STA 202

COLLECTION OF DATA

Before any statistical work can be done figure or data must be collected, it's very important for us to collect the right and correct data because any mistakes, error or bias in data collection will affect the result, Decision taken and conclusion a research work

DEMOGRAPHY.

Is a branch of the social sciences that deals with study of human population, their structures change (through birth, death and migration) and they relationship with natural environment.

DEMOGRAPHIC INDICATORS.

these include population size, growth rate, birth rate death rate, total fertility rate, type expectancy demographic changes affect all areas of human activities that is economic, social, cultural and political.

Definition of Some Demographic terminologies.

(ii) Crude birth rate: is the annual number of type birth per 1000 people.

(iii) Crude death rate: is the annual number of death per 1000 people

(iii) Fertility rate: is the annual number of type birth per 1000 women of child bearing age. 15 – 49 years.

(4) Life expectancy: Is the number of years an individual at a given age is expected to live at present mortality level.

(6) Total fertility rate: Is the number of live birth of women completing her reproductive life if her child bearing at each age reflect current specific fertility rate.

(7) The replacement level fertility: Is the average number of children a woman must have in order to replace herself with daughter in the next gener

(8) The growth reproduction rate: Is the number of daughters who will be born to a woman completin her reproductive life at current age specific fertility rate.

(9) The next reproduction rate: Is the expected number of daughters per new born prospective mother who may or may not survive to and through the age of ch bearing.

(10) Stable population: Is a population that has a constant crude birth rate and death rate for a pe period of time and the percentage of people in every age limit remains constant.

formulae:

(i) Crude birth rate: $\frac{\text{Total number of birth}}{\text{Total number of population}} \times 1000$ (per 1000).

(ii) Crude death rate: $\frac{\text{Total number death}}{\text{Estimated population}} \times 1000$ (per 1000).

(iii) Age specific death rate (ASDR) = $\frac{\text{death rate}}{\text{Total population}} \times 1000$ (per 1000).

P.T.O

(vi) Age specific birth rate: (ASBR).

$$= \frac{\text{Birth rate}}{\text{Total fertile female population (per 1000)}} \times 1000$$

(v) Standard death rate: (SDR)

$$= \frac{\text{Total expected death}}{\text{Total standard population (per 1000)}} \times 1000$$

(iv) Expected death = Age specific death rate × standard p

1000

(per 1000).

(viii) Standardized birth rate (SBR)

$$SBR = \frac{\text{Total expected birth}}{\text{Total standard fertile female population (per 1000)}} \times 1000$$

(vii) Expected birth = Age specific birth rate × standard fe

female popula

1000.

(per 1000).

Example: The data below shows the age group, estimate population recorded death and standard popula tion as recorded by a demographic.

Age group	Estimated population in million	Recorded death in thousands	standard population in millions.
0 — 20	25	360	24
21 — 40	38	250	32
41 — 60	26	150	16
61 and above	8	650	8
sum	100	1446	80

Topic: Final.

(a) Crude death rate
(b) Age specific death rate
(c) Standardized death rate

Solution.

(a) Crude death rate (CDR) =

$$\frac{\text{Total number of death}}{\text{Estimated population}} \times 1000$$

$$= \frac{144\,000}{100,000,000} \times 1000$$

$$= 14.4 \text{ per } 1000$$

(b) Age-specific death rate.

Age group	Estimated population	Record death	Age specific death rate
0 – 20	25	360	$\frac{360}{28,000,000} = 12.857$
21 – 40	58	250	$\frac{250,000,000}{28,000,000} \times 1000 = 6.579$
41 – 60	26	150	$\frac{150,000}{26,000,000} \times 1000 = 6.923$
61 and above	8	650	$\frac{650,000}{8,000,000} \times 1000 = 81.2$
total	100	1440	

(c)

Age group	Age-specific death rate	Standardized	Expected deaths
0 – 20	12.857	24	308,568
21 – 40	6.579	32	210,528
41 – 60	6.923	16	110,768
61 and above	81.25	8	650,00
Sum		80	1,279,864

$$SDR = \frac{1,279,864}{80,000,000} \times 1000$$

$$= 15.9 \text{ per } 1000$$

$$= 16 \text{ per } 1000$$

308568
210528
110768
650000

Exercise:

(1) The data below shows the age group, estimated population, estimated fertile female population, birth and standard fertile female population as recorded by a statistician in a rural community in Nigeria.

