FULL NOTES ⇒ STATISTICS

- PROBABILITY

5.1 - PRESENTATION OF DATA

➔ **DISCRETE DATA** ⇒ Exact number valves. Can be counted.

EG - 1 ⇒ Number of peaple, test scores, shoe sizes, ...

➔ **CONTINUOUS DATA** ⇒ Takes numerical valves within a range, usually measured,

EG - 1 ⇒ Height, weight, temparature, ...

➔ **PRESENTATION METHODS** ⇒

➔ **TABLES** ⇒ We start with a list :

EG - 1 ⇒ Test scores, out of 10 : (n = 30)

8 7 8 4 1 9 9 10 6 8 6 7 8 2 10
10 9 8 6 8 2 8 7 3 7 9 5 9 6 8

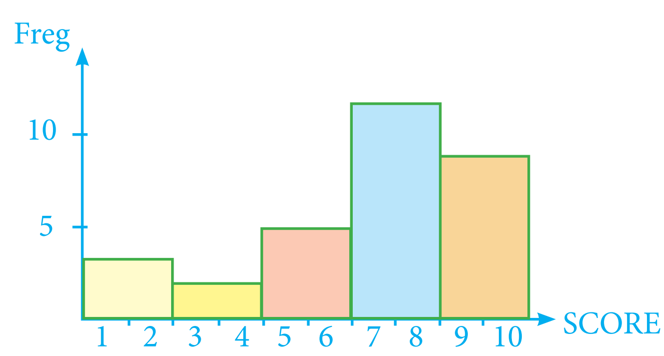
SCORE	TALLY	FREQUENCY
1	I	1
2	II	2
3	I	1
4	I	1
5	I	1
6	IIII	4
7	IIII	4
8	IIII III	8
9	IIII	5
10	III	3
TOTAL :		30

This data can be shown with a frequency table (night), and a groyed frequency table w / equal class size :

SCORE	TALLY	FREQ	RELATIVE FREQ %
1 - 2	III	3	10 %
3 - 4	II	2	6.7 %
5 - 6	~~IIII~~	512	16.7 %
7 - 8	~~IIII~~ ~~IIII~~ II	8	40 %
9 - 10	~~IIII~~ III		26.7 %
TOTAL :		30	100 %

The next step here would be to display the data in a HISTOGRAM

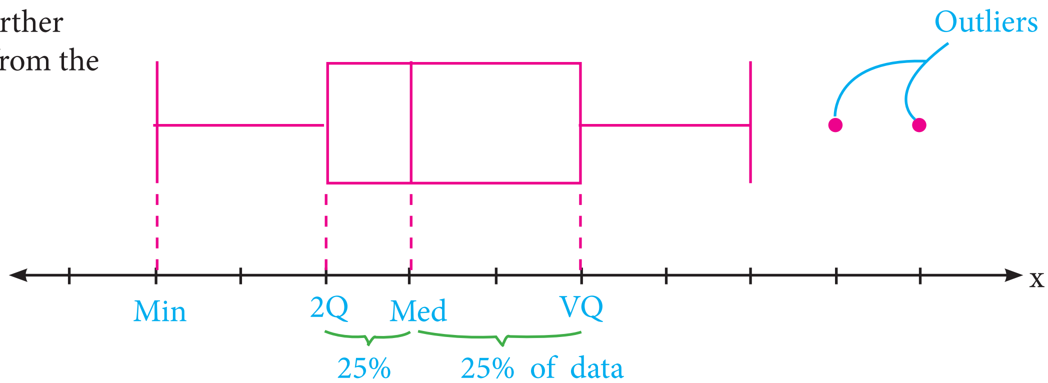
↳ We can say here that the 'modal class' is a score of 7 to 8.



➔ **BOX & WHISKER PLOTS** ⇒ This is another way of showing your data. It shows :

the medin, upper / lower quartiles (see 5.2) and 'outliers'

Qutliers : These are data points that lie further chan 1.5 x IQR from the median.



