population size

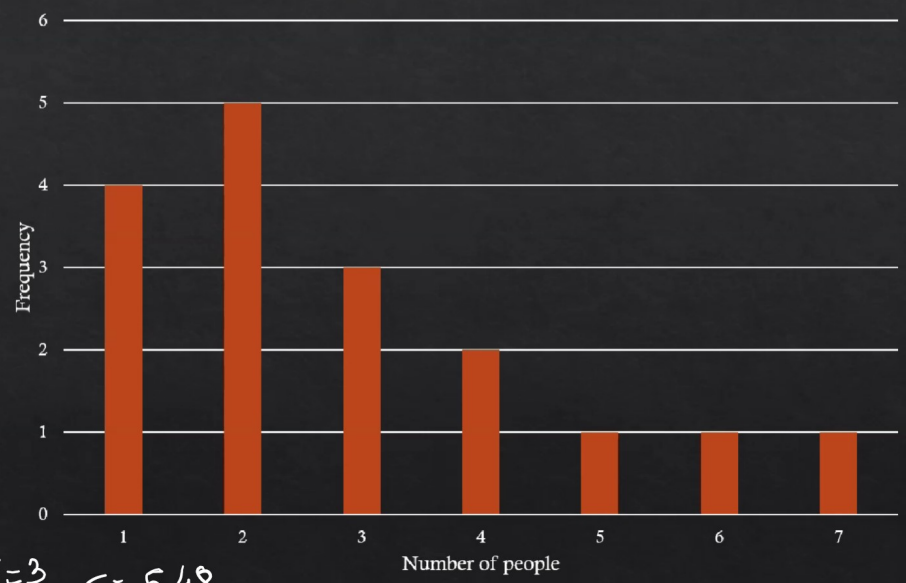
standard deviation

$\gamma_m > 0$ positive skew (right-skewed)

$\gamma_m < 0$ negative skew (left-skewed)

{ Note:- In python we can automatically calculate skewness. Remembering formula is not important }

Number of people
1
2
2
1
3
4
3
2
5
1
2
7
3
4
5
6
2
1



$\bar{x} = 3$, $\sigma = 5.48$

$$\gamma_m = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\sigma^3} = \frac{4}{5.48} = 0.73$$

$\gamma_m = 0.73$

Since γ_m is closer to 1, the data is right skewed.

* Measures of Spread

* Range and IQR {Inter-Quartile Range}

* Range

min value ↓ , ↓ max value

data = {1, 2, 3, 4, 5, 6}

Range = max val − min val

= 6 − 1

= 5

* Inter Quartile Range (IQR)

