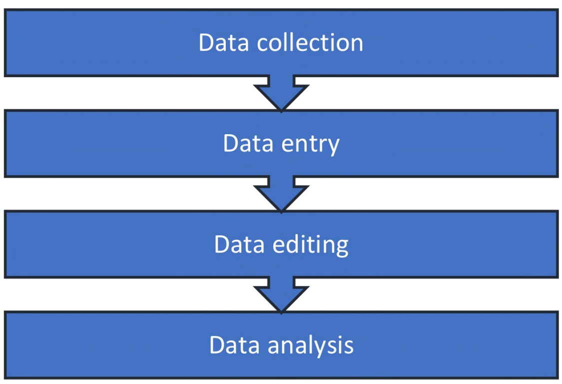
* Full Summary :-

Chapter (1) :- General over view :-

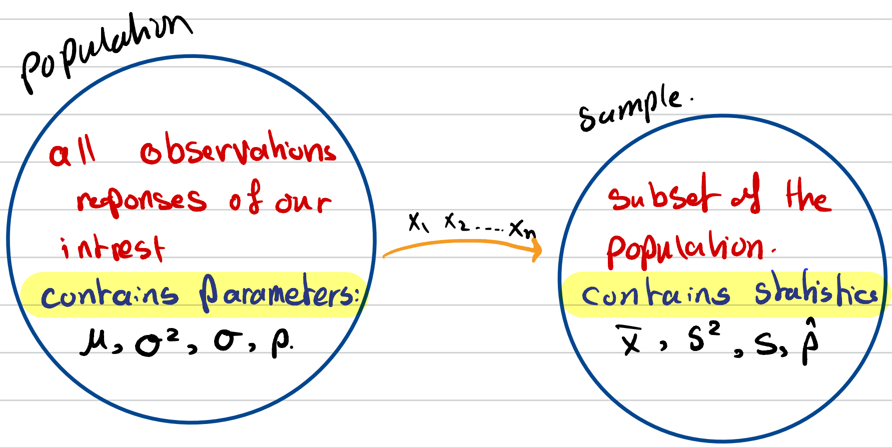
Statistics

- Mathematical.
- Applied.
  - ⇓ Bio statistic.

How the research process go ?



* Data sets :-



Data.

Discrete — "countable"

continuous. — "Measurable"

Statistics.

Descriptive — Describing, classifying summerizing & presenting the data.

Inferential. — making conclusions or inferences about the pop. based on sample.

Chapter (2):- Descriptive statistics.

① Measures of central tendency.

② Measures of variation.

→ Measures of central tendency:-

① Mean ($\bar{x}$)

- affected by outliers
- found by:

$$\bar{X} = \frac{\sum x_i}{n}.$$

② Mode (m).

- the most frequent value.
- used for qualitative data.

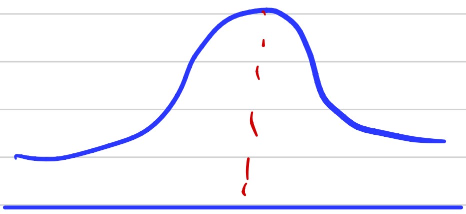
one mode → unimodal.
two modes → bimodal.
three modes → multimodal.

③ Median (Q_2).

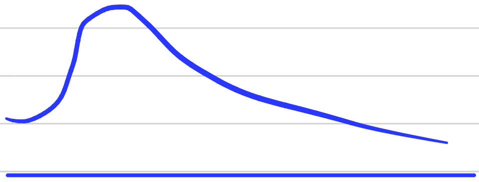
- the value in the middle & not affected by outliers.
- $Q_2 \to \frac{n+1}{2}$ { fraction → average. / whole no. → keep it.