Age group	estimated P (millions)	est fer f p (thousands)	Res f f b p	Standard fer fe p (millio
0 — 14	300	—	—	—
15 — 24	350	195	38	52
25 — 34	540	260	96	198
35 — 44	100	85	25	50
45 and above	230	—	—	—
Sum				

Require.

(i) Calculate the Crude birth rate

(ii) Compute the age specific and standardized birth

(2) The table below gives the population and the number of death in two towns A and B.

Age-group	Population (Thousands)	Number of death (1000) A	Population (1000) B	Death (1000) B	Standard Population (1000)
0 — 2	2,300	150	3000	300	100
2 — 10	1,400	30	12,000	60	150
10 — 20	11,000	40	14,000	100	350
20 — 30	30,000	200	18,000	90	250
30 and above	6,300	500	5,000	450	200

Dated: 1/1 1/20....

Task: Calculate:
(i) the Crude death rate for the two towns.
(ii) the death rate for each groups for each town.
(iii) the standardize age group for each towns.

REGRESSION: is defined as the amount of change in one variable associated with a unit change in another variable. To be more specific for a variable Y which depend on another variable X the Independent variable and equation relating Y on X is called a Regression equation.

Regression equation is there for a statistical tool a technic use to predict or explain the relationship between Y and X. As plain from the definition presented above the two types of variable are involve in regression analysis.
One variable changes in values and therefore induce a change in a unit value of another variables i.e a cause and effect relationship. X is called the Independent variables explanatory variables, exogenous variables control or Regressor. Y is called the dependent variables, Respond, Endogenous, Uncontrol or the Regress variable.

Regression can be simple or multiple linear or non-linear we shall discuss only the simple Regression that is only two variables namely.
The Simple regression model

$$Y_i = b_0 + b_1 X_1 + e_i$$

where Y = dependent variable.
X_i = the Independent variables
b_0 = constant / intercept

b_1 = the regression coefficient

e_1 = error terms.

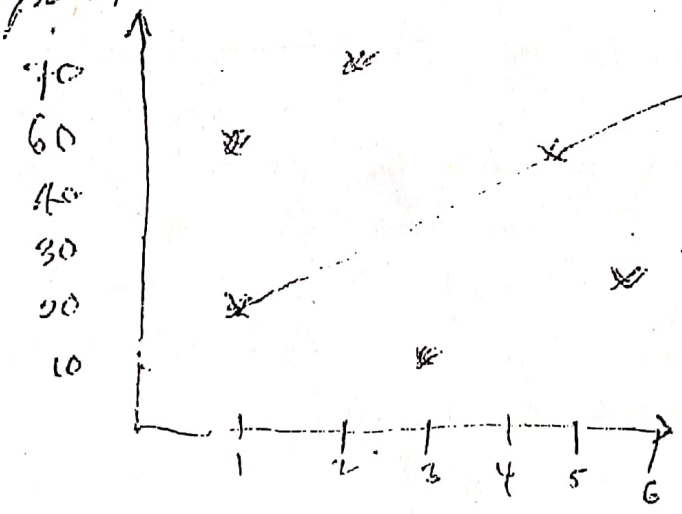
$$b_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

$$b_0 = \bar{Y} - b_1\bar{X}$$

where $\bar{Y} = \frac{\sum Y}{N}$ and $\bar{X}_i = \frac{\sum x}{N}$.

BIVARIATE DATA: A bivariate data (x, y) is a set of values which appears in pairs it's a data in which the values of y depend on the value of x.

SCATTERED DIAGRAM: when the points (x, y) of a bivariate data are plotted in a rectangular coordinate, the resulting diagram is called a scatter diagram



Simple CORRELATION.

Definition: is a statistical techniques use to measure the degree, strength and direction of the relationship or association between two or more variables.

The main purpose of correlation is to find out how

Date: 1/1 1/20....

Aim:

Strong or Weak a relationship is or to find out if a relationship is positive or negative, it's denoted by γ or r.

TYPES of CORRELATION.

