Unit-3

Population:

A group of individuals under study is called a population.

Sample: (To generalise population)

A part of the population is called a sample.

Measure of Central tendency
- " " average
- " " location

- Arithematic Mean
- Median
- Mode
- Harmonic Mean
- Geometric Mean

HM:-

$$HM = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots \frac{1}{x_n}}$$

reciprocal of mean of reciprocals

H.M is calculated for a given set of observations $x_1, x_2, \ldots x_n$, $HM = \frac{1}{\frac{\frac{1}{x_1} + \frac{1}{x_2} + \cdots \frac{1}{x_n}}{n}}$

GM:-

$$GM = \sqrt[n]{x_1 . x_2 \ldots x_n} = (x_1 . x_2 . x_3 \ldots x_n)^{1/n}$$

$$\log G = \frac{1}{n} \log (x_1 x_2 x_3 \ldots x_n)$$

$$= \frac{1}{n} (\log x_1 + \log x_2 + \ldots \log x_n)$$

$$G = \text{Anti log} \left[\frac{1}{n} (\log x_1 + \log x_2 + \ldots \log x_n)\right]$$

Average - Balancing the data

→ NSSO - National Sample Survey Organisation (under Ministry of Finance Govern. of India)

Measure of skewness

Always tells you about whether the data is symmetric or not

Measure of kurtosis :-

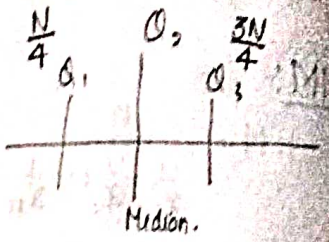
The kurtosis always deals with the study of the frequency curves and its shapes.

Measure of central tendency
Measure of dispersion → Range, S.D, Q.D, M.D = $\frac{\Sigma|x-M|}{N}$ (Mean deviation)
Measure of skewness
Measure of kurtosis

(M, mean, log, median)

Quartile deviation :-

Q.d. $\frac{Q_3 - Q_1}{2}$ third quartile - first quartile / 2



median - python prog.

$$Md = l + \frac{\frac{N}{2} - m}{f} \times c$$

Standard deviation :-

$$S.D = \sqrt{\frac{1}{n}\Sigma(x_i - \bar{x})^2} \quad \text{(ungrouped data)}$$

$$S.D = \sqrt{\left(\frac{\Sigma f_i x_i^2}{\Sigma f_i}\right) - \left(\frac{\Sigma f_i x_i}{\Sigma f_i}\right)^2} \quad \text{(grouped data)}$$

The square of the standard deviation is called variance.

$$Var = \frac{1}{n}(S.D)^2$$

Variation in the series of observations.

Ex:- 15, 20, 15, 16, 17, 19, 25, 30, 45, 20 - Find Standard deviation.

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
15	-7.2	51.84
15	-7.2	51.84
16	-6.2	38.44
17	-5.2	27.04
19	-3.2	10.24
20	-2.2	4.84
20	-2.2	4.84
25	2.8	7.84
30	7.8	60.84
45	22.8	519.84

$\bar{x} = \frac{222}{10} = 22.2$. $\sum (x_i - \bar{x})^2 = 777.6$

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

$$= \sqrt{\frac{1}{10} \times 777.6}$$

$$= \sqrt{77.76}$$

$$= 8.81816.$$

Percentile :-

Percentile is used to understand and interpret data.

The n^{th} percentile of a set of data is the value at which n percent of data is below it.

Ex:- Calculate 45^{th} percentile of the above data.