(i) For data with even count

{2, 5, 1, 3, 6, 4}

sort the data

{1, (2), 3 | 4, (5), 6}

IQR = 5 − 2 = 3

(ii) For data with odd count

{2, 5, 1, 1, 3, 6, 4, 6, 4}

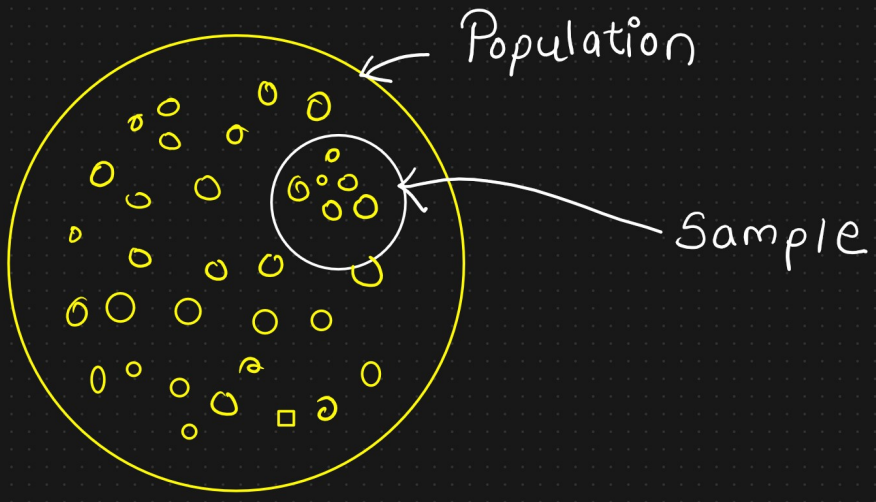
sort the data

{1, (1, 2), 3, 4, 4, (5, 6), 6}

$\frac{1+2}{2} = 1.5$ $\qquad$ $\frac{5+6}{2} = 5.5$

IQR = 5.5 − 1.5 = 4

* Sample and Population



Nomenclatures in Population

μ = Average of Population

N = Size of Population

σ^2 = Variance of population

Nomenclatures in Samples

$\bar{x}$ = sample average

n = Sample size

s^2 = Sample Variance

* Variance and Standard deviation

Two most sensitive measures of Data spread

① Variance

ⓐ Population Variance

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

Let's understand the formula

$(x_i - \mu)$ calculates the distance of each value from mean. This is called as mean centering.

Now some of these distances can be +ve or -ve. Therefore they are squared so that they can become +ve and they do not cancel each other.

Also squaring gives more weight to outliers (data points far away from mean)

ⓑ Sample Variance

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

$(n-1)$ is degree of freedom. It is taken because s^2 will give a proper estimate of σ^2.

In Conclusion, higher the variance the more the distance between each values. The lesser the variance more compact is data, i.e values are very close to each other.

② Standard Deviation (σ)

$$\sigma = \sqrt{\sigma^2}$$

Std dev or Variance is used depending on use cases.

* Scaling and Shifting

Scaling :- Multiply/Div values of data by certain value

for e.g:- {1,1,1,1} scaled by 2 becomes {2,2,2,2}

Shifting :- Add/Sub values o data b certain value.

for e.g:- {1,2,3,4} shifted by 2 becomes {2,4,5,6}

Let's check effects of scaling & shifting

{1, 2, 4, 5, 7, 5}

{3, 5, 7, 8, 10, 8} ← Shifting by 3

{2, 4, 8, 10, 14, 10} ← Scaling by 2

		Shifting by 3	Scaling by 2
mean	4	+3	×2
median	4.5	+3	× 2
mode	5	+3	× 2
range	6	No Change	× 2
IQR	3	No Change	× 2
Vasiance	4	No Change	$\times 2^2$

* Statistical Moments

Standardized statistical moments

Mean – Center

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Variance – Spread

$$m_2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

Skewness – Dispersion asymmetry

$$m_3 = \frac{1}{n\sigma^3}\sum_{i=1}^{n} (x_i - \bar{x})^3$$

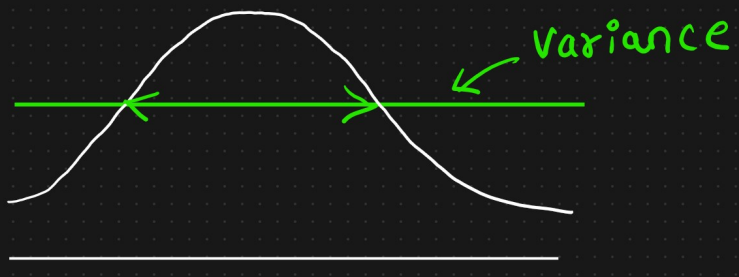
Kurtosis – Tail "heaviness"