* Comparison the mean & the median:-

① symmetric. ② skewed to right. (positive)

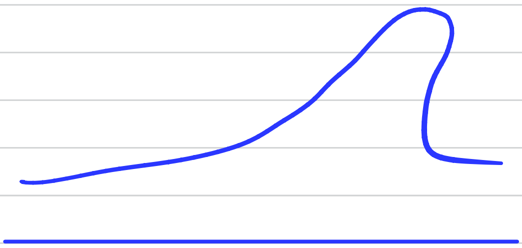


$\bar{x} = Q_2$



$\bar{x} > Q_2$.

③ skewed to left. Negative.



$\bar{x} < Q_2$.

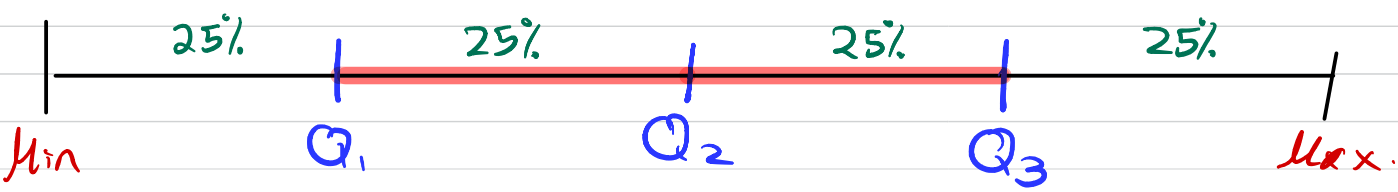
* Measures of spread:- (non-negative values)

① Range ② IQR ③ variance ④ std.

⇒ Range = Max − Min.

easy to calculate but very sensitive to outliers.

⇒ Inter-quartile-range = $Q_3 - Q_1$



takes the middle 50%, not affected by outliers.

$Q_1 \rightarrow \frac{25}{100} \cdot n = \frac{n}{4}$.

$Q_3 \rightarrow \frac{75}{100} \cdot n = \frac{3n}{4}$.

- fraction → next int.
- whole no. → $\frac{K^{th} + (K+1)^{th}}{2}$

Don't forget to order data values.

* **Percentiles:** The value that has K% of data below it

$P_K \rightarrow \frac{K}{100} \cdot n$

- fraction → next int.
- whole no. → $\frac{K^{th} + (K+1)^{th}}{2}$

⇒ **Variance & std:**

Deviation $D_i = x_i - \bar{x}$.

Sum of Deviations $= 0$ for any data set.

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

$$S^2 = \frac{\sum x^2}{n-1} - \frac{(\sum x)^2}{n(n-1)}$$ (better)

$std = \sqrt{variance}$.

affected by outliers.

* Outliers :-

Any value less than $Q_1 - 1.5IQR$
More than $Q_3 + 1.5IQR$
is called outlier

Any value less than $Q_1 - 3IQR$.
More than $Q_3 + 3IQR$.
is called extreme outlier.

* coefficient of variation :

$$C.V = \frac{S}{\bar{x}} \cdot 100\%$$

used to compare variation in two data sets of different units, since It's unitless.

* Coding (linear transform).

$$y = ax + b.$$

* Measures of central tendency

$\bar{y} = a \cdot \bar{x} + b$

$Q_2(y) = a \cdot Q_2(x) + b$.

$mode(y) = a \cdot mode(x) + b$.

Affected by addition & multiplication.

* Measures of spread.

$Range(y) = |a| \cdot Range(x)$.

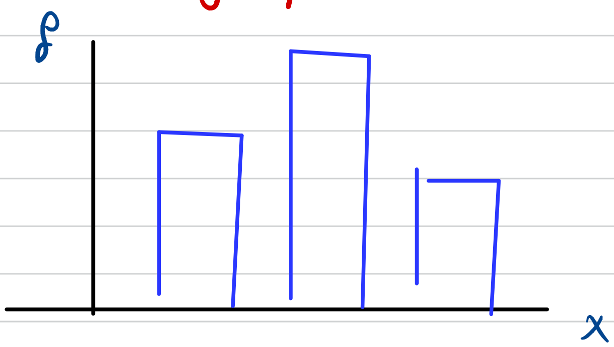
$S_y = |a| \cdot S_x$.

