a) Statistics the exploration and analysis of data by Roxy and Jay Derore

b) Statistics by Schum series.

Course Content

- Probability
- Space Theory
- Conditional Probability and it's dependences
- Random Variables
- Discrete and Continous Distribution
- Mean and Variance
- Binomial, Posson, Hypergeometry, Exponential and Normal Distribution and their characteristics
- Centre Limit Theory
- Elementary Sample Theory for Normal Population
- Statistical Inference Mean, Population and Variance
- Simple Linear Regressions
- Chi square Test and Engineering Application.

Probability and Statistics

Probability is the measure of a likelyhood that a particular event will occur in any one trial or experiment carried out in a predescriped condition.

Notation;

The probability that a certain event A will occur is denoted by $P(A) = P(A$ and also equal to [Psuccess]

What occurs — Success

What does not Occur — Failure

Success or Failure

When an event occur in any one trial, it is called success but when it fails to occur, it is called a FAILURE;

In N, if any trials there are X successes there will also be [N−X] failures.

Probability of Success $\rightarrow \frac{X}{N} = P_{success}$

Probability of failure $\rightarrow \frac{N-X}{N} = P_{failure}$ or $1-X$

It's applied that $P(A) + P(\text{Not}A) = 1$.

i.e $\frac{X}{N} + \frac{(N - X)}{N} = 1$

or

$P(A) + P(\bar{A}) = 1$

Types of Probabilities:

i) Empirical or Experimental

ii) Classical or Theoretical

i) Experimental Probability: Is based on the previously known result, in the probability there is an expectation.

Expectation: Is the product of the no of trials and probability that event A will occur in any given trial.

ii) Classical Probabilities: Based on theoretical number of ways in which it is possible for an event to occur.

Classical probability of an event, A occurring is defined by:

$P(A) = \frac{\text{No of ways in which A occurs}}{\text{Total number of all possible outcome.}}$

Probability of having a H in a dice or dice = 1/6

Mutually Exclusive and Mutually Inexclusive Events

Mutually Exclusive events are events which cannot occur together.

Mutually Inexclusive: are events that occurs simultaneously or when there is a pair of event.

Additional Law of Probability

For a given event, A and B are mutually exclusive if A and B are mutually exclusive it's probability will be either A or B but not both i.e.

$P(A \text{ or } B) = P(A) + P(B)$

OR = Addition

BUT = Multiplication

Independent Event and Dependent Event

Events are Independent when the occurence of one event does not affect the probability of the occurence of the second / another event

Events are dependent when the occurance of one event affects the probability of the occurence of the other event

- Independent: It means it can be replaced.
- Dependent: It cannot be replaced

e.g Do this at home

Conditional Probability

Conditional probability of an event B occuring when given an event A has already taken place vis-a-vis Incase of an event, B stated above the probability will be given as $P\left(\frac{B}{A}\right)$, i.e calculate the probability of the event separately then divide it and it can occur vis-a-vis

2nd class

Binomial Distribution

Binomial Formular $\binom{n}{x} p^{x} (q)^{n-x}$

$b(x; n, p) = {}^{n}C_{x} p^{x} * (1-p)^{n-x}$

Value;

b = binomial probability

x = total number of successes

P = probability of successes in an individual trial

n = no of trials.

$${}^{n}C_{x} = \frac{n!}{x!(n-x)!}$$

Example: A coin is tossed 10 times what is the probability of getting six heads?

Sol

$P = 0.5$; $q = 1 - 0.5 = 0.5$; $n = 10$; $x = 6$

$P_6 = {}^{n}C_{x} * p^{x} * [1-p]^{n-x}$

$= {}^{10}C_6 * (0.5)^6 * [1-0.5]^{10-6}$

$= \frac{10!}{6![10-6]!} * 0.5^6 * 0.5^4$

$= 0.205$

Properties of Binomial

mean = np

Variance = npq

Standard deviation = $\sqrt{npq}$

q – probability of failure

p – probability of success

n – no of trials.

$mn = np = 10 \times 0.5 = 5$

variance $= npq = 10 \times 0.5 \times 0.5 = 2.5$

$SD = \sqrt{variance} = 1.58$.

Poisson Equation.

This occur when we have a rare event i.e when the probability of success is very low.

