Statistical Methods

*) Collecⁿ, Classificaⁿ & the analysis of data is the statistical method which helps to make decisions.

*) Central Tendency Measure :- (Average calculation)

1. Arithmetic mean : AM → ~~If data is~~
2. Median
3. Mode.

Arithmetic Mean : If data is available for 1, 2, 3, 4, 5, 6, 7, 8, 9 →

$$AM = \frac{1+2+3+\ldots+9}{9} = 5.$$

In general ⇒ $$A.M = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

For Big data calculaⁿˢ, use freqⁿ distribuⁿ table.

& $$\bar{x} = \frac{\sum f_i x_i}{N = \sum f_i}$$

Ex: Find A.M of -

x	f	$x \times f$
0	4	0
1	5	5
2	12	24
3	12	36
4	13	52
5	16	80
6	18	90
7	13	91
8	12	96
9	5	45
10	6	60
		579

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{579}{113} = 5.12$$

<u>Note:</u> To reduce the calculaⁿ, in the freqⁿ distribuⁿ table define the no. $d = x - A$, where A = Assumed mean or mid value. The A.M is calculated by

$$\bar{x} = A + \frac{\sum fd}{\sum f}$$

Still to reduce further calculaⁿˢ divide the data into class intervals.

Let $u = \frac{x-A}{h} = \frac{d}{h}$ where h = width of interval.

then $\bar{x} = A + h\frac{\sum fu}{\sum f}$

Q. Marks obtained in Maths paper are classified as : Find A.M.

C.I.	No. of students	Mid Value of C.I.	$u = \frac{x-A}{h}$ Let A=45	fu
0-10	8	5	-4	-32
10-20	20	15	-3	-60
20-30	14	25	-2	-28
30-40	16	35	-1	-16
40-50	20	45	0	0
50-60	25	55	1	25
60-70	13	65	2	26
70-80	10	75	3	30
80-90	5	85	4	20
90-100	2	95	5	10
	$\sum f = 133$			$\sum fu = -25$

$\bar{x} = A + \frac{h\sum fu}{\sum f}$

$\bar{x} = 45 + \frac{10\times(-25)}{133}$

$\bar{x} = 43.12$

*) Joint Mean :

$\bar{z} = \frac{n_1\bar{x} + n_2\bar{y}}{n_1+n_2}$ where $\bar{x}$ = A.M. of set 1

$\bar{y}$ = A.M. of set 2.

OR $\bar{z} = \frac{\sum fx + \sum fy}{\sum f + \sum f}$

(for set 1) (for set 2)

or $= N_1 + N_2$

Q. Group 1

C.I.	Marks
0-10	5
10-20	6
20-30	18
30-40	18
40-50	9

Group 2

C.I.	Marks
0-10	8
10-20	15
20-30	18
30-40	13
40-50	6

Find joint Mean

Ans : Group 1

Mid Value	$u = \frac{x-A}{h}$	$f\times u$
5	-2	-10
15	-1	-6
25	0	0
35	1	15
45	2	18
		~~40~~ 17

Group 2

	$u = \frac{x-A}{h}$	fu
5	-2	-16
15	-1	-15
25	0	0
35	1	13
45	2	2
		-16

$\bar{x} = A + h\frac{\sum fu}{\sum f}$

$= 25 + \frac{10\times(17)}{\cancel{45}\ 80}$

$= 21.6$

Ex: The mean weekly salary paid to 300 employees of a fi[rm] is Rs 1470. These are ~~200~~ 200 male employees & remaining female. If mean salary of males is 1505. Find mean s[alary] of females.

Soln:
$$\bar{z} = \frac{200 \times 1505 + 100 \times \bar{y}}{300}$$

$$1470 = \frac{301000 + 100\bar{y}}{300}$$

$$441000 = 301000 + 100\bar{y}$$

$$\bar{y} = 1400$$

*) coefficient of variation : $C.V. = \frac{\sigma}{\bar{x}} \times 100$ (used for [comparison] of data).

Q. Goals scored by 2 teams given as

Goals	0	1	2	3	4
Matches A	27	9	8	5	4
B	17	9	6	5	3

Find which tea[m is more] consistent.

Soln:
$\bar{x}_A = 1.0566$ $\sigma_A = 1.30914$

$\bar{x}_B = 1.2$ $\sigma_B = 1.30766$

$$C.V. = \frac{\sigma_A}{\bar{x}_A} \times 100 = 123.90119$$

$$C.V. = \frac{\sigma_B}{\bar{x}_B} \times 100 = 108.9716$$

Team B is more consistent as C.V. for B<A.