5.2 STATISTICAL MEASURES

➔ **CENTRAL TENDANCY :**

MODE ⇒ Most frequent valve / class :

MEDIAN ⇒ Order data, then obsenre the $\left(\frac{n+1}{2}\right)^{th}$ valve :

EG - ⇒

For : 1, 2, 1, 6, 5

1

2

$\frac{15}{5} = 3$

For : 18, 17, 14, 15
15, 11, 10, 4

15

Halfway between
14 & 15 = 14.5

$\frac{104}{8} = 13$

MEAN ⇒ $\frac{\text{TOTAL}}{\text{FREQVENCY}} = \frac{\sum x}{\sum f}$

↳ **For grouped frequency** ⇒ You can only 'estimete' the total by multiplying midpoints by Freg. and summing these

↳ $\frac{\sum (f.x)}{\sum f}$

EG - 1 ⇒

Add column

SCORE	TALLY	FREQUENCY
1 - 2	3	4.5
3 - 4	2	7
5 - 6	5	27.5
7 - 8	12	90
9 - 10	8	76
TOTAL :	30	205

Modal class

MEAN : $\frac{205}{30} = 6.83$

→ **QUARTILES** ⇒ Similar to how the median splits the data into 50% part, guartiles are 4 egual ports of 25%, after ordering

LOWER QUARTILE (LQ / Q1) ⇒ Point that is greater than 25% of the data

UPPER QUARTILE (UQ / Q3) ⇒ Point that is greater than 75% of the data

→ **PERCENTILES** ⇒ An extension of this breakdown of the data is spliting it into 1% sections, called 'percentiles'

DİSPERSION

→ **RANGE** ⇒ HIGHEST VALVE - LOWEST VALVE = RANGE

→ **INTERQUARTILE RANGE** ⇒ $IQR = Q_3 - Q_1$

→ **VARIANCE** ⇒ Another measure of spread of data. It is essentially the average of the differences between each valve and the mean.

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

→ **STANDARD DEVIATION** ⇒ Related meaure, just calculated by doing the square root of the variance. 5.0. is often used more frequently

CALCULATOR :

TI - nspire ⇒ Enter valves in 'List & Spreadsheets', then on a calculator page, press MENN → 6 : Statistick → 1 ...... 1 : One - var stats . S. D. is Qx

TI - 84 ⇒ Press STAT → 1 : EDIT → ENTER, then fill column with valves. Then press STAT → CALC → 1 : 1 var stats ENTER. S. D. is Qx