- Pearson's product moment correlation co-efficient (PPMCC).
- Spearman's Rank correlation coefficient.
- Kendall Tau correlation coefficient.
- Partial and Multiple correlation formulae.

$$\gamma_s = \frac{\Sigma xy}{\sqrt{(\Sigma x^2)(\Sigma y^2)}}$$

where $X = x - \bar{x}$ or $\gamma = \frac{\Sigma xy}{\sqrt{\Sigma x \Sigma y}}$

$$r_{xy} = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{(N\Sigma x^2 - (\Sigma x)^2)(N\Sigma y^2 - (\Sigma y)^2}}$$

If $\gamma = 0.1$ — 0.49, a Week positive correlation

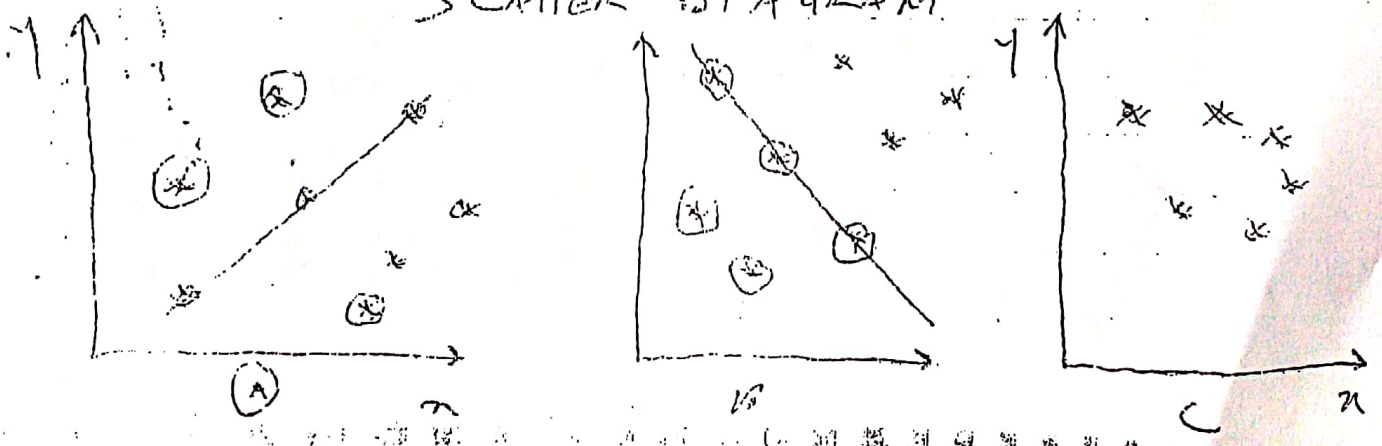
If $\gamma = 0.5$ — 0.59, a Moderate or average Correlation.

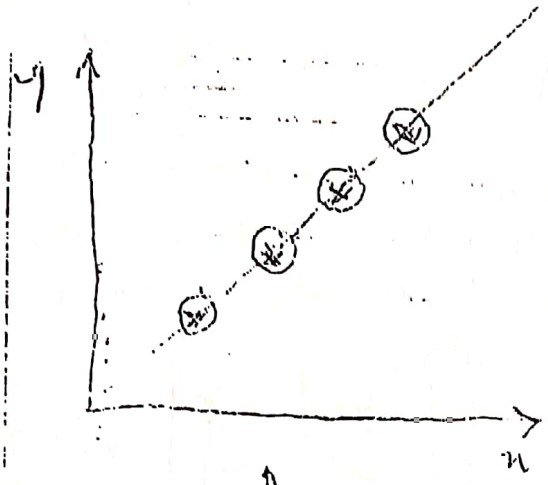
If $\gamma = 0.6$ — 0.99, a Strong positive correlation

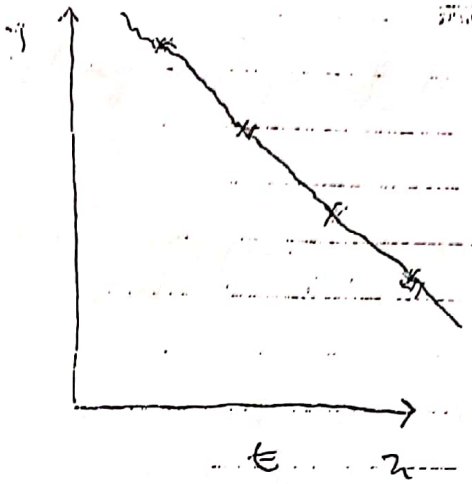
If $\gamma = 0$ no correlation,

If 1, there exis a perfect positive correlation and Vice Versa is negative Correlation.