*) Moments, Skewness & Kurtosis.
r^{th} moment about the mean of a distribution is μ[...]
where $N = \Sigma f$
& $\bar{x} = A.M.$

if $r=0 \Rightarrow \mu_0 = \frac{\Sigma f}{N} = 1$

$r=1 \Rightarrow \mu_1 = \frac{1}{N}\Sigma f(x-\bar{x})$

$= \frac{\Sigma f x}{N} - \frac{\Sigma f \bar{x}}{N}$

$\mu_1 = \bar{x} - \bar{x} = 0$

& $r=2 \Rightarrow \mu_2 = \Sigma f(x-\bar{x})^2$

$= \sigma^2 =$ variance

$\mu_3 = \frac{1}{N}\Sigma f(x-\bar{x})^3$ etc.

In actual calculation moments are obtained by using

$\mu_r' = \frac{1}{N}\Sigma f(x-A)^r$ A = assumed mean

for $r=0$, $\mu_0' = 1$

for $r=1 \Rightarrow \mu_1' = \bar{x} - A$

for $r=2 \Rightarrow \mu_2' = \frac{1}{N}\Sigma f(x-A)^2$ = mean square deviation.

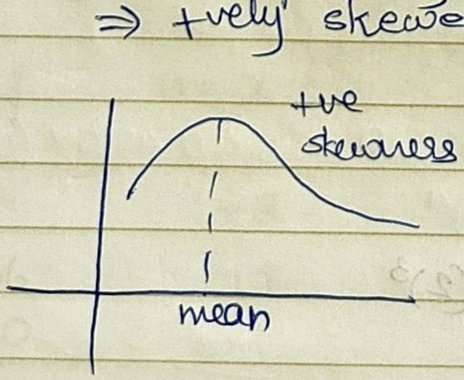
In general, using relⁿ betⁿ μ_r & μ_r'

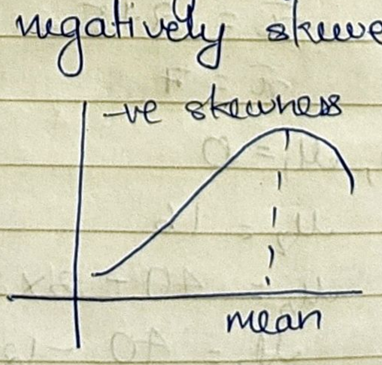
$\therefore \mu_0 = 1$, $\mu_1 = 0$ & $\mu_2 = \mu_2' - \mu_1'$ $\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$

$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$

*) Skewness: If frequency curve stretches right of its mean.

$\Rightarrow$ +vely skewed otherwise negatively skewed.



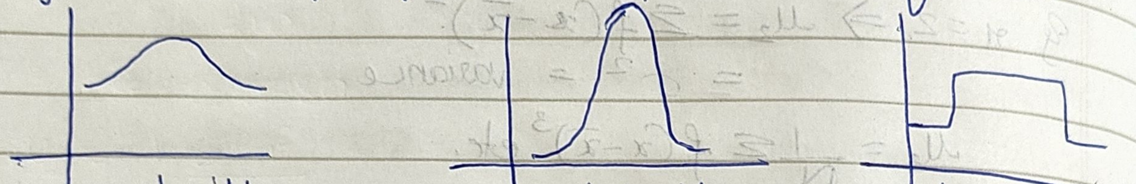


$$\text{skewness} = \frac{3\,(\text{Mean} - \text{Median})}{\text{standard deviation}}$$

coeff. of skewness $\beta_1 = \dfrac{\mu_3^2}{\mu_2^3}$

*) Kurtosis:

To get an idea for flatness or peakedness of curve



mesokurtic   leptokurtic $\beta_2 > 3$ platykurtic $\beta_2 < 3$

measured by $\beta_2 = \dfrac{\mu_4}{\mu_2^2}$

for normal distribution $\beta_2 = 3$

Q. The 1st 4 moments of a distribution are about 5 are 2, 20, 40 & 50. Obtain 1st 4 central moments, mean, st. dev (σ) & coff β_1 & β_2.

Sol^n: Here A (assumed mean) = 5

& given $\mu_1' = 2$, $\mu_2' = 20$, $\mu_3' = 40$, $\mu_4' = 50$

To find $\mu_1, \mu_2, \mu_3, \mu_4$ & σ, $\bar{x}$, β_1 & β_2.