$$m_4 = \frac{1}{n\sigma^4}\sum_{i=1}^{n} (x_i - \bar{x})^4$$

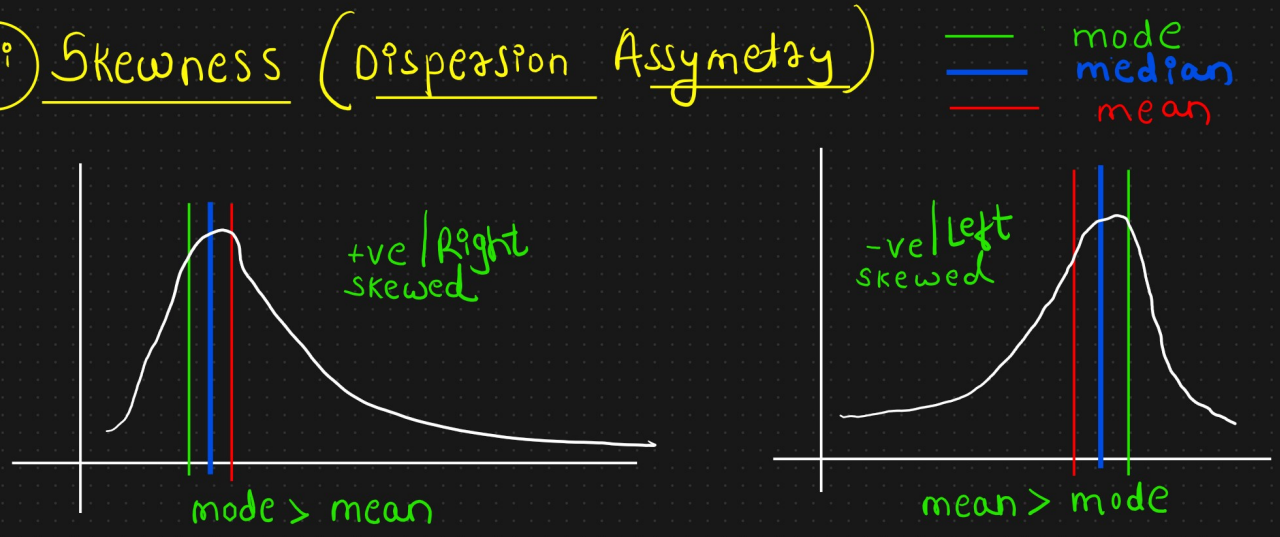
It gives us the idea of how data is spread

(i) Mean

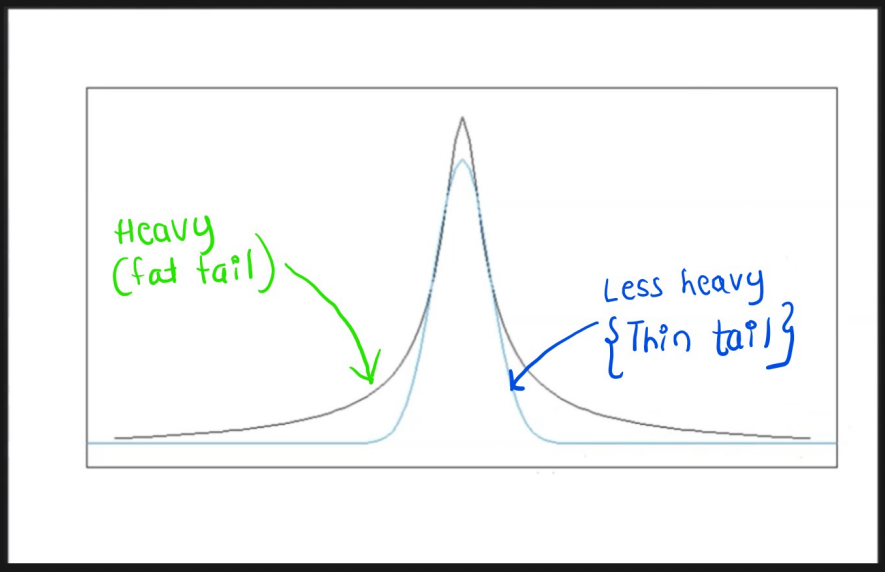
(ii) Variance



(iii) Skewness (Dispersion Assymetry)



(iv) Kurtosis (Tail "Heaviness")