SCATTER DIAGRAM







↑ = positive correlation.
↓ = Negative correlation
C = No correlation
P = Perfect positive
E = Perfect negative.

Spearman's Rank correlation coefficient is a non-parametric equivalent of the pearson's product moment correlation coefficient. It's used for data cannot be quantified, the data are ranked first computation.

The Spearman's Rank formulae is given as

$$\text{Rank} = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

Where, d_i = Rank X_i − Rank y_i

$d_i = U_i - V_i$

x y U_i V_i $V - V_i$ d_i^2

Example 2:

x	15	12	18	21	16
y	12	14	16	18	18

(i) Plot a scatterred diagram for the information above.

(ii) Regress y on x

(iii) Compute the correlation co-efficient of x and y and interpret your result.

Solution

X	Y	XY	X^2	Y^2
15	12	180	225	144
12	14	168	144	196
18	16	288	324	256
21	15	315	441	324
16	18	288	256	324

$\Sigma x = 82$, $\Sigma y = 78$

$\Sigma xy = 1302$

$\Sigma x^2 = 1590$

$\Sigma y^2 = 1244$

(ii) $Y = b_0 + b_1 x$

$$b_1 = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{N\Sigma x^2 - (\Sigma x)^2}$$

$$= \frac{5(1302) - 82(78)}{5(1390) - (82)^2}$$

$b_1 = 0.504$

$b_0 = \bar{Y} - b_1\bar{x}$

$\bar{Y} = \frac{\Sigma y}{N} = \frac{78}{5} = 15.6$

$\bar{x} = \frac{\Sigma x}{N} = \frac{82}{5} = 16.4$

$b_0 = 15.6 - (0.504)(16.4)$

$b_0 = 7.32$

$Y = 7.32 + 0.504x$

(iii) $r_{xy} = \frac{(5)(1302) - (82)(78)}{\sqrt{(5)(1390) - (82)^2}\sqrt{(5)(1244) - (78)^2}}$

$$1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

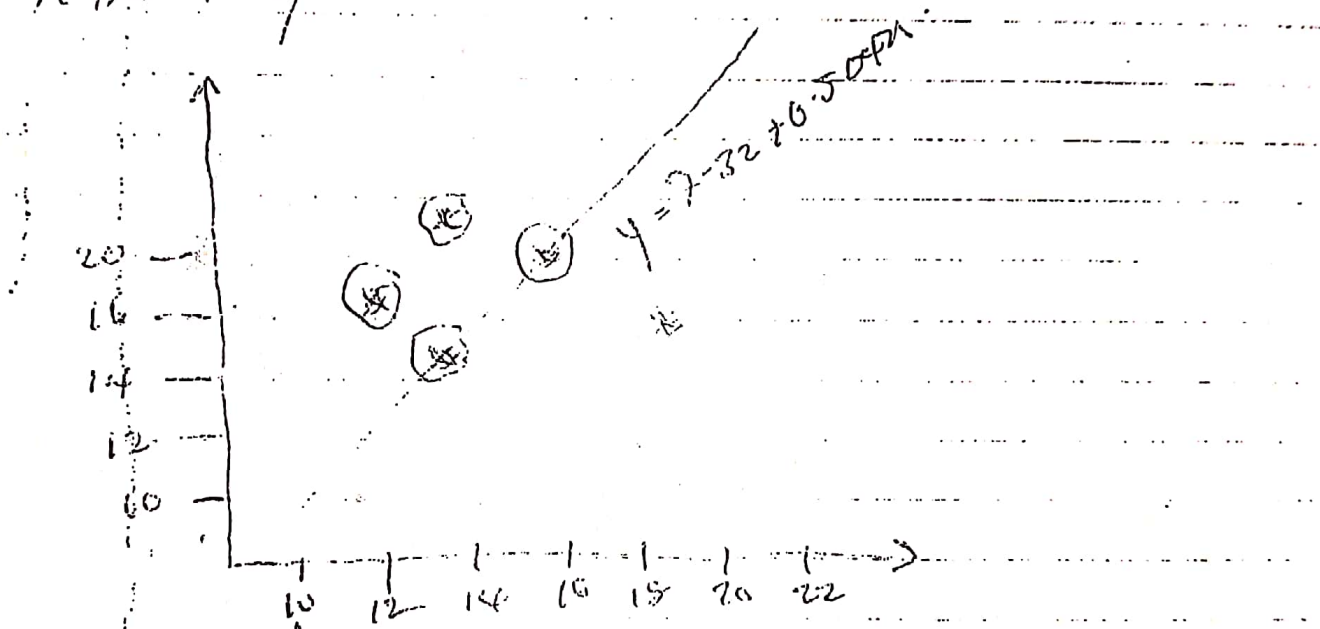
Dated ... 1/20....