Since $\mu_1' = \bar{x} - A$

$\bar{x} = 2 + 5$

$\bar{x} = 7$

Also, $\mu_1 = 0$

$\mu_2 = 16$

$\mu_3 = 40 - 3 \times 20 \times 2 + 2 \times (2)^3$

$\mu_3 = 40 - 120 + 16$

$\mu_3 = -64$

$\mu_4 = 40 - \cancel{4 \times 40 \times 2}\ 4 \times 40 \times 2 + 6 \times 20 \times (2)^2 - 3 \times (2)^4$

$4 \times 4 \times 16 \times 16$

$\mu_4 = 162.$

$\sigma^2 = 16$

$\sigma = 4$

$\beta_1 = 1 \Rightarrow$ +ve skewness.

$\beta_2 = \frac{(-64)^2}{(16)^3}$

$\beta_2 = 0.63 < 3$

Q. Calculate 1st 4 moments about the mean if distribution table is

x	2	2.5	3	3.5	4	4.5	5
f	4	36	80	90	70	40	10

Find β_1 & β_2.

Soln: Let A = assumed mean = 3.5

$\mu_r' = \frac{1}{N} \Sigma f(x-A)^r$

Let $u = \frac{x-A}{h} \Rightarrow x - A = uh$

$\mu_r' = \frac{1}{N} \Sigma f(uh)^r$

using $h = 0.5$

$\mu_r' = h^r \frac{\Sigma f(u)^r}{\Sigma f = N}$

$\therefore \mu_1' = h \frac{\Sigma fu}{\Sigma f} \qquad \mu_2' = h^2 \frac{\Sigma fu^2}{\Sigma f}$

x	f	$u = \frac{x-A}{h}$	fu	fu^2	fu^3	fu^4
2	4	-3	-12	36	-108	324
2.5	36	-2	-72	~~164~~ 144	~~308~~ -288	~~606~~ 576
3	60	-1	-60	60	-60	60
3.5	90	0	0	0	0	0
4	70	1	70	70	70	70
4.5	40	2	80	160	320	640

5	10	3	30	90	270	810
$\Sigma f = 310$			$\Sigma fu = 36$	$\Sigma fu^2 = \cancel{5840} = 560$	$\Sigma fu^3 = 204$	$\Sigma fu^4 = 2480$.

$$\mu_1' = \frac{h \Sigma fu}{\Sigma f}$$

$$= \frac{0.5 \times 36}{310}$$

$$= 0.0580$$

$$\mu_2' = \frac{h \Sigma fu^2}{\Sigma f}$$

$$= \frac{(0.5)^2 \times 560}{310}$$

$$= 0.4516$$

$$\mu_3' = \frac{h^3 \Sigma fu^3}{\Sigma f}$$

$$= 0.08225$$

$$\mu_4' = 0.5$$

$$\mu_1 = 0$$

$$\mu_2 = \mu_2' - \mu_1'^2 = 0.44822$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$$

$$= 0.08225 - 3 \times 0.4516 \times 0.0580 + 2 \times (0.0580)^3$$

$$= 0.00\cancel{40}398$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$$

$$= 0.5 - 4 \times 0.08225 \times 0.0580 + 6 \times 0.4516 \times (0.0580)^2 - 3(0.0580)^4$$

$$= 0.4899.$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0.000175 \qquad \beta_2 = \frac{\mu_4}{\mu_2^2} = 2.43874.$$

*) Karl Pearson's coeff. of correlation

$$r(x,y) = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

$$cov(x,y) = \frac{1}{n} \Sigma (x_i - \bar{x})(y_i - \bar{y}).$$

For bivariate data

$$r = \frac{\Sigma xy - n\bar{x}\bar{y}}{\sqrt{(x^2 - n\bar{x}^2) \times (y^2 - n\bar{y}^2)}}.$$

Q. Data given as

Export	10	11	14	14	20	22	16	12	15	13.	$\to x$
Import	12	14	15	16	21	26	21	15	16	14	$\to y$.

Calculate coeff. of correlation bet^n import & export values.