We use the following formula $L_p = (n+1)\frac{p}{100}$

15
15
16
17
19
20
20
25
30
45

$$L_{45} = (10+1)\frac{45}{100}$$

$$= 11 \times 0.45$$

$$= 4.95$$

$$P_{45} = 17 + (0.95)2$$

$$= 18.8$$

Find 75th percentile

$L_{75} = (10+1)\frac{75}{100}$

$= 11 \times 0.75$

$= 8.25$

$P_{75} = 25 + (0.25)5$

$= 26.25$

$L_{25} = (10+1)\frac{25}{100}$

$= 2.75$

$P_{25} = 15 + (0.75)1$

$= 15.75$

Range - 30

Q.D - 5.25

S.D - 8.81

```
n = int(input(" "))
x = len(n)
n.sort( )
if x%2 == 0:
    median 1 = n[x//2]
    median 2 = n[x//2 - 1]
    median = (median 1 + median2)/2
else
    median = n[x//2]
print ("median is :", median)
```

Covariance :

Variation btw 2 observations.

(x_1, y_1)

$\vdots$

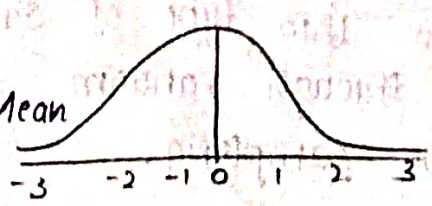
(x_n, y_n)

$\text{Covariance} = \frac{1}{n} \Sigma (x_i - \bar{x})(y_i - \bar{y})$

Skewness:-

$S_k = \frac{\text{Mean-Mode}}{\sigma}$ $(-1 < S_k < 1)$

Mode = 3Median - 2 Mean



$= \frac{\text{Mean} - (3\text{Median} - 2\text{Mean})}{\sigma}$

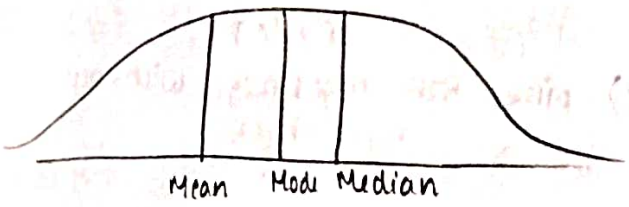
$S_k = \frac{3(\text{Mean} - \text{Median})}{\sigma}$

Skewness = 0

(Mean = Median = Mode)

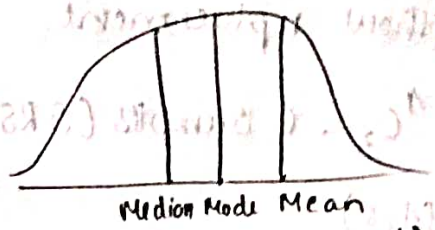
↓

symmetrical distribution



(+vely skewed)

$(0 < S_k < 1)$



(-vely skewed)

$(-1 < S_k < 0)$

- To know the characterstics of data we find skewness, mean, median . . .

Sampling:-

Sampling can be done in 3 types

- Random Sampling
- Purposive Sampling
- Mixed Sampling

- Random Sampling:-

In this type each unit in the population has an equal chance to include in the sample.

- Puposive Sampling:-

Here the sampling units are selected from the population for the purpose.

• Mixed Sampling:-

In this type of sampling some of the units are selected through random & remaining from purposive sampling.

→ In Random Sampling there are 2 types

- Sin without replacement
- with replacement.

Ex:- 2 numbers are to be drawn from 4 (4,5,6,7)

without replacement

$^4C_2$ = 6 samples (SRSWOR) - simple Random Sample with out replacement

(4,5)
(4,6)
(4,7)
(5,6)
(5,7)
(6,7)

with replacement

4^2 = 16 samples (SRSWR) - simple Random Sample with replacement.

(4,4)	(5,4)	(6,4)	(7,4)
(4,5)	(5,5)	(6,5)	(7,5)
(4,6)	(5,6)	(6,6)	(7,6)
(4,7)	(5,7)	(6,7)	(7,7)

problem:- (1M)

A population consist of 6 units draw the samples with the two schemes simple Random sampling with and without replacement of a sample 2.

Population = 1,2,3,4,5,6.

without replacement

$^6C_2 = \frac{6 \times 5}{2} = 15$ samples.