$$= \frac{6510 - 6396}{\sqrt{(6950 - 6724)(6120 - 6084)}}$$

$$r_{avg} = \frac{114}{\sqrt{(226)(36)}} = \frac{114}{\sqrt{8136}}$$

$$r_{avg} = \frac{114}{174.90} = 0.6511$$

There exist a strong positive correlation between y and x.



Exercise.

1) The data below shows the number of road accident and the resulting death in nigeria between 1990 and 1997

Year	1990	1991	1992	1993	1994	1995	1996
Number of accident	17	18	23	25	27	33	54
Death	3	4	4	5	5	6	8

1997	1998
34	36
7	9

Exponential Smoothing is a rule of thumb technique for smoothing time series data using the exponential window function....

Exponential Smoothing is one of many window functions commonly applied to smooth data in signal processing as low pass filters to remove high frequency no

(1) Draw the scattered diagram
(2) Fixed the regression equation Cue of Y on X
(3) Predict the number of death in 1997 if 2750 are
enrolled occured.

(iv) The following is a set of data of the Federal government area expenditure on agric and health sector in the period of 2001 to 2009.

Period	2001	2002	2003	2004	2005	2006	2007	08	09
Agric Ex	4.4	4.5	12.4	13.4	24.6	38.8	18.5	42.0	20.5
Health Ex	12.0	14.1	19.4	18.3	29.1	62.1	83.5	85	95.5

(i) Estimate the best line of fit between Y and X (ii) What is the degree of association between X and Y using Pearson and Spearmen's correlation coefficient.

TIME SERIES

Is defined as the collection of observation repeated over a long period of time.

Time series is study for the interest, for the future occurance it also forecast future level economic activities by studying the behavior of data in the past.

COMPONENT OF TIME SERIES.

Time series is noted for it's variation from time to time, this variation can be attributed to climatic, social, economics or accidental factors this factors can be describe as the future or four component of time series describe bel

(1) The Trend is a part which time series appears in column over a long period of time it's upward and downward movement in time series.

(2) Seasonal Variation: This is brought by climatic social factors, Seasonal Variation appears with so irregularities in the same season of a year, it's graph reflect upward and downward movement of time series graph.

(3) Cyclical Variation: This time series exist over a fixed period due to some principal cause

(4) Irregular Variation: after trend and cyclical variation have been removed from the set of data we are left with series of residual which may or not random they are sporadic or unpredictable. Like war, fire disaster, flood and election.

MATHEMATICAL REPRESENTATION OF TIME SERIES

The following notation are needed for mathematical combination of time series.

(a) The time series is denoted by Capital Y.
(b) The trend component is denoted by capital T.
(c) The Cyclical Variation is denoted by C.
(d) The Seasonal Variation is denoted by capital S
(e) The irregular Variation is denoted by I.

There are two mathematical model of time series, the Additive and Multiplicative model

(a) Additive model is defined by $Y = T + S + C + I$
(b) Multiplicative models are.

$$Y = T \times S \times C \times I = TSCI.$$

Exercise.

The Data below shows the quarterly production of cassava in Nigeria between 2010 and 2013.

Quarters

Year	1	2	3	4
2010	5.8	19.2	1.8	21.1
2011	9.5	22.2	2.0	34.8
2012	20.6	22.5	1.0	40.1
2013	28.7	21.5	1.7	25.5

Compute the trend value using

(i) the least square method.

(ii) the moving average

(iii) find the seasonal index.

The Seasonal index = $\frac{\text{Total Value of Quarter}}{\text{Total average}}$

Example

Year	1	2	3	4
2013	13	20	17	25
2014	15	15	16	23
2015	12	22	14	21
2016	11	17	10	24
Total	51	74	57	93

Quarter average = $\frac{51 + 74 + 57 + 93}{4}$

$= \frac{275}{4} = 70$

Seasonal index for $Q_1 = \frac{51}{70} = 0.73$.

Q_2 ... Q_4 ... $\frac{74}{70} = 1.13$.