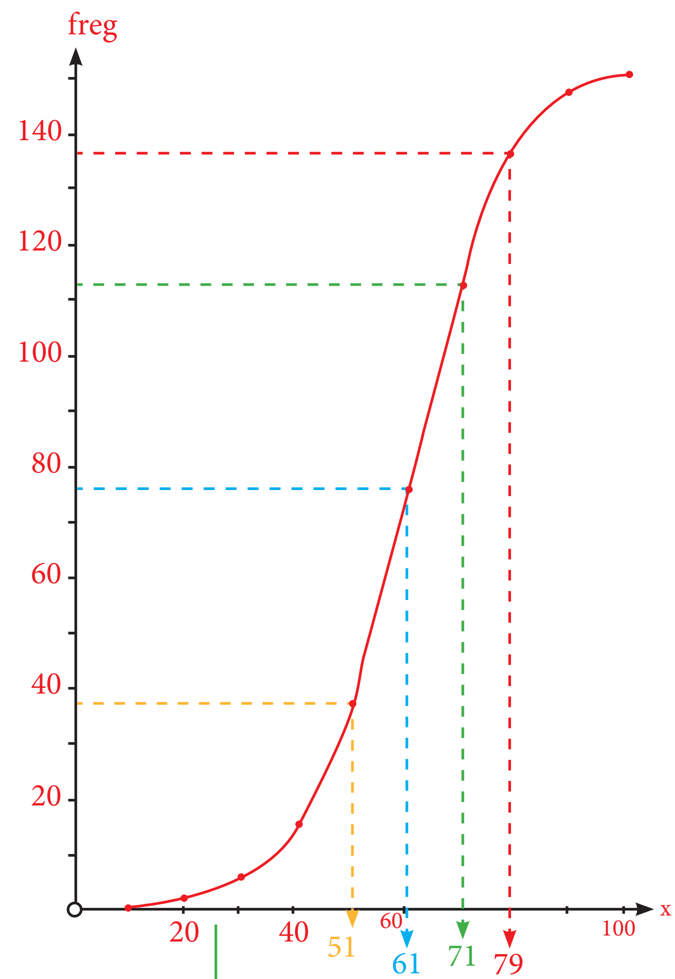
5.3 CUMULATIVE FREQUVENCY

⇒ It is hard to see what proportion of the data is above or below a certain valve. Adding on the freguency of each successive class, we build a 'cumultative freg' colomn. We plot this against the UB of each class, them analyse :

Add column

EG - 1 ⇒

SCORE	FREQ,	U.B	C.F.
$10 \le x < 20$	2	20	2
$20 \le x < 30$	5	30	7
$30 \le x < 40$	7	40	14
$40 \le x < 50$	21	50	35
$50 \le x < 60$	36	60	71
$60 \le x < 70$	40	70	111
$80 \le x < 80$	27	80	138
$80 \le x < 90$	9	90	147
$90 \le x < 100$	3	100	150



↳ E.G. ⇒ Median = 61

LQ = 51
UQ = 71
IQR = 71 - 51 = 20

90^{th} percentile = 79

➔ ANALYSIS ⇒

FINDING MEDIAN

↳ Start at $\left(\frac{n+1}{2}\right)^{th}$ position on y - axis, trace across to the curve, then down to the x - axis, then read your median. (See blue line)

LOWER QUARTILE

↳ Start at $\left(\frac{n+1}{2}\right)^{th}$ position then follow the same process as above.

UPPER QUARTILE

↳ Start at $\left(\frac{3(n+1)}{4}\right)^{th}$ position then follow the same process as above.

FINDING PERCENTILES

↳ Similar rocess, but with $\left(\frac{p(n+1)}{100}\right)$ if you are trying to find the pth percentile

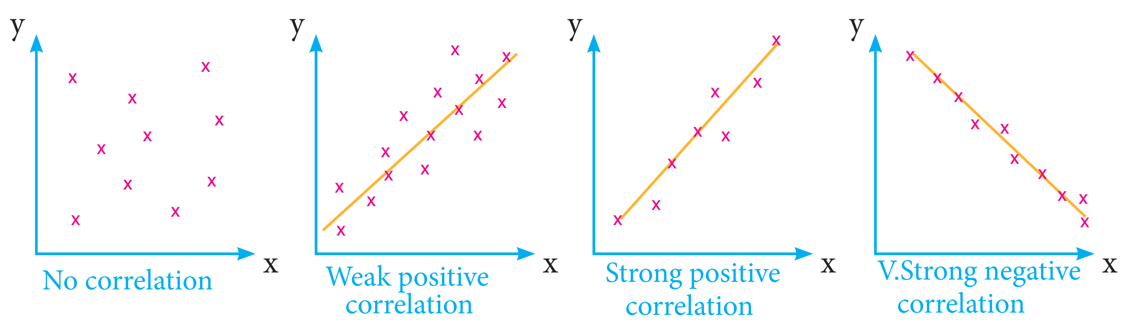
↳ E.G.(above) : 90 th percentile :

$$\frac{90(\text{ısı})}{100} = 1.35.9$$

5.4 CORRELATION

➔ **CORRELATION** ⇒ Testing the correlation between two sets of data, x and y, means that you are testing whether a change in x causes a similar change in y, and to what extent.