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

with replacement

$6^2 = 36$ samples

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

• We have 3 techniques in Random Sampling

1- Simple Random Sampling
2- Stratified Random Sampling
3- Systematic Sampling

2- Stratified R.S :- (no restrictions for population)

If the population is not homogeneous and it it is divided into k sections such that $N_1, N_2 \ldots N_k$ are the respective population sizes in k sections of the pop. Then if I have to draw a sample of size n from the population then I can draw $n_1, n_2, \ldots n_k$ samples from the respective sections in the population.

then

$$N_1 + N_2 + \ldots + N_k = N$$

$$n_1 + n_2 + \ldots + n_k = n$$

Then the sample is called stratified R.S.

3- Systematic Sample :- (the population must be in systematic order)

In systematic sampling the population size must be known population must be arranged in an systematic order. So, that it is convenient to draw sample units from population.

Example :-

If I want to check 10 bills out of 100 bills generated by the shop keeper then divide the population size by sample size i.e, $\frac{100}{10} = 10$. $\left(\frac{N}{n} = k\right)$

Now the first unit in systematic sample is selected at random say 5.

For every next sample unit add k to the random number.

The next sample units from the bills are 5+k i.e, 15 et
he sample units are 5, 15, 25, 35, 45, 55, 65, 75, 85 & 95
- There is restriction on sample. The population size
ust be known. Should be in systematic order.

In any experiment if the result is unique then it is a
on-Random " (Example: Chemistry & Physics experiments)

If the result is coming from several outcomes then
t is a Random Experiment.

Random Variable:-

A variable associated with a random experiment is
a Random variable.

or

A function that maps the sample space of a random
experiment to the real line $(-\infty, \infty)$.

Probability Distribution:-

A probability distribution is always defined for the values
assumed by a R.V. If $x_1, x_2 \ldots x_n$ are the values
" " " 'X' for each value of the R.V
if we assign probabilities & sum of the Prob. is
equal to 1. Then such a pair $(x_i, P(x_i))$ is called
a prob. dist.

$$X : x_1, x_2 \ldots x_n$$

$$P(X=x) : P(x_1), P(x_2) \ldots P(x_n)$$

$$\{x, P(x_i)\}$$

$$\sum_{i=1}^{n} P(x_i) = 1$$

Ex:- Write the prob. dist. if a single die is thrown. write for the face of the die (X).

X	1	2	3	4	5	6
P(X=x)	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

$$\sum_{1} P(x) = 1.$$

Binomial distribution:-

A discrete R.V 'X' is said to have a binomial distribution if its probability mass function is given by

$$P(x) = {}^{n}C_x p^x q^{n-x}, \quad x = 0, 1, 2 \ldots n$$
$$p+q = 1.$$

Bernoulli R.V:-

A R.V 'X' is said to be a bernoulli R.V if $X = 1$ with probability of success p.

$= 0$ with probability of failure q

such that, $p+q = 1$

Problem:

A python program is said to have 5% bugs. A sample of 10lines were inspected what is the probability of getting 3 bugs in the sample.

$p = 5\% = 0.05$ $\quad q = 1 - 0.05 = 0.95$

$n = 10$ $\quad x = 3$

$$P(x) = {}^{10}C_3 (0.05)^3 (0.95)^{10-3}$$

$$P(3) = {}^{10}C_3 (0.05)^3 (0.95)^7$$

$$= 0.010347$$

2. A bolt manufacturing company claims that there will be 3% defectives in their total output. an inspection committee inspected an sample of 5 bolts what is the prob that there will be 3 defectives in sample

$p = 0.03 \qquad q = 0.97$

$n = 5 \qquad x = 3$

$$P(3) = {}^5C_3 (0.03)^3 (0.97)^{5-3}$$

$$= {}^5C_3 (0.03)^3 (0.97)^2$$

$$= 2.54043 \times 10^{-4}$$

```
from scipy.stats import binom
import matplotlib.pyplot as plt
n = 5
p = 0.03
x = list(range(n+1))
dist = [binom.pmf(r, n, p) for r in x]
plt.plot(x, dist)
plt.show()
```

plt.bar(x, dist) → bar graph.