$S^2_y = a^2 \cdot S^2_x$.

affected by multiplication only.

* Graphic Methods :-

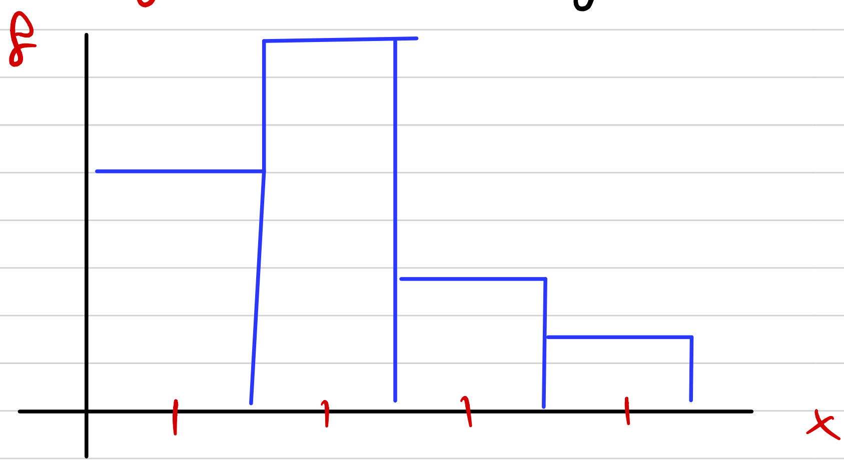
① Bar-graph :-



→ used for discrete data only.

→ good for qualitative (non-numeric data).

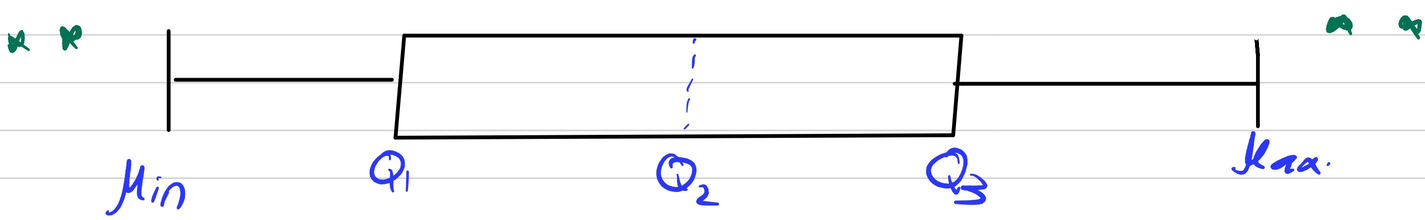
② Histogram :- (using Minitab only).



→ we can know the shape of data distribution.

→ for quanitative data only.

③ Box & whiskers plot :-

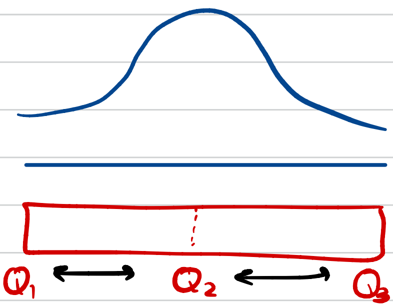


Represent the five number summary:

Min, Q_1, Q_2, Q_3, Max.

* skewness using box plot:

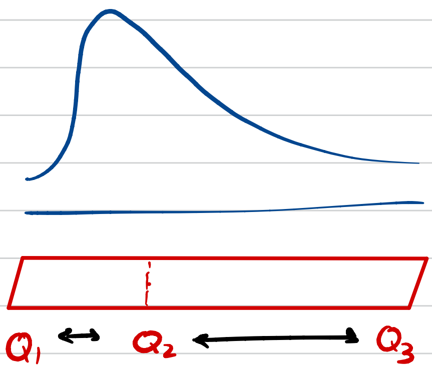
① symmetric.