Date: / /20...

$Q_3 = \frac{51}{30} = 1.13$

$Q_4 = \frac{43}{30} = 1.43$

Test of Hypothesis.

(1) Statistical hypothesis: is an assertion about the distribution of one or more random variables. If the statistical hypothesis completely specified the distribution it is called a simple statistical hypothesis. If it doesn't it is called the composite statistical hypothesis. Example.

$H_0 : \mu = 75$

Vs

$H_1 : \mu \neq 75$

$H_1 : \mu < 75$

(2) A Test of statistical hypothesis is a rule which when experimental simple value have been obtained lead to a decision to accept or reject the hypothesis under consideration.

There are five steps to be taken when one is interesting in testing the hypothesis there are:

(i) formulate the hypothesis

(ii) State the level of significant.

e.g. $\alpha = 0.05$ or $\alpha = 0.01$

$\alpha = 0.1$

Example: Test at the $\alpha = 0.01$ level Significance that the means of 51, 5 is less than 6.

$H_0 : \mu = 6$

Vs $H_1 : \mu < 6$.

of Mean 99% assurance level of the hypothesis you want to take

(iii) Compute the test statistic for the formulae.

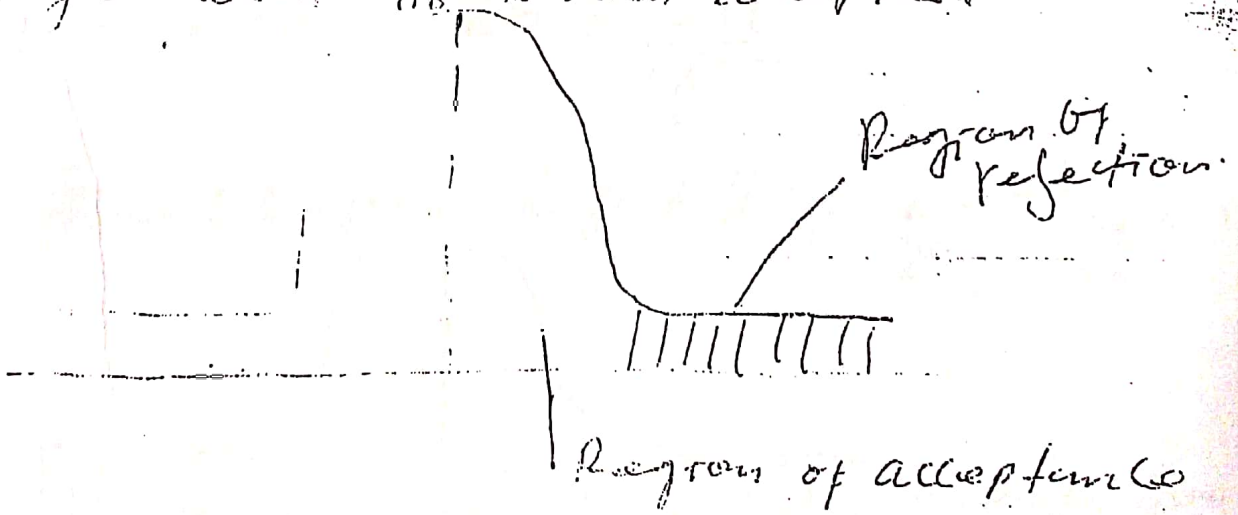
(iv) Compare the test statistic value with that in the critical region

(v) Arrived at a decision.

3. Level of significance: It's the quantity of risk of the type (1) error which we are ready to tolerate by making a decision of $\alpha = 0.05$, 0.01 etc

4. P-Value: This is another approach by which we can find the smallest alpha level at which H_0 is rejected. The criteria for using this is reject H_0 if P-value is less than α.

(5) CRITICAL Region: Is a test the area under the probability density curve is divided into two groups or region i.e the region of rejection and acceptance region. The region of rejection is the region where H_0 is rejected the region of acceptance is the region where H_0 is been accepted.



(6) The Tailed Test: The critical region is always on the tail of the distribution curve. It may be on both tailed or on one tailed depending upon the alternative hypothesis.