Poisson distribution:

A discrete R.V 'X' is said to have poisson distribution if its pmf is given by

$$P(x) = \frac{e^{-\mu}\mu^x}{x!}, \quad x = 0,1,2,\ldots\infty, \quad \mu > 0$$

The mean and variance of dist. are equal and equal to μ

Example:

In a bustest junction the mean no. of accidents is '2' what is the probability that there will be 4 accidents on a particular day. Since the no. of accidents is a poisson variate.

no. of accidents – poisson variate

$\therefore$ X is the no. of accidents.

$$P(x) = \frac{e^{-2}2^x}{x!}$$

$$P(4) = \frac{e^{-2}2^4}{4!} = 0.09.$$

A newly published book have 500 pages and it is found to be 2 typographical errors per page on an average. obtain the no. of pages with 0, 1, 2, 3, 4 errors.

The no. of errors in a page is a poisson variate

and given as average 2 errors

$\mu = 2$

The poisson probability mass function

$$P(x) = \frac{e^{-2}2^x}{x!}$$

no. of pages with 0 errors = $P(0).500 = 67.66 \simeq 68$

" " " 1 " = $P(1).500 = 135.3 \simeq 135$

" " " 2 " = $P(2).500 = 135.5 \simeq 135$

" " " 3 " = $P(3).500 = 90.22 \simeq 90$

" " " 4 " = $P(4).500 = 45.1 \simeq 45$

```
from scipy.stats import poisson
import matplotlib.pyplot as plt
x = poisson.rvs(mu=2, size 10)
plt.hist(x, density= True ,edgecolor = 'black')
plt.show( )
```

Continous Random Variable :-

A R.V which assumes values btw certain limits

Ex :- height and weight

* Normal distribution :-

A continous R.V is said to have normal distribution if its probability density function is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \begin{matrix} -\infty < x < \infty \\ -\infty < \mu < \infty \\ \sigma > 0 \end{matrix}$$

where μ & σ are the arithematic mean & standard deviation of the distribution.

$$\text{if } Z = \frac{x-\mu}{\sigma}$$

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \quad -\infty < z < \infty$$

where z is known as standard normal variate.

The normal distribution having the following characterstics :-

1. The curve is bell shape and symmetrical about the line $x = \mu$.
2. Mean, Median & Mode of the distribution coincide
3. As x increases numerically, $f(x)$ decreases rapidly, The max. probability occuring at the pant $x = \mu$.
4. $\beta_1 = 0$ and $\beta_2 = 3$ (measure of kurtosis)
(skewness parameter)
5. Since $f(x)$ being the probability can never be $-ve$. No portion of the curve lies below the x-axis.
6. Linear combination of independent normal variates is also

a normal variate.

7. X-axis is an asymptote to the curve (never touches x axis)

8. The points of inflection of the curve are given by,

$x = \mu \pm \sigma$ (third derivative)

In kurtosis we have 3 types of curves

i. Leptokurtic : The curve which is peaked.

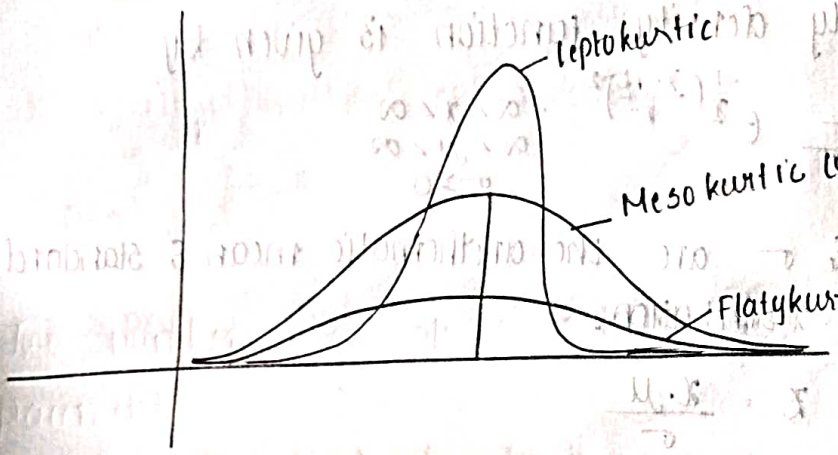
ii. Platykurtic : The curve which is almost flat

iii. Mesokurtic : curve which is neither peak nor flat.