$Q_3 - Q_2 = Q_2 - Q_1$

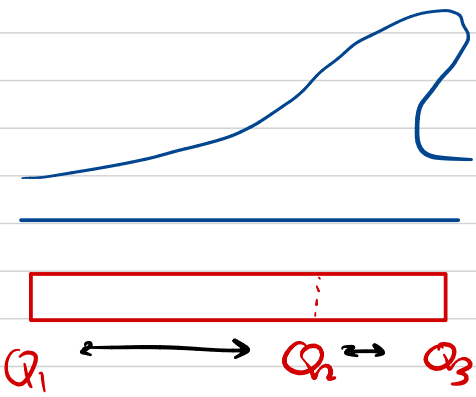
equal distances.

② skewed to right:



$Q_3 - Q_2 > Q_2 - Q_1$

③ skewed to left.



$Q_3 - Q_2 < Q_2 - Q_1$

* chapter (3) : probability

⇒ Some basic definitions:

Sample space : the set of possible outcomes. (S) or (Ω).

N(S) or N(Ω): number of elements in S.

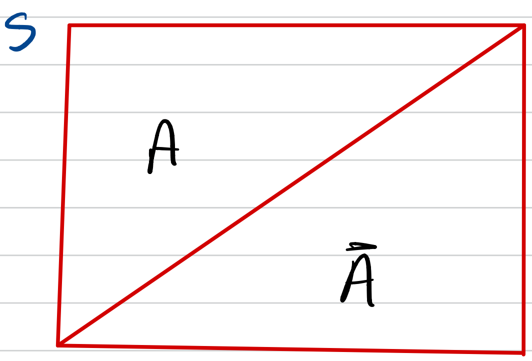
outcome: the result of a single trial.

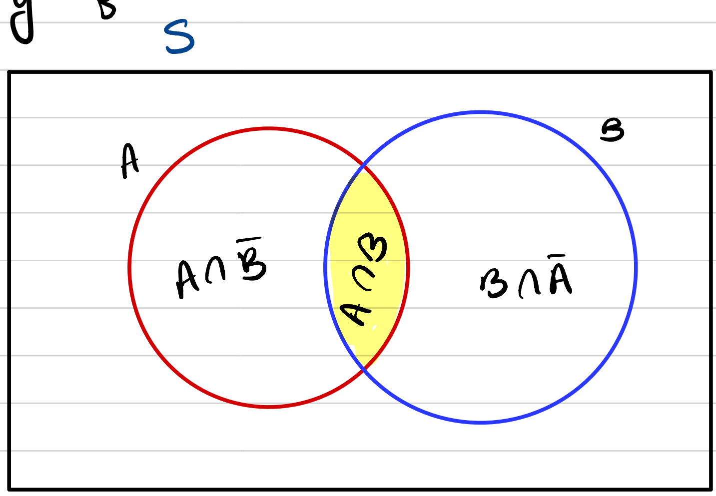
Event: any subset of the sample space. A, B, C ...

probability: The chance of getting an event.

$$P(A) = \frac{N(A)}{N(S)} = \frac{\text{no. of elements in } A}{\text{no. of elements in } S}$$

* Rules of probability :





① $P(A) + P(\bar{A}) = 1.$

→ $P(A) = 1 - P(\bar{A})$

→ $P(\bar{A}) = 1 - P(A).$

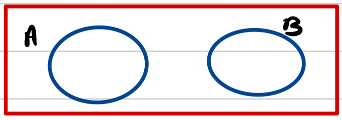
② $P(A \cup B) = P(A) + P(B) - P(A \cap B).$

③ $P(A \cap \bar{B}) = P(A) - P(A \cap B).$

$P(B \cap \bar{A}) = P(B) - P(A \cap B).$

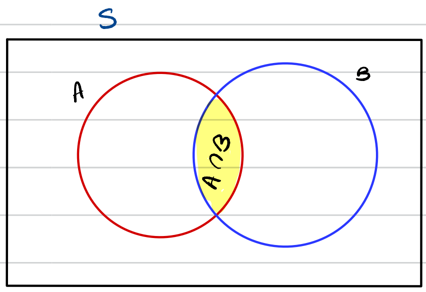
④ If A & B mutually Exclusive (disjoint) :-