The poisson probability is

$$P(x;\mu) = \frac{[e^{-\mu}][\mu^{x}]}{x!}$$

where

x = actual no of success that results from experiment.

$e = 2.71828$

μ = mean of population

~~Properties~~

$\bar{x}$ = Mean of sample

s = Standard deviation of sample

σ = standard deviation of population

Properties

Mean $= \mu$

Variance $= \mu$

$SD = \sqrt{\mu}$

Ex2: Average no of hoes sold by OmoAkin's Company is 2 hoes per day what is the probability that exactly 3 hoes will be sold tomorrow.

Sol

$\mu = 2$ [Since 2 hoes are sold per day]

$x = 3$

$e = 2.71828$

$$P[x;\mu] = \frac{[e^{-\mu}][\mu^{x}]}{x!}$$

$$= \frac{[2.71828^{-2}][2^{3}]}{3!}$$

$= 0.180$

Hypergeometric Distribution

When an e is taken from a sample that is drawn from a population hypergeometic occurs.

If a population of size "N" contains K items of successes, the failure will be "N−K", then the probability of the hypergemetric random variable x and the no of success in a random sample of size n is

$$P(x=k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

$$= \frac{[{}^{K}C_{k}][{}^{N-K}C_{n-k}]}{{}^{N}C_{n}}$$

Where: K = no of success in population

k = no of observed successes

N = Population size

n = no of draws

Ex3: A deck of cards contains 20 cards 6 red cards and 14 black cards. 5 cards are drawn randomly without replacement. What is the probabilities that exactly 4 red cards are drawn

Sol

Exactly = 4

$N=20$ $K=6$ [Dealing with red]

$n=5$ $k=4$

$N-K = 20-6 = 14$

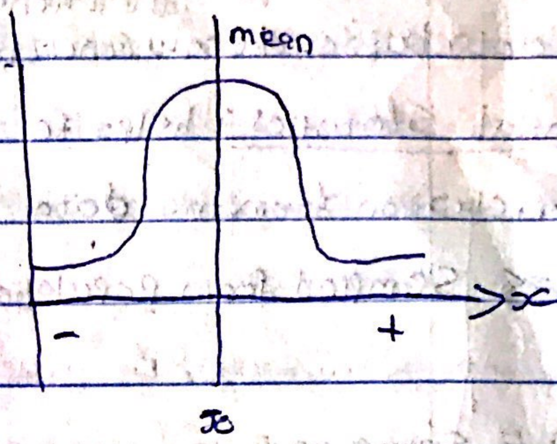
$n-k = 5-4 = 1$

$$P(x=k) = \frac{\binom{6}{4}\binom{14}{1}}{\binom{20}{5}}$$

$$= \frac{[{}^{6}C_{4}][{}^{14}C_{1}]}{[{}^{20}C_{5}]} = \frac{15 \times 14}{15504}$$

$= 0.0135$

Normal Distribution



$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Z = size

$\bar{x}$ = Mean of sample

μ = Mean of population

n = size of sample

σ = standard deviation of population

3rd C(s)

Statistical Inference

Types of Statistics

a) Descriptive Statistics

b) Inferential Statistics

Descriptive Statistics:- helps to realize relationship and patterns using graphs and curves between various data but do not draw conclusion beyond data at hand.

B) Inferential Statistics :- helps to draw conclusion from the data which was sampled from population

Random Sampling

It's a situation when every member of the population is given an equal chance to be selected at a sample e.g to determine the quality of palm-kernels for sale - they are samples taken without bias. Most theories of applied statistics are based on random sampling.

Disadvantages

- Samples collected may not be a good representation of a population
- In agriculture, the same treatment may be assigned to adjacent plot of land by chance.

Stratified Sampling

Population is first divided into two or more groups (strata); from here random sampling is then conducted e.g interview of people in a town to get their opinions. Selected units are usually more represented than that of random sampling.

Cluster Sampling

Similar to stratified sampling but in this case units of population exists already in particular groups.

Advantages

- Sample are usually representative of population

Disadvantages: It may be difficult to locate cluster boundaries.

Systematic Sampling

It's one in which elements are selected from a determined interval from a population frame.

¶. To determine the extent of certainty of how well a sample is taken and how the population is represented. It's expected that a sample must be a good representation of a population from which it's drawn. A sampling error occurs because there's no how sample could be same and perfectly represent a population.