Sol^n:

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{147}{10} = 14.7$$

$$\bar{y} = \frac{\Sigma y_i}{n} = \frac{170}{10} = 17$$

$$\Sigma x^2 = 2291$$

$$\Sigma y^2 = 3036$$

$$\Sigma xy = 2638.$$

$$r = \frac{2638 - 10 \times 14.7 \times 17}{\sqrt{(2291 - 10 \times (14.7)^2) \times (3036 - 10 \times (17)^2)}}$$

$$= \frac{139}{\sqrt{130.1 \times 166}}$$

$$r = 0.9458$$

*) Regression lines:

*) To find lines of regression:

① Eqⁿ of line of regression of y on x ⇒

$$(y-\bar{y}) = r\frac{\sigma_y}{\sigma_x}(x-\bar{x})$$

② Eqⁿ of line of x on y ⇒ $(x-\bar{x}) = r\frac{\sigma_x}{\sigma_y}(y-\bar{y})$

here $r\frac{\sigma_y}{\sigma_x}$ = coeff. regression = b_{yx}, σ_x = S.D. of x

$r\frac{\sigma_x}{\sigma_y}$ = " " = b_{xy}, σ_y = S.D. of y

r = coeff. of correlaⁿ.

$\bar{x}$, $\bar{y}$ are A.M ⇒ satisfy st. line eqⁿ.

③ Coeff. of correlaⁿ r = G.M. of regression coeff.

$= \sqrt{b_{xy}\, b_{yx}}$ where $b_{xy} = r\frac{\sigma_x}{\sigma_y}$

$b_{yx} = r\frac{\sigma_y}{\sigma_x}$.

Q. Find the lines of regression for the data

x	10	14	19	26	30	34	39
y	12	16	18	26	29	35	38

& estimate y for x = 14.5, estimate x for y = 29.5.

Solⁿ.

$\bar{x} = \frac{172}{7} = 24.57.$ $\Sigma xy = 4904.$

$\bar{y} = \frac{174}{7} = 24.85$ $\Sigma x^2 = 4910$

n = 7. $\Sigma y^2 = 4910$

4910

4910 $(24.5)^2$

$$\sigma x = \frac{1}{n}\left[\Sigma(x_i^2) - \bar{x}^2\right] = 615.67.$$

$$= \frac{1}{10}\left[(100 - (24.57)^2) + (14^2 - 24.5^2) + \right.$$

$$r = \frac{4907 - 10 \times 24.57 \times 24.85}{\sqrt{(4910 - 10 \times (24.57)^2)(4910 - 10 \times (24.85)^2)}}$$

$= 0.995$

$\sigma y = 613.21.$

$\therefore b_{yx} = r\frac{\sigma y}{\sigma x} = 0.9194$

$b_{xy} = r\frac{\sigma x}{\sigma y} = 1.0749.$

lines of regression

$y - 24.8 = 0.9194(x - 24.5)$ — ①

& $x - 24.5 = 1.074(y - 24.8)$ — ②.

put $x = 14.5$ in ① $\Rightarrow y =$

put $y = 29.8$ in ② $\Rightarrow x =$

Probability Theory

*) Sample Space: Set of all possible outcomes in an event.

1) Tossing a coin $\Rightarrow S = \{H, T\}$
2) Throwing a die $\Rightarrow S = \{1, 2, 3, 4, 5, 6\}$.
3) Throwing 2 dice $\Rightarrow S = \{(1,1)\ (1,2)\ (1,3)\ (1,4) \ldots\}$.

*) Def^n of Prob: If A is event of happening then $P(A) = \frac{m}{n} = \frac{\text{favorable ways}}{\text{Total outcomes} = s} = p$ denoted by

$q = \therefore P(\bar{A}) = \frac{n-m}{n} = 1 - \frac{m}{n} = 1 - p$ where $\bar{A}$ = not happening of event.

$\therefore \boxed{p + q = 1}$

Independent Events: Happening of A has nothing to do with happening of B & vice-versa.
But if event B is dependent upon happening of event A, then they are dependent event.

Note: $^nC_r = \frac{n!}{r!(n-r)!}$

Ex: Find the probability of drawing a queen in a pack of 52 cards.

Sol^n: $P(A) = \frac{m}{n} = \frac{4}{52} = \frac{1}{13}$

Ex: Find the prob. that a leap year selected at random will contain 53 sundays.

Sol^n: $\because$ leap year contains 366 days, 52 weeks
$\therefore$ 52 sundays will be there.
Sunday combina^n = Sat. sun & mon-sun.
$\therefore P(53 \text{ sundays}) = \frac{2}{7}$

$4 \times 4 \times 4$

Q. From a pack of 52 cards, 3 cards are drawn at random. Find the probability that they form king, queen & jack combinaⁿ.