➔ **GRAPHS** ⇒ This is one way of showing what the strengths of correlation mean in reality :



➔ **CORRELATIONS COEFFICINT** ⇒ An accunake lekpor of caycuyakioh conneyakioh ' r ' measured on a scale form -1 (perfect negative corr.) to +1 (perfect positive correlation)

↳ **BY HAND** ⇒ $r = \dfrac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 (y-\bar{y})^2}}$ (You will use a calculator in exams)

↳ **BY HAND** ⇒

TI - 84

→ STAT | 1 : Edit ENTER

→ Put x valves in L1, y in L2

→ MODE | Check diagnastics are ON

→ 2^{nd} QUIT

→ STAT | CALC | 4 : Lin Reg (ax + b)

→ ENTER until info shown

TI - ASPIRE

→ Add 'Lists & Spreadsheet'

→ Name columns : ' x ' & ' x '

→ Type data

→ MENU | 4 : Stats 1 : Stat Calc. | 3 : Lin. Reg (mx + b)

→ Choose ' x ' for X list, ' y ' for Y list

→ ENTER , then abserve

➔ CORRELATION ⇒ On the same calculator screen as the r valve, you will see a valve for ' a ' & ' b '. This creates an equation in the form $y = ax + b$.

This is a more advanced version of what you may have seen as a 'line of best fit', which is more of an estimate.

You may well be asked to use this equation to estimate y valves given x valves

5.5 PROBABILITY

➔ DEFINITIONS ⇒ ' TRIAL ' → Each time an experiment is repented.

' OUTCOMES ' → Possible results of one trial.

' SAMPLE SPACE (U) ' → Set of all possible outcomesin an experiment

➔ PROB. RULES ⇒ If A is a set of results from an experiment with all equally likely results, then :

→ Prob. Of A occuring = P(A) $\frac{\text{EVENTS IN A}}{\text{EVENTS IN U}} = \frac{n(A)}{n(U)}$

EG - ⇒ Probability of roolling a multiple of 3 on a fair dice :

↳ **SOL :** P (Mult. of 3) = $\frac{2}{6} = \frac{1}{3}$

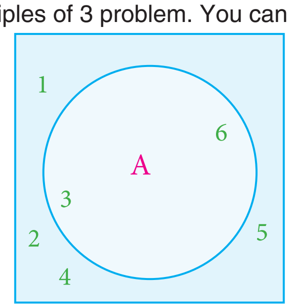
⇒ Two events are complementary if exactly one the the two events must occur, so :

→ $P(A) + P(A') = 1$

EG - ⇒ Rolling a 6 and not rolling 6 are complementary, as the probabilities add to 1 :

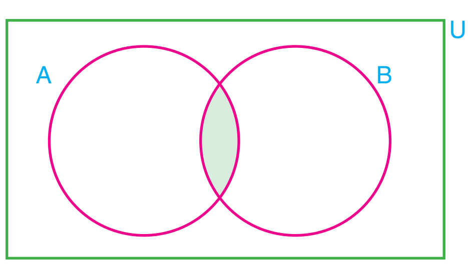
↳ **SOL :** $\frac{1}{6} = \frac{5}{6} = 1$

➔ VENN DIAGRAMS ⇒ The venn diagram on the left represents the earlier multiples of 3 problem. You can see where the 2 out of 6 probability is derived from :

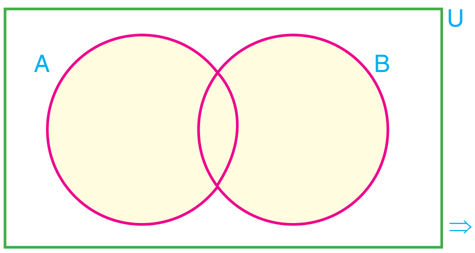


➔ INTERSECTION ⇒ The venn to the right represent th 'intercektion' between sets A & B, denoted $A \cap B$. These are the elements common to both A and B :

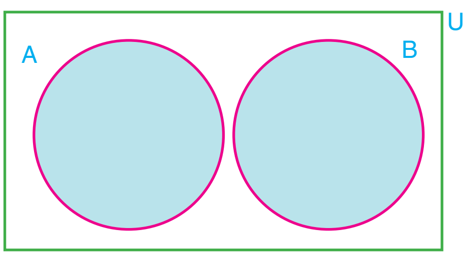
$A \cap B$ ⇒



➔ INTERSECTION ⇒ Denoted A ∪ B, represents elements in A or B or both :