- Sampling Distribution

Linear Regression

This is the mathematical tool use for division of a independent variable and other hand, it is used to test the strength of dependent variable.

It is a mathematical model use in prediction of variable x and also used to obtain the strength of independent variable y

For 2 constant

$y = a + bx$ — (i)

$\Sigma y = an + b\Sigma x$ — (ii)

$\Sigma xy = a\Sigma x + b\Sigma x^2$ — (iii)

a and b are constants.

Values for x and y will be given on the table.

For 3 variables

$y = a + bx_1 + cx_2$ — (i)

$\Sigma y = an + b\Sigma x_1 + c\Sigma x_2$ — (ii)

$\Sigma x_1 y = a\Sigma x_1 + b\Sigma x_1^2 + c\Sigma x_2$ — (iii)

$\Sigma x_2 y = a\Sigma x_2 + b\Sigma x_1 x_2 + c\Sigma x_2^2$ — (iv)

a, b and c are constant.

n = no of variables of sample.

Example 1

n	x	y	$\sum xy = x \cdot y$	x^2
1	4	5	20	16
2	6	4	24	36
3	7	5	35	49
4	8	6	48	64
$\sum$	25	20	127	165

$\sum y = a\cancel{n} + b\sum x$

$20 = 4a + 25b$ —— (i)

$\sum xy = a\sum x + bx^2$

$127 = 25a + 165b$ —— (ii)

$a = \frac{25}{7}$ $b = \frac{8}{35}$

$y = \frac{25}{9} + \frac{8}{35}x$

From $y = a + bx$

Analysis of Variance

Dispersion from beneath

$\sum = \frac{\frac{\bar{x} - \cancel{x}}{s}}{\sqrt{n}}$

$\bar{x}$ = mean sample

x = mean of population

= SD of sample n = sample size.

Degree of freedom (DF) $= n-1$

$SD = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}$

F ratio $= \frac{\text{Larger Variance}}{\text{Smaller Variance}}$

Chi-Square Test (x^2)

$x^2 = \frac{\sum[0-E]^2}{E}$.

O = Observation Value

E = Expected Value.

Degree of freedom = no of observation -1

$\therefore DF = n-1$

	O	E	x^2
1	6	5	$\frac{(6-5)^2}{5} = 1/5$
2	8	6	$\frac{(8-6)^2}{6} = 2/3$
3	10	6	$\frac{(10-6)^2}{6} = 16/6$
			$1/5 + 2/3 + 16/6 = 3.54$

$x^2 = 3.54$ [Calculated Value]

$df = n-1$

$n = 3$

$df = 3-1 = 2$

Note:- If the tabulated value is greater than the calculated values it will be accepted, there is no significance. When calculated value is less than tabulated value, it is accepted when otherwise, it is not accepted.

Grouping Method for χ^2

Ex

Observed Value

	S orange	M orange	L orange	Total
Red	5	9	10	24
Green	4	8	9	21
Brown	3	7	4	14
Total	12	24	23	59

Expected value

	S orange	M orange	L orange
Red	4.881	9.763	9.356
Green	4.271	8.542	8.186
Brown	2.847	5.695	5.458

$$\text{Expected Value} = \frac{\text{Row Total} \times \text{Column T}}{\text{Total of Total}}$$

For Red C_1, C_2 and C_3

$$C_1 = \frac{24 \times 12}{59} = 4.881$$

$$C_2 = \frac{24 \times 24}{59} = 9.763$$

$$C_3 = \frac{24 \times 23}{59} = 9.356$$

For ~~Red~~ Brown

$$C_1 = \frac{14 \times 12}{59} = 2.847$$

$$C_2 = \frac{14 \times 24}{59} = 5.695$$

$$C_3 = \frac{14 \times 23}{59} = 5.458$$

For Green

$$C_1 = \frac{21 \times 12}{59} = 4.271$$

$$C_2 = \frac{21 \times 24}{59} = 8.542$$

$$C_3 = \frac{21 \times 23}{59} = 8.186$$

Calculating the x^2 for each observation use O value and E value

For Red, we have

$\frac{(5-4.881)^2}{4.881} + \frac{(9-9.763)^2}{9.763} + \frac{[10-9.356]^2}{9.356}$

$x^2 = 0.107$

For greens

$\frac{(4-4.271)^2}{4.271} + \frac{(8-8.542)^2}{8.542} + \frac{[9-8.186]^2}{8.186}$