$P(A \cap B) = 0.$



⑤ If A & B independent :-

$P(A \cap B) = P(A) \times P(B).$



Also: $P(\bar{A} \cap \bar{B}) = P(\bar{A}) \times P(\bar{B})$

$P(\bar{A} \cap B) = P(\bar{A}) \times P(B)$

$P(A \cap \bar{B}) = P(A) \times P(\bar{B}).$

⑥ conditional probability :-

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}.$$

$A \cup B = B \cup A$

$A \cap B = B \cap A.$

order doesn't matter.

Note: If A & B independent :-

→ $P(A \mid B) = P(A)$

→ $P(B \mid A) = P(B).$

Note: Relative Risk of B given A.

$$\frac{P(B \mid A)}{P(B \mid \bar{A})}$$

→ $= 1$ independent.

→ > 1 dependent.

Note: $P(\bar{A} \cap \bar{B}) = \overline{P(A \cup B)} = 1 - P(A \cup B)$

$P(\bar{A} \cup \bar{B}) = \overline{P(A \cap B)} = 1 - P(A \cap B).$

* probability table :-

→ good when we need to find intersections
$P(A \cap B)$, $P(A \cap \bar{B})$ ....

	A	$\bar{A}$	total
B	$P(A \cap B)$	$P(\bar{A} \cap B)$	$P(B)$
$\bar{B}$	$P(\bar{B} \cap A)$	$P(\bar{A} \cap \bar{B})$	$P(\bar{B})$
total	$P(A)$	$P(\bar{A})$	1

* Screening tests :-

Actual / Disease.

Test / symptome

	Positive (D)	Negative ($\bar{D}$)
Positive (T^+)	True Positive (TP).	False Positive (FP)
Negative (T^-)	False Negative (FN)	True Negative (TN).

→ $P(PV+) = P(D \mid T^+) = \frac{P(D \cap T^+)}{P(T^+)} = \frac{TP}{TP+FP}$

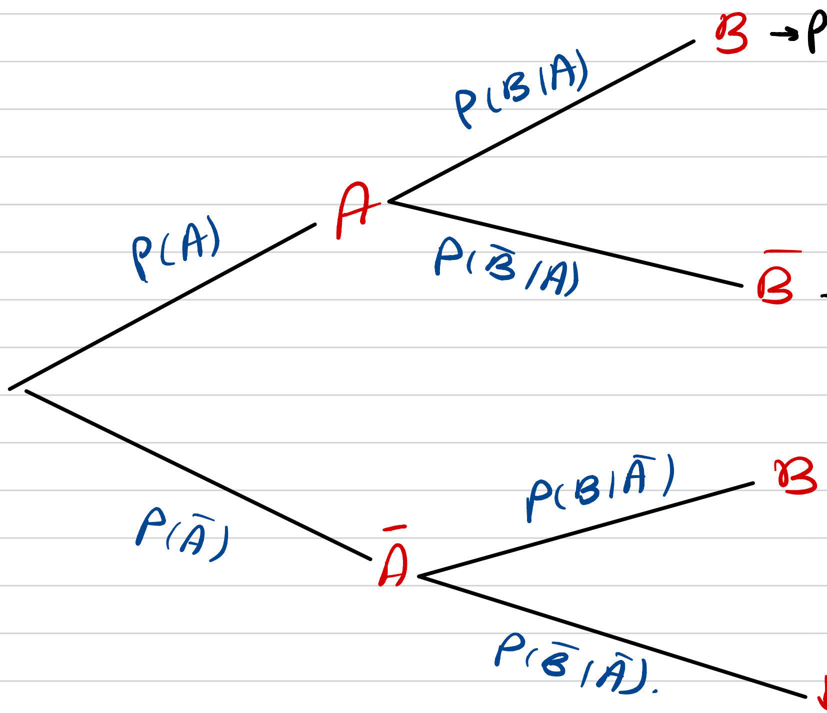
→ $P(PV-) = P(\bar{D} \mid T^-) = \frac{P(\bar{D} \cap T^-)}{P(T^-)} = \frac{TN}{TN+FN}$

→ sensitivity $= P(T^+ \mid D) = \frac{P(T^+ \cap D)}{P(D)} = \frac{TP}{TP+FN}$.