⇒ A ∪ B

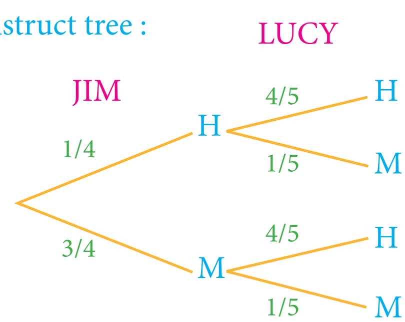


- * These sets are disjoint
- * We say A ∩ B = ∅, where ∅ is the empty set
- * A & B are mutually exclusive - meaning you can have both accurning at the same time

➔ PROBABILITY TREES ⇒ When you have two or more trials, and the possible outcomes are not too numerous, wo can use 'probability trees' :

EG - ⇒ In archery, Jim hits a target $\frac{1}{4}$ of the time, andd Lucy hits it $\frac{4}{9}$ of the time.

i) Construct tree :



OUTCOME	PROB.
H & H	1/4 x 4/5 = $\boxed{4/20}$
H & M	1/4 x 1/5 = $\boxed{1/20}$
M & H	3/4 x 4/5 = $\boxed{12/20}$
M & M	3/4 x 1/5 = $\boxed{3/20}$

ii) Prob of at least one hit ? 1^{st} , 2^{nd} & 3^{rd} options : 4/20 + 1/20 + 12/20 = $\boxed{17/20}$

5.6 COMBINED PROB :

➔ **UNION PROB.** $\Rightarrow$ $\boxed{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$

➔ **INTERSECTION** $\Rightarrow$ $\boxed{P(A \cap B) = P(A) \times P(B)}$ [if independent]

IN F.B.

➔ **CONDITIONAL** $\Rightarrow$ $\boxed{P(A / B) = \frac{P(A \cap B)}{P(B)}}$

$\rightarrow$ where P(A / B) means : the prob. of a occurring, given that B has occurred.

➔ **INDEPENDENT** $\Rightarrow$ Two event are independent if one event occurring does not affect the probability of the other event occuming.

It can also be shown formally by checking whether :

$P(A / B) = P(A)$ holds

OR

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

5.7 DISCRETE RANDOM VARIABLES

➔ **RANDOM VARIABLE** $\Rightarrow$ This represents, in number form, the possible outcomes which could occur for some random experiment.

➔ **DISCRETE** $\Rightarrow$ A set of distinct possible valves, which you can count.

➔ **CONTINUOUS** $\Rightarrow$ Valves are measured between a certain range.

➔ **PROBABILITY DISTRIBUTIONS** $\Rightarrow$ For any random veriable, there is a prob. dist. which desribes the probability of eac valve occuring

↳ **NOTATION** $\Rightarrow$ The prob. that the variable X takes valve x is denoted : $P(x = x)$

EG - $\Rightarrow$ Tossing 2 coins, counting how many ' heads ' occur

↳ $P(x = 0) = 1/4$, $P(x = 1) = \frac{2}{4} = \frac{1}{2}$, $P(x = 2) = 1/4$

➔ **RULE** $\Rightarrow$ For someting to be a valid probability distribution fînction, the sum of the probabilities needs to equal 1.

i. e : $\sum P(x) = 1$

EG - $\Rightarrow$ (From above ↑) : $\frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1$

EG - ⇒ Find k :

x	0	1	2
$P(x = x)$	0,3	k	0,5

→ 1 - 0,3 - 0,5 = 0,2 , k = 0,2

➔ EXPECTED VALVE ⇒ Take n trials of an experiment, where in each of the trials the event has prob. of p of occurring, then the number of times we. expenct the event to accur is $n \times p$

The expected outcome for the random variable X is the mean result, μ

$$E(X) = \mu \sum_{i=1}^{a} x_i \, q_i$$

EG - ⇒ In a magazine store, 23% of customers purchased 1 magazine, 38% bought 2, 21% bought 3, 13% bought 4, and 5% bought 5. Calculate the expected number of magazines bought