Soln: n = total ways of drawing 3 cards $= {}^{52}C_3$

$$= \frac{52!}{3!\,(49)!}$$

$$= \frac{52 \times 51 \times 50}{3 \times 2 \times 1}$$

$$= 22100$$

Drawing a king card $= {}^4C_1$

" queen " $= {}^4C_1$

" jack " $= {}^4C_1$

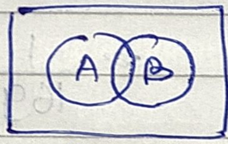
$\therefore$ m = All three combination $= {}^4C_1 \times {}^4C_1 \times {}^4C_1$

$m = 64$.

$\therefore$ $P(A) = \frac{m}{n} = \frac{64}{22100} = 0.00289$.

#) Algebra used for prob.

1) If A & B are any 2 events.



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

If $A_1, A_2, A_3 \ldots A_n$ are mutually exclusive events.

then $P(A_1 \cup A_2 \cup A_3 \ldots \cup A_n) = P(A_1) + P(A_2) + P(A_3) + \ldots P(A_n)$.

2) For compound prob:

$P(A \cap B) = P(A) \cdot P(B/A)$.

where B/A = Occurence of B when A has already occured.

For independent events $\rightarrow$ $P(B/A) = P(B)$ or $P(A/B) = P(A)$.

then

$P(A \cap B) = P(A) \cdot P(B)$

$|||^{ly}$ $P(A_1 \cap A_2 \cap A_3) = P(A_1) \cdot P(A_2) \cdot P(A_3) \ldots$

*) Baye's Thm:

If $A_1, A_2, \ldots A_n$ are mutually exclusive events whose union is the sample ~~set~~ space, & A is any event.

Then $$P(A_k/A) = \frac{P(A_k) \cdot P(A/A_k)}{\sum_{k=1}^{n} P(A_k) \cdot P(A/A_k)}$$

Q. 2 cards are drawn from pack of 52 cards. Find the prob. that they are both kings if

1) 1^{st} ~~packdo~~ card drawn is replaced.
2) 1^{st} card drawn is not replaced.

Soln. let A_1 = king in the 1^{st} draw

A_2 = king in the 2^{nd} draw.

$A_1 \cap A_2$ = king in both the draws.

1) $P(A_1 \cap A_2) = P(A_1) \cdot P(A_2)$

$= \frac{4}{52} \times \frac{4}{52}$

$\boxed{= \frac{1}{169}}$

2) ~~$P(A) = \frac{4}{52} \times \frac{3}{51}$~~ $P(A_1 \cap A_2) = P(A_1) \cdot P(A_2/A_1)$

$= \frac{4}{52} \times \frac{3}{51}$

$= \frac{3}{663}$

$= \frac{1}{221}$

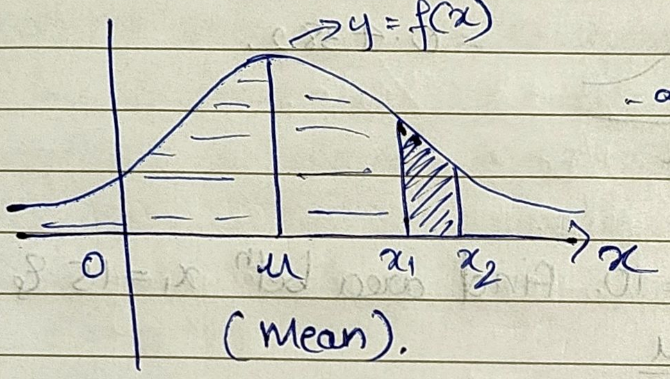
Q. A factory has 2 machines, machine 1 is producing ~~30~~ 30% of output and machine 2 produces 70%. If 5% of items produced by machine 1 are defective and 1% of items produced by machine 2 are defective. If the defective item is drawn at random. What is the probability that it is by machine I or II?

Ans: Let A_1 = drawing an item from Machine I.

A_2 = drawing an item from Machine II.

Let B = be event of drawing defective item from either I or II.

*) Normal Distr.



$$\int_{-\infty}^{\infty} y\,dx = 1$$

Binomial & poisson's distr. is used for discrete variable,

Normal Distr. is for continuous variable giving distr. curve.

∴ Area under the curve $y = f(x) \Rightarrow \int_{-\infty}^{\infty} y\,dx = 1$

Left side of $\mu = 0.5$, rt. side of $\mu = 0.5$.