→ specificity $= P(T^- \mid \bar{D}) = \frac{P(T^- \cap \bar{D})}{P(\bar{D})} = \frac{TN}{TN+FP}$.

* Tree diagram (Bayes theorem):

- → used when we have conditional probabilities
- → used when we have two partitions, two stages



$B \rightarrow P(A \cap B) = P(A) \times P(B|A).$

$\bar{B} \rightarrow P(A \cap \bar{B}) = P(A) \times P(\bar{B}|A)$

$B \rightarrow P(\bar{A} \cap B) = P(\bar{A})\, P(B|\bar{A})$

$\bar{B} \rightarrow P(\bar{A} \cap \bar{B}) = P(\bar{A})\, P(\bar{B}|\bar{A}).$

$$P(B) = P(A \cap B) + P(\bar{A} \cap B)$$

↳ This is called total probability rule.

✱ Chapter (4) :- Discrete probability distribution

Random Variable :- A function that assigns numiric values to different events in S.

Random variable.

- **Discrete**: countable, finite.
- **Continuous**: Measurable, infinite.

✱ probability distribution :

A table has the values of x & their prob.

x	x_1	x_2	x_3	....
$P(x)$	-	-	-	-

$\sum P(x) = 1.$

✱ probability mass function (P.M.F) :-

A function $f(x) = P(X = x_i)$ is a P.M.F if :-

① $P(X = x_i) \geq 0$ ② $\sum P(X = x_i) = 1.$

✱ The Expected value :- It's the weighted mean.

$$E(x) = \mu = \sum x \cdot P(x = x_i).$$

✱ The Variance :-

→ $Var(x) = \sigma^2 = E(x - \mu)^2.$

→ $Var(x) = \sigma^2 = E(x^2) - (E(x))^2$ (way better)

$Std = \sigma = \sqrt{Var(x)}.$

Note : $E(x^2) = \sum x^2 \cdot P(x = x_i).$

* The Binomial distribution :-

x is said to follow binomial if :-

1. n independent trials.
2. we have two outcomes.
3. prob. of success p is the same.

$\Rightarrow$ $X \sim Bin(n, p)$. $\quad x = 0, 1, \dots n$.

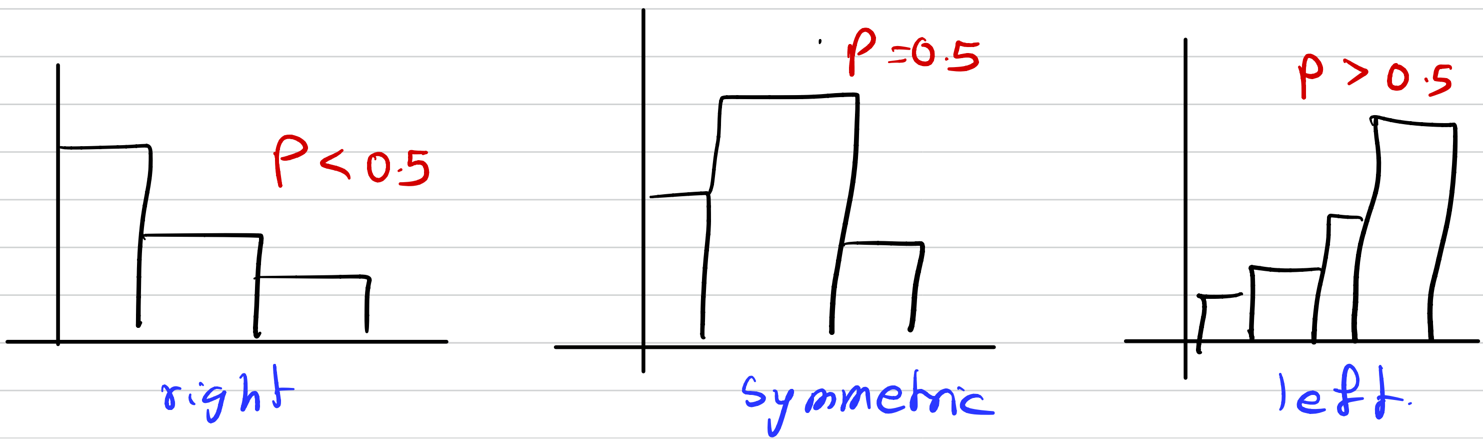
n : number of trials.