$x^2 = 0.133$

For Brown

$\frac{[3-2.847]^2}{2.847} + \frac{[7-5.695]^2}{5.695} + \frac{[4-5.458]^2}{5.458}$

$x^2 = 0.697$

x^2_{tag} = Red + Green + Brown

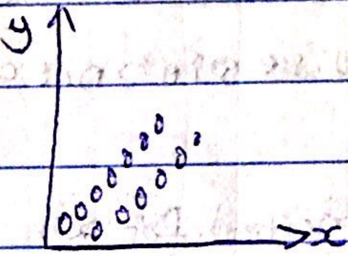
$= 0.107 + 0.133 + 0.697$

$x^2 = 0.937$

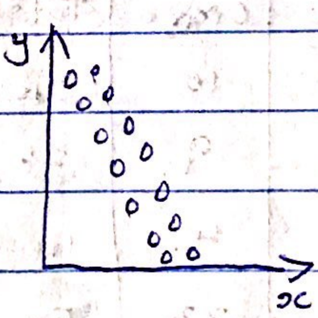
$Df = [3-1][3-1] = 4$ [Row and C]

Measure of Association / Relationship

The relationship between two variables is called association.



+ve relationship observation.



-ve Relationship

In probability, the valid no can only be 1 or -1.

Linear Correlations & Linear Regression

Karl pearson product-moment Correlation Coefficient

$$r = \frac{\text{Covariance}(x, y)}{\sigma_x \; \sigma_y} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

σ_x = SD of variable x

σ_y = SD of variable y.

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}}.$$

Sampling Distribution

The sampling distribution of a statistics is the probability distribution of that statistics

Statistics & Parameters

Statistics comes from a sample while parameter comes from a population.

Parameter is a measurement or quantity that describes the population.

Statistics is the measurement or quantity that describes the sample

E.g of Parameters is Population mean, variance, standard deviation, median, mode, etc.

E.g of statistics are sample mean, variance, standard deviation.

Continuation of Linear Correlation

Correlation is a measure of the amount of association existing between variable. It is used to determine the linear correlation coefficient.

$$r = \frac{\Sigma XY}{\sqrt{(\Sigma X^2)(\Sigma Y^2)}}$$

$x \rightarrow$ is the value of deviations of co-ordinates x from $\bar{x}$ [mean value]

$y \rightarrow$ The value of deviation of coordinates y from $\bar{y}$

$X = x - \bar{x}$

$Y = y - \bar{y}$

The determination of this give the value of r lying between

+1 and -1, +1 [Perfect direct Correlation]
-1 [Perfect Inver Correlation]
0 [no Correlation]

The smaller the value of r the Lesser the Correlation. [0.7-1].

To find $\bar{x}$,

$\{\bar{x}\} = \frac{\sum x}{n}$.

overall mean or normal mean.

Eg In an experiment to determine the relationship between force of a wire and the resulting extension the following data was obtain.

Force(N). 10 20 30 40 50

Extension 0.22 0.40 0.61 0.85 1.20

60 70

1.45 1.70.

Sol

x = Force, y = extension.

$X = x - \bar{X}$

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

x	y	X	Y	XY	X^2	Y^2
10	0.22	-30	0.699	20.97	900	0.489
20	0.40	-20	-0.519	10.38	400	0.269
30	0.61	-10	-0.309	3.09	100	0.095
40	0.85	0	-0.009	0	0	0.005
50	1.20	10	0.281	2.81	100	0.074
60	1.45	20	0.531	10.62	400	0.282
70	1.70	30	0.781	23.43	900	0.610

Here $X = [x - \bar{x})$

$\bar{x} = \frac{\sum x}{n} = \frac{10+20+30+\cdots+70}{7}$

for Let's say 10, $x = [10-40]$
$= -30$

do for all and same for y

$\therefore r = \frac{\sum XY}{\sqrt{\sum X^2 \cdot \sum Y^2}}$

$\sum XY = 71.30$, $\sum X^2 = 2800$,

$\sum Y^2 = 1.829$

$= \frac{71.30}{\sqrt{5121.2}}$

$r = 0.997$

shows a very good direct Correlation.

Scanned with CamScanner

Note:

r in the ranges 0.7 to 1 and -0.7 to -1 show that a fair amount of Correlation exists.