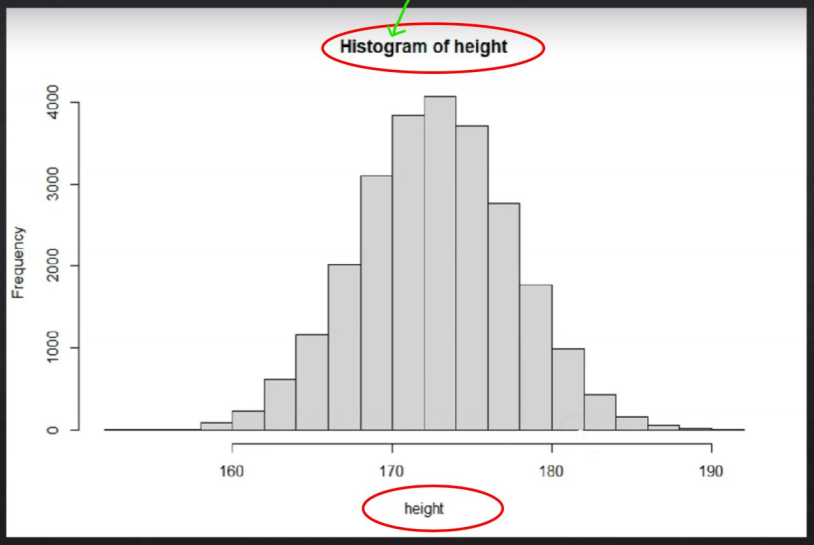


* Distributions

Distribution tells us how often values appear in a Dataset.

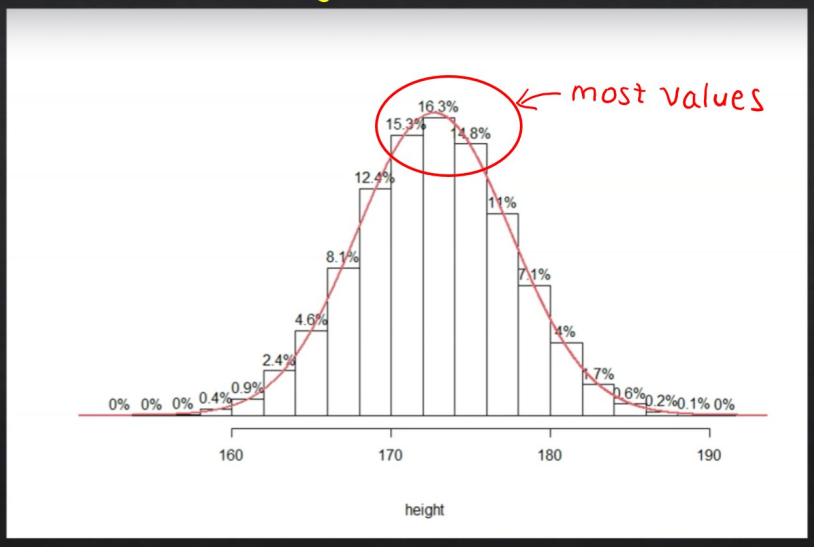
it is a frequency plot

Height in cm	Weight in kg	category
167.0895172	51.25259682	160-170
181.6485353	61.90967151	180-190
176.2727044	69.41191679	180-190
173.2700704	64.5623282	170-180
172.1809444	65.45214141	170-180
174.4924194	55.92909436	170-180
177.2970859	64.18099264	180-190
177.8372928	61.89833169	180-190
172.4726379	50.9712785	170-180
169.6271028	54.73378632	170-180
168.8786414	57.81114206	170-180
171.7631654	51.77445546	170-180
173.4882055	56.97612287	170-180
170.4759703	55.54780416	170-180
173.4302681	52.65605864	170-180
180.5725665	63.50187334	180-190
168.8108488	58.74132504	170-180
174.3690516	64.85167512	170-180
180.9249405	62.55159619	180-190