p : prob. of success.

q : prob. of failure. $\quad q = 1 - p$

$\Rightarrow$ we want to find :-

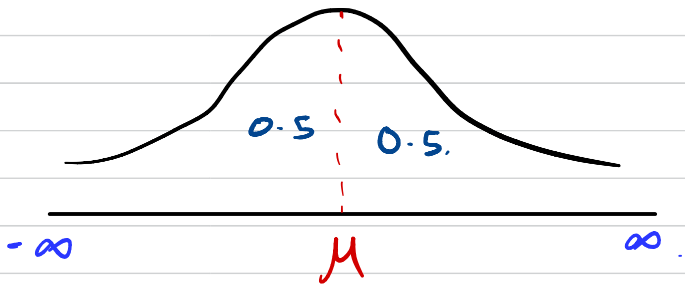
1. $P(X = x_i) = \binom{n}{x_i} p^{x_i} \cdot q^{n - x_i}$
2. $P(x \leq k)$ or $P(x \geq k)$ --- use tables.
3. $E(x) = n \times p$.
4. $var(x) = n \times p \times q$.
5. $Std = \sqrt{var(x)}$.



* Chapter (5) : Normal distribution.

$X \sim N(\mu, \sigma^2)$.

1) symmetric.
2) mean = mode = median.
3) prob. = area.
4) total area = 1.



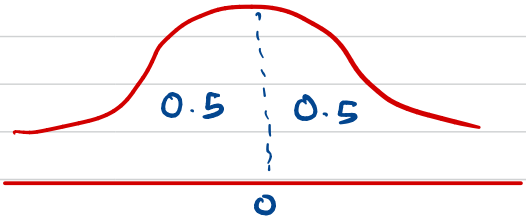
The standard normal :-

$Z \sim N(0, 1)$.

1) $P(Z \leq K)$ use tables.
2) $P(Z > K) = 1 - P(Z \leq K)$

or $P(Z > K) = P(Z < -K)$

3) $P(Z = K) = 0$.
4) $P(a < Z < b) = P(Z < b) - P(Z < a)$.



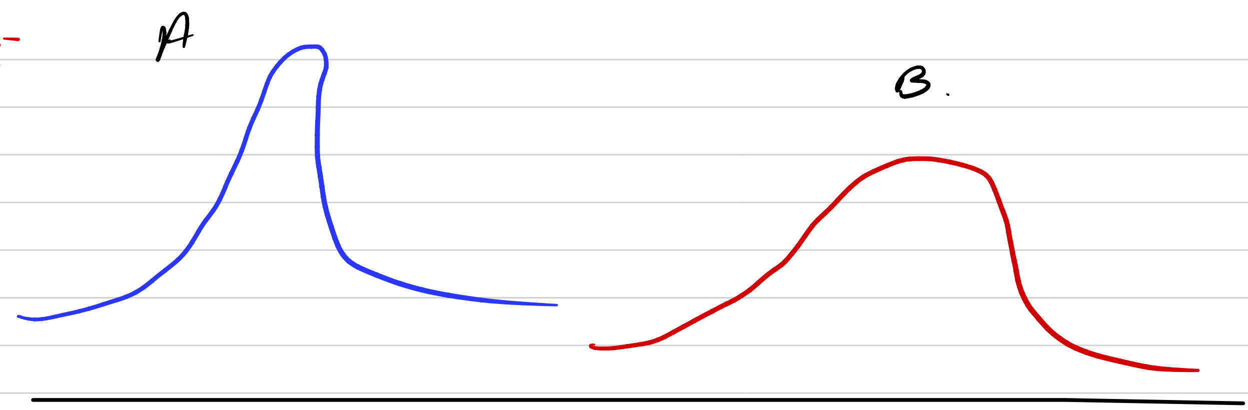
Note : $P(-a < Z < a)$ using D column in tables.