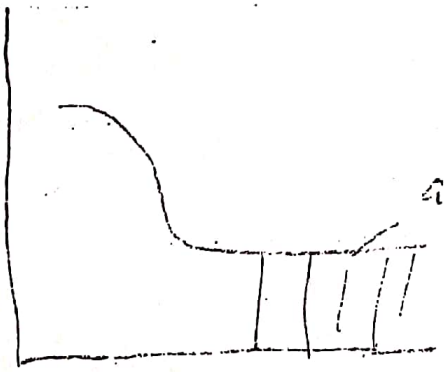
1. One - tailed test (to the right) hypothesis:

$H_0 : \mu = 3$

$H_1 : \mu > 3$

Test statistic $Z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

Critical region: Reject H_0, if $Z > Z_{\alpha, n-1}$



a Rejection region

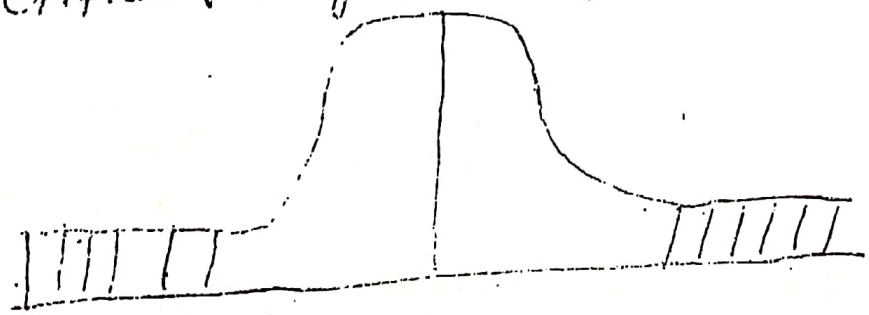
2. One - tailed test (to the left) hypothesis:

$H_0 : \mu = 3$

$H_1 : \mu < 3$

Test statistics $Z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

Critical Region: Reject H_0 if $Z < -Z_{\alpha, n-1}$ or $Z_{\alpha/2, n-1}$



Z Rejection region

$-Z_\alpha$ Rejection region

Date: / / 1/20....

Region:

Probability of rejecting H_0 when it is false is denoted by $1 - \beta$.

Example (1):- The data below represent the scores obtained by 20 students in Sta202 and Csc202.

STA 202	67	62	52	54	51	61	54	59	57	55
CSC202	60	55	57	68	66	52	53	54	56	58

Assuming the data is normally distributed with means 60 and variance 25. Test at 0.05 level of Significant that the means scores in both exams is less than 60:

Solution

(1) Hypotheses

$H_0: \mu = 60$

$H_1: \mu < 60$

(2) $\alpha = 0.05$

(3) Test statistic

$$z = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/n}} \quad \text{or} \quad \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

(4) the means from both $\bar{x} = \frac{\sum x}{n} = \frac{1146}{20}$

$$\bar{x} = 57.3$$

$$z = \frac{57.3 - 60}{\sqrt{25}/\sqrt{20}} = \frac{-2.7}{1.12}$$

$$z_{cal} = -2.42$$

$$z_{0.05,19} = -1.73$$

(5) Decision: Since $z_{cal} = -2.42$ is less than -1.73 we reject the null hypothesis and accept the alternative hypothesis H_1:

The conclusion reached here is the mean is less than 60.

TESTING DIFFERENCE BETWEEN TWO MEANS (WHEN VARIANCE ARE UNKNOWN).

formulae:

1. $\bar{X}_1 = \frac{\sum x_1}{N_1}$, 2. $\bar{X}_2 = \frac{\sum x_2}{N_2}$

3. $S_1^2 = \delta_1^2 = \frac{N\sum x_1^2 - (\sum x_1)^2}{N_1 - 1}$

4. $S_2^2 = \delta_2^2 = \frac{N_2\sum x_2^2 - (\sum x_2)^2}{N_2 - 1}$

5. pooled Variance

$$S_p^2 = \frac{(N_1 - 1)S_1^2 + (N_2 - 1)S_2^2}{N_1 + N_2 - 2}$$

6. $$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_p^2}{N_1} + \frac{S_p^2}{N_2}}}$$

Example 2: The Information below represent the test scores obtained in sta202 and math206

9	18	15	3	16	21	13	7		
4	12	5	10	8	11	25	19	4	17

$\alpha = 0.01$ level of significant whether score of sta202 is equal to the means MTH206.