or,

Normal curve

$\beta_2 = 3$ (always)



It seems that the x-axis is touching the curve but it will never touch the x-axis. Therefore, x-axis is an asymptote to the normal probability curve.

- From the area property of the normal probability curve

$P(|Z| \leq 3) = 0.9973$.

Z always lies between -3 and 3.

Ex:- The mean and standard deviation of a normal distribution are 12 & 3 respectively. Find the prob. of

$N(\mu, \sigma^2)$

i. $x \geq 10$

ii. $x \geq 15$

iii.

mean of $X = 12$

S.D of $X = 3$.

$N(\mu, \sigma^2)$ $\quad X \sim N(12, 3^2)$

$$Z = \frac{x - \mu}{\sigma}$$

$\therefore P(X \geq 10) \Rightarrow$

$$Z = \frac{10 - 12}{3} = \frac{-2}{3} = -0.666$$

Exponential distribution:-

A continous R.V is said to have exponential distribution its pdf $f(x) = \mu . e^{-\mu x}$ $\quad x > 0$, $\mu > 0$

$$\text{mean} = \frac{1}{\mu}$$

$$\text{variance} = \frac{1}{\mu^2}$$

Testing of Hypothesis:-

Since the population is very large to study the characterstics of population we take a sample from the population through this sample we try to estimate or test the population parameters.

Parameter:-

A population const. is known as parameter.

Ex: Population arthimatic mean of a characterstic which can be measured.

Statistic:

A function of sample of the observations is known as statistics.

ex:- Sample mean, Sample Variance. Sample standard deviation.

If there is a population of size 'N' and if we want to draw sample of size 'n' we can have $N_{C_n} = K$ samples, K samples can be drawn. For each sample, if we calculated sample mean then it is called a sampling dist. of the statistic ($\bar{x}$) that is given in following table.

Sample no.	Sample mean
1	$\bar{x}_1$
2	$\bar{x}_2$
⋮	⋮
K	$\bar{x}_k$

Ex:- Two numbers are to be selected from 4 without replacement

(4, 5, 6, 7)

$${}^4C_2 = \frac{4 \cdot 3}{1 \cdot 2} = 6.$$

Sample no.	Sample observations	Sample mean ($\bar{x}$)
1	(4, 5)	4.5
2	(4, 6)	5
3	(4, 7)	5.5
4	(5, 6)	5.5
5	(5, 7)	6
6	(6, 7)	6.5

Standard deviation of sample means gives Standard error.

Standard error of the stample mean $\bar{x}$ → S.E($\bar{x}$).

~~A population consists of~~

The above table represents the sampling dist. of statistic, sample mean ($\bar{x}$)

Standard Error :-

The S.D of the sampling dist. of the statistic is known as standard error. It is denoted by S.E (Statistic)

$$\bar{x} = \frac{1.5+5+5.5+5.5+6+6.5}{6} = 5.5$$

$$S.E = \sqrt{\frac{1}{n}\Sigma(x_i - \bar{x})^2}$$

$$= \sqrt{\frac{1}{6}(1+0.25+0+0+0.25+1)}$$

$$= \sqrt{0.41666}.$$

$$= 0.645$$

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
1.5	-1	1
5	-0.5	0.25
5.5	0	0
5.5	0	0
6	0.5	0.25
6.5	1	1
$\bar{x} = 5.5$		2.5

∴ Standard Error = 0.645