If we display them in percentage

Height in cm	Weight in kg	category
167.0895172	51.25259682	160-170
181.6485353	61.90967151	180-190
176.2727044	69.41191679	180-190
173.2700704	64.5623282	170-180
172.1809444	65.45214141	170-180
174.4924194	55.92909436	170-180
177.2970859	64.18099264	180-190
177.8372928	61.89833169	180-190
172.4726379	50.9712785	170-180
169.6271028	54.73378632	170-180
168.8786414	57.81114206	170-180
171.7631654	51.77445546	170-180
173.4882055	56.97612287	170-180
170.4759703	55.54780416	170-180
173.4302681	52.65605864	170-180
180.5725665	63.50187334	180-190
168.8108488	58.74132504	170-180
174.3690516	64.85167512	170-180
180.9249405	62.55159619	180-190

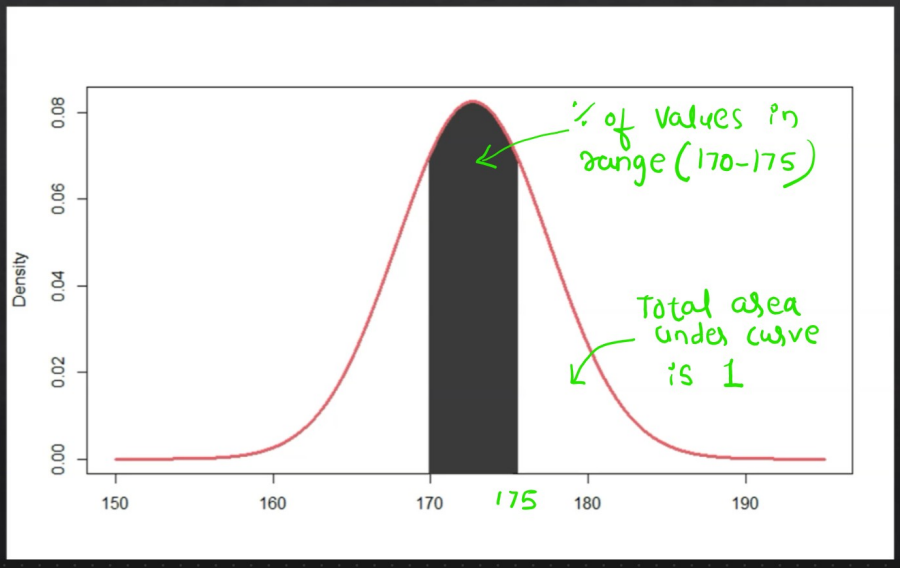


This will help us to find certain values which fall in a particular range.

If we plot a distribution chart using density function, we can calculate % values in certain range by calculating area covered by the values in the distribution.

{ what density function is? is discussed further }

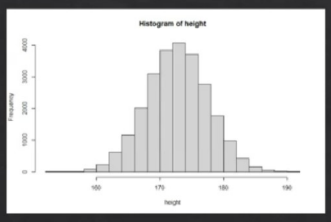
Height in cm	Weight in kg	category
167.0895172	51.25259682	160-170
181.6485353	61.90967151	180-190
176.2727044	69.41191679	180-190
173.2700704	64.5623282	170-180
172.1809444	65.45214141	170-180
174.4924194	55.92909436	170-180
177.2970859	64.18099264	180-190
177.8372928	61.89833169	180-190
172.4726379	50.9712785	170-180
169.6271028	54.73378632	170-180
168.8786414	57.81114206	170-180
171.7631654	51.77445546	170-180
173.4882055	56.97612287	170-180
170.4759703	55.54780416	170-180
173.4302681	52.65605864	170-180
180.5725665	63.50187334	180-190
168.8108488	58.74132504	170-180
174.3690516	64.85167512	170-180
180.9249405	62.55159619	180-190



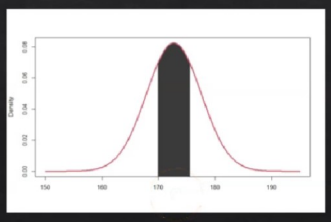
This above graph can be used to calculate probability of a certain person falling in a certain category.
for eg:- If a person enters our room what is the probability that its height is bet^n certain range.