↳

x_i	1	2	3	4	5
p_i	0.23	0.38	0.21	0.13	0,05

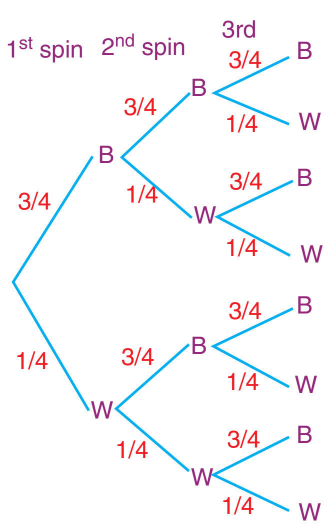
$$\mu = \sum x_i \, p_i$$
$$= 1(0.23) + 2(0.38) + 3(0.21) + 4(0.13) + 5(0.05)$$
$$= 2.39 \text{ magazines}$$

5.8 BINOMIAL DISTRIBUTION

➔ BINOMIAL EXPERIMENTS ⇒ These are experiments where there are just 2 possible results : success or failure (event occuring or not). This is then repated in a number of independent trials. For each trial, prob. of success is p, and prob. of failure is there fore $1 - p$

➔ OPENING PROBLEM ⇒ A ' spinner ' has three blue edges and one white edge so it has 5/4 prob. of getting blue, and 1/4 of getting white.

If we call spinning a blue a ' success ', so p = 3/4, then we can analyse the prob. of getting 0,1 , 28,3 success from 3 spins.



PROBABILITIES ⇒

$P(3 \text{ blues}) = P(x = 3) = \left(\frac{3}{4}\right)^3 = 0.4219$

$P(0 \text{ blues}) = P(x = 0) = \left(\frac{1}{4}\right)^3 = 0.0156$

Those two are easy to calculate, but 1 blue and 2 blues both have three different paths leading to that total. So we have a 'multiplier' of x 3 :

$P(1 \text{ blues}) = P(x = 1) = \left(\frac{3}{4}\right)\left(\frac{1}{4}\right)^3 \underline{\times 3} = 0.1406$

BWW, WBW or WWB

$P(2 \text{ blues}) = P(x = 2) = \left(\frac{3}{4}\right)^2\left(\frac{1}{4}\right) \underline{\times 3} = 0.4219$

Notice that 0.0156 + 0.1406 + 0.4219 + 0.4219 = 1

➔ **FUNCTION** ⇒ Above is just an example for 3 trials. We need to know how calculate probabilities

. for any n. The multipilier effect is illustrated in the $\binom{n}{r}$ past of the following function :

For n trials, the probability that there are r successes & n - r failures is :

$$\boxed{P(x = r) = \binom{n}{r} p^r (1 - p)^{n-r}} \quad \text{for } r = 0,1 \ldots\ldots n$$

EG - ⇒ 8 rolls of a dice, prob. of rolling 5 sixes ?

n = 8

r = 5

p = 1/6

$$P(x = 5) = \binom{8}{5} \left(\frac{1}{6}\right)^5 \left(\frac{5}{6}\right)^3 = 56.\left(\frac{1}{6}\right)^5 \left(\frac{5}{6}\right)^3 = 0{,}00417$$

See $\boxed{1.3}$

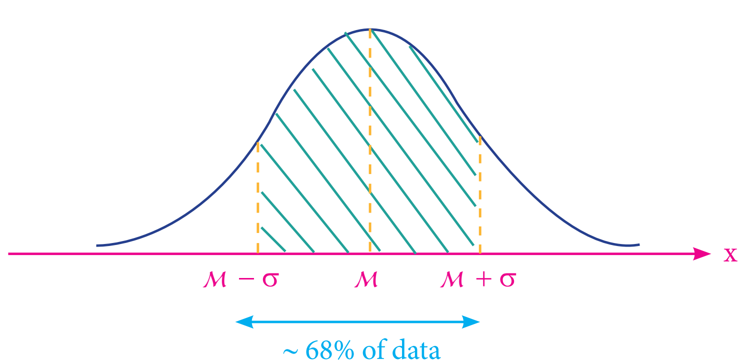
5.9 NORMAL DISTRIBUTION

When you have a large single set of valves, a histogram / bar chart becomes unfeasıble. So an alternative is to approximate this into one bell - shaped curve