& Prob. distr. given by

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

σ = S.D. μ = mean

Put $z = \frac{x-\mu}{\sigma}$ (newscale for x).

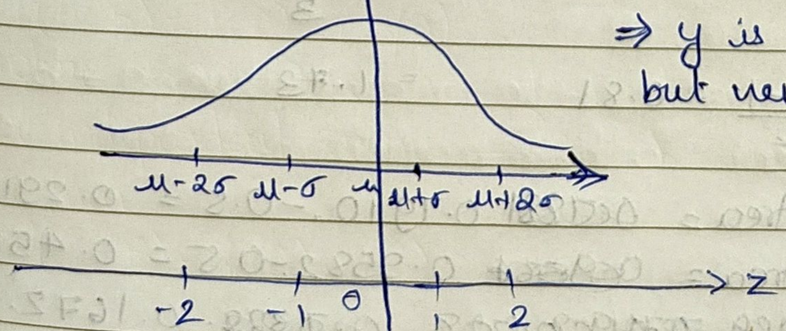
$$\therefore P(z) \Rightarrow y = \frac{1}{\sigma\sqrt{2\pi}} e^{-z^2/2}$$

$-\infty \le z \le \infty$

as x increasing

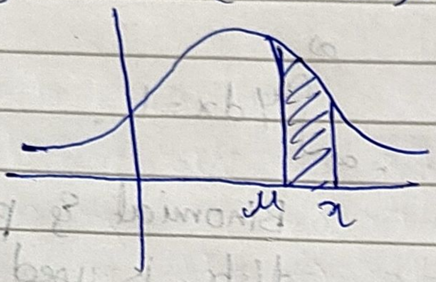
⇒ y is decreasing.

but never touch x-axis.



To solve the problem:
With μ & S.D. σ using z table.

Q. Find area under $x = 1.54$. Corresponding area for $z \rightarrow$ given by 0.9382 (z-table).



$0.9382 - 0.5$
$= 0.4382.$

Q. $\mu = 20$, $\sigma = 10$. Find area bet^n $x_1 = 15$ & $x_2 = 40$.

Sol^n: $z = \frac{x - \mu}{\sigma}$

$$= \frac{x_1 - \mu}{\sigma} = \frac{15 - 20}{10} = -0.5.$$

$$z = \frac{x_2 - \mu}{10} = \frac{40 - 20}{10} = 2.$$

Area corresponding to z_1 = $0.6915 - 0.5 = 0.1915$.
Area corresponding to z_2 = $0.9772 - 0.5 = 0.4772$.
$\therefore$ Total area required = 0.6687.
Total area required $0.1915 + 0.4772$
$= 0.6687.$

Q. Find normal distri. for $\mu = 1$, $\sigma = 3$

1) $3.43 \leq x \leq 6.19$ 2) $-1.43 \leq x \leq 6.19$

Sol^n: $z_1 = \frac{x_1 - 1}{3}$

$$= \frac{3.43 - 1}{3} = 0.81$$

$$z_2 = \frac{6.19 - 1}{3} = 1.73.$$

z_1 corr. Area = $0.7910 - 0.5 = 0.2910$
z_2 corr. Area = $0.9582 - 0.5 = 0.4582.$
Total area = 0.1672.

2) $z_1 = \frac{-1.43-1}{3} = -0.81$ $\qquad z_2 = \frac{6.19-1}{3} = 1.73$

$\cancel{0.7881}$ 0.7910 $\qquad 0.2910$

Area corr. to $z_1 = \cancel{0.7910} - 0.5 = \cancel{0.2910}$ $\cancel{0.2881}$

Area corr. to $z_2 = 0.9582 - 0.5 = 0.4582$

Total area 0.7492

Q. In statistics examinaⁿ 2000 appearead. Average marks obtai-ned by students are 50% with S.D. 5%. How many students are expected to obtain more than 60% of marks if marks are distributed normally.

Solⁿ:
$\mu = 50\% = 0.5$
$\sigma = 0.05$

$P(\geq 0.6) \Rightarrow \qquad z = \frac{0.6 - 0.5}{0.05} = 2.$

Area corr. to $z = 0.9772 - 0.5 = 0.4772.$

For 2000 student $\qquad \hookrightarrow 0.5 - 0.4772$

$= 0.\cancel{4772}\,0228 \times 2000 \qquad = 0.0228.$

$= 45.6$

$\approx 46.$ (approx).