What to remember

- What is a histogram: Frequency plot

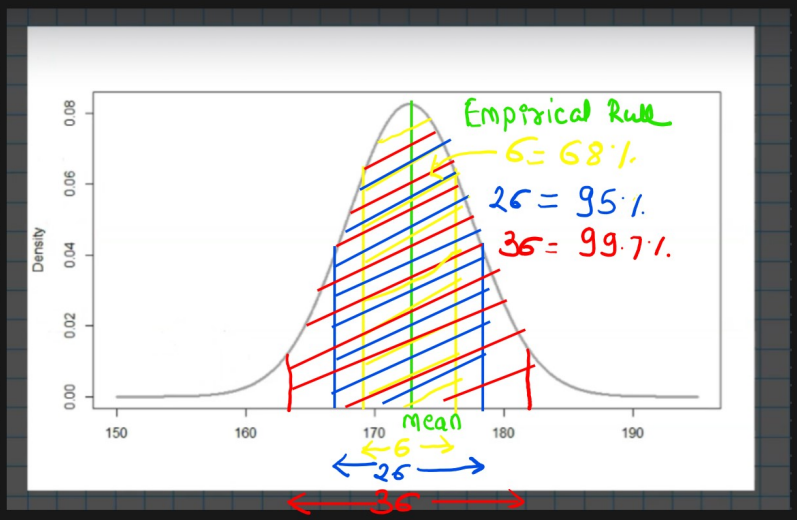


- Values in our data are always distributed in a certain way
- A distribution has a density function



- We get the probability that a random data point takes a value from a certain range if we calculate the are under the according density function

* Normal Distribution



Also called as
'Gaussian Distribution'
or
'Bell Curve'

Normal Distribution depends only on
ⓐ mean
ⓑ Standard deviation

* Empesical Rule:-

It states about what percentage of values lie when we shift σ, 2σ, 3σ towards left or right from the mean. under a normal distribution curve.

* Z-Scores

For the distribution on right →

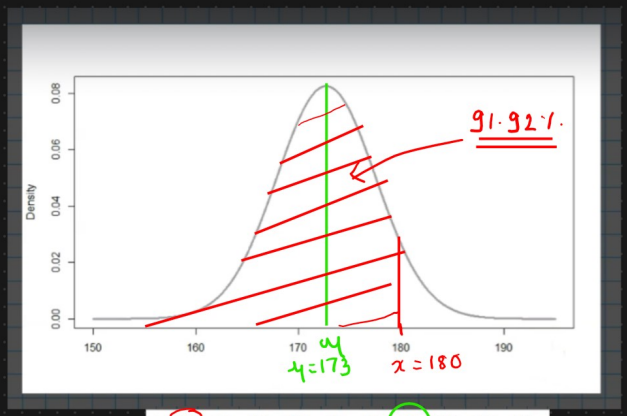
$x = 180$ $\quad \sigma = 5$

$\mu = 173$

$$Z_{score} = \frac{x - \mu}{\sigma}$$

$$z = \frac{180 - 173}{5}$$

$$z = 1.40$$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Z-Table

from Z-table area under the curve till $x = 180$ is 0.9192 i.e 91.92 %

This calc^n implies that around 91.92% values are smaller than 180.

In second case, suppose we want to know which value is in top 4% ie we need to know value of x if area is 96%.

Then locate value in z table to 0.96 or slightly greater. {0.9608}

from table we get $z_{score} = 1.76$

$$z = \frac{x - \mu}{\sigma} \quad \therefore x = (z \times \sigma) - \mu$$

$$\therefore x = (1.76 \times 5) - 173$$

$$\therefore x = 181.8$$

$\therefore$ All values starting from 181.8 are in top 4%.

There's another z-table for -ve values of z ie for x values on left of mean.

Later on we will work a lot more with the normal distribution and also learn about more distributions. But before we do that let's learn about probability theory!