P.T.O

Solution

1. Hypothesis

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

2. $\alpha = 0.05$

3.

n_1	x_1	x_1^2	n_2	x_2	x_2^2
1	9	81	1	14	196
2	18	324	2	12	144
3	15	225	3	5	25
4	3	9	4	10	100
5	16	256	5	8	64
6	21	441	6	11	121
7	13	169	7	25	625
8	7	49	8	19	361
			9	4	16
			10	17	289

$n_1 = 8, \quad \Sigma x_1 = 102, \quad \Sigma x^2 = 1554$

$n_2 = 10, \quad \Sigma x_2 = 125, \quad \Sigma x_2^2 = 1991$

$$\bar{X}_1 = \frac{\Sigma x_1}{n_1} = \frac{102}{8} = 12.75$$

$$\bar{X}_2 = \frac{\Sigma x_2}{n_2} = \frac{125}{10} = 12.5$$

$$\int_1^2 = \frac{N_1 \Sigma x_1^2 - (\Sigma x_1)^2}{N_1 - 1} = \frac{8(1554) - (102)^2}{8-1}$$

$$\int_1^2 = \frac{2028}{7} = 289.7$$

Date: / /20....

Topic:

$$S_2^2 = \frac{N_2 \Sigma x_2^2 - (\Sigma x_2)^2}{N_2 - 1} = \frac{10(1941) - (125)^2}{10 - 1}$$

$$= \frac{3755}{9} = S_2^2 = 420.5$$

$$S_p^2 = \frac{(N_1 - 1)S_1^2 + (N_2 - 1)S_2^2}{N_1 + N_2 - 2} = \frac{7(289.7) + 9(420.5)}{16}$$

$$S_p^2 = 363.$$

$$t_{cal} = \frac{12.75 - 2.5}{\sqrt{\frac{363}{5} + \frac{363}{10}}}$$

$t_{cal} = 0.03.$

4. Decision Rule: Reject H_0, if $t_{cal} > t_{\frac{\alpha}{2}, n_1+n_2-2}$

$t_{0.05, 16} = 2.12.$

5. Since $T_{cal} = 0.03 < t_{tab} = 2.120$ We can not reject the null hypothesis.

6. The conclusion reached there is that the mean score in Sta202 is equal to mean score in MTH206.

Date: / / 20....

Topic:

MOST POWERFUL TEST

If a simple hypothesis is tested against a simple alternative and suppose we have a the same level of significant the Test will be smallest size type II error is the most powerful Test of the two that significant.

A Uniformly most powerful (UMP) is a hypothesis which has the greatest power that is less than or all the possible test of a given size α.

for example the NEYMAN PEARSON LEMMA the likely hood ratio test etc. are UMP. for testing simple hypotheses.

Best critical region: Consider the test of a simple Null hypothesis ($H_0: \theta = \theta_0$) versus a simple alternative $H_1: \theta = \theta_1$ let c be a critical region of size α. the best critical region is that region that has a greatest power among all the critical region of size α.

the Neyman pearson Lemma gives sufficient condition for a best critical region of size α

NEYMAN PEARSON LEMMA:

Let $x_1, x_2 \ldots x_n$ be a random sample of size n from the distribution with PDF $f(x; \theta)$ where θ_0 and θ_1 are two possible values of θ. let the joint PDF of $x_1, x_2 \ldots x_n$ be denoted by the likely hood function $L(\theta) = L(\theta; x_1, x_2 \ldots x_n) = f(x_1, \theta)$

$f(x, \theta) \ldots f(x_n, \theta)$ if and only if (IFF) there exist a positive constant k and critical region c of the sample space such that

Date: / / 1/20....

Topic:

MOST POWERFUL TEST.

If a simple hypothesis is tested against a simple alternative and suppose we have a the same level of significance the Test will be smaller size type II error is the most powerful Test of the two that significant.

A Uniformly most powerful (UMP) is a hypothesis which has the greatest power that is less than or all the possible test of a given size α.

for example the NEYMAN PEARSON LEMMA the likely hood ratio test etc. are UMP. for testing simple hypotheses.

Best critical region: Consider the test of a simple Null hypothesis ($H_0: \theta = \theta_0$) versus a simple alternative $H_1: \theta = \theta_1$ let c be a critical region of size α. the best critical region is that region that has a greatest power among all the critical region of size α.

the Neyman pearson Lemma gives sufficient condition for a best critical region of size α

NEYMAN PEARSON LEMMA.

Let $x_1, x_2 \