* Transforming to standard normal :-

If $X \sim N(\mu, \sigma^2)$ then.

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

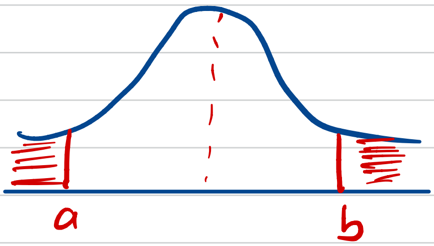
Note :-



var (B) heigher than var (A).

✱ very important notes

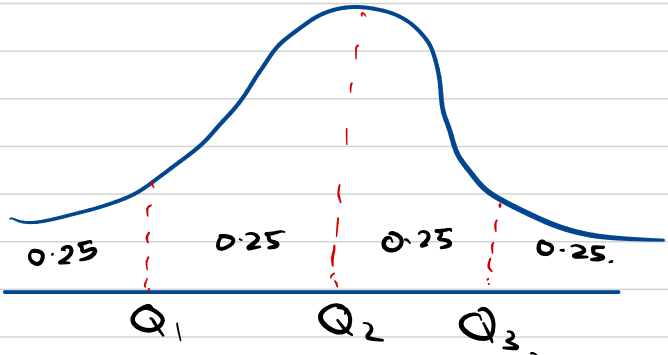
①



$P(x < a) = P(x > b)$

$\therefore \mu = \frac{a+b}{2}$.

②



$P(x < Q_1) = 0.25$

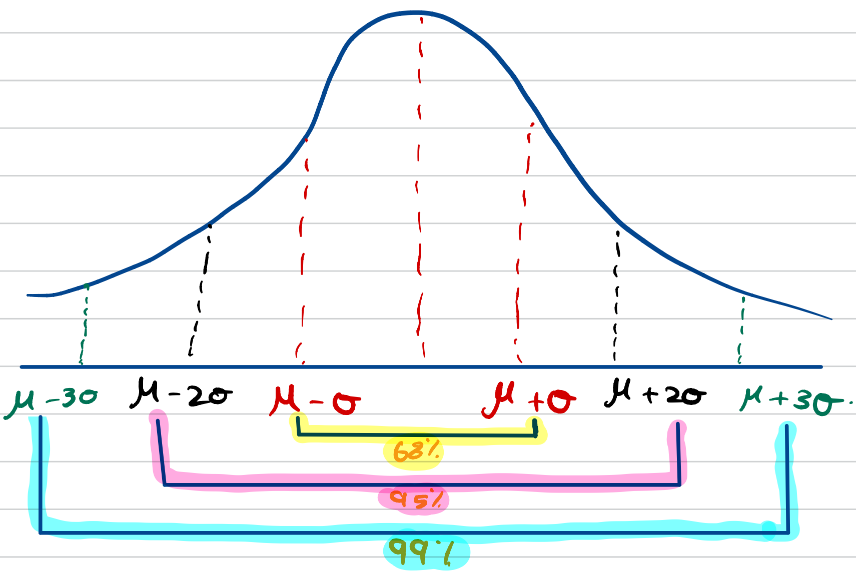
$P(x < Q_2) = 0.50$

$P(x < Q_3) = 0.75$.

In general :-

$P(x < P_K) = K\%$

③



Arwa M. Bader.