➔ **PROPERTIES** ⇒

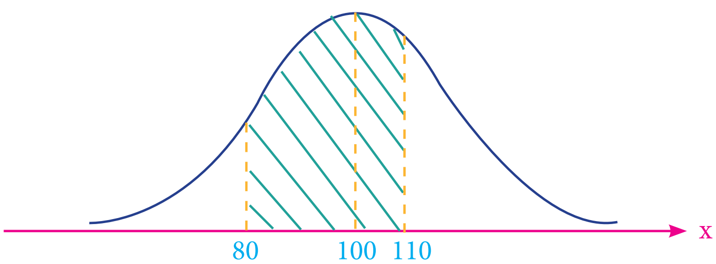
* SYSMETRICAL BELL - SHAPED CURVE
* THE AREA UNDER THE CURVE IS EQUAL TO 1 (or 100%)
* IT IS DEFINED BY ITS MEAN (μ) & STANDART DEV. (σ)
* THE MEAN IS IN THE CENTER
* ~ 68% OF THE DATA IS WITHIN 1σ OF THE MEAN
* ~ 95% IS + 2σ OF THE MEAN ~99% IS $\pm$ 3σ

➔ **GENERAL DIAGRAM** ⇒



➔ NORMAL PROB. CALCULATIONS ⇒ When you are given the $\mathcal{M}$ & σ of a curve, you can find the probability that a randomly selected valve would lie within a certain range of x :

EG - ⇒ A set of 2000 IQ scores is normally distributed with $\mathcal{M}$ = 100 & σ = 10 . Find the probability ıf picking an IQ between 80 & 110 :



SOL : (TI - nspire)

→ MENU | 5 : Prob. | 5 : Dist. | 2 : Normal CDF

→ Enter L.B. = 80, U.B =110, $\mathcal{M}$ = 100,σ = 10

→ Answer will be shown as decimal

(TI - 84)

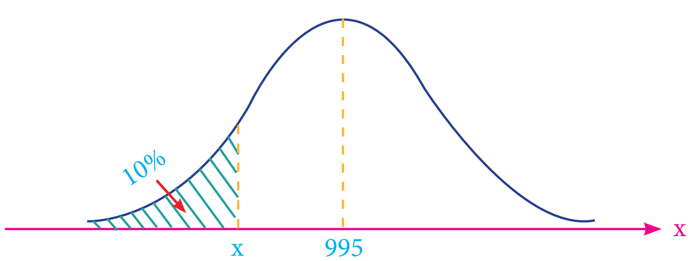
→ 2nd DISTR | 2 : normal cdf ENTER

→ Rest is the same as above ...

NOTE : If no US / LB given, you may have to use 10^{99} & -10^{99} respectively.

➔ INVERSE NORM CALCULATIONS ⇒ This style of question is the opposite. You will be given an area / prob., and asked to find a boundary :

EG - ⇒ The vol. of milk coutans has : $\mathcal{M}$ = 995 mL & σ = 5 mL. 10% of coutans are < x mL. Find x :



SOL : (TI - nspire)

→ MENU | 5 : Prob. | 5 : Dist. | 3 : Inverse.

→ Enter Area = 0,1 $\mathcal{M}$ = 995 ,σ = 5

→ The Answer (998.6 mL) will be an upper boundary. To do find a Lower boundary do [1 - answer]

(TI - 84)

→ 2nd DISTR | 3 : Invnorm | ENTER

→ Rest is the same as above ...

➔ INVERSE NORM CALCULATIONS ⇒ In certain cases, it is helpful to work with a normal distribution with $\mathcal{M}$ = 0 , σ = 1. This is called the Z - distribution, or Z ~ N (0,1)

➔ METHOD ⇒ To use the Z - dist, you must also transform each x - valve to what we call a z - score

$z = \frac{x - \mu}{\sigma}$ [This also represents how many S.D's from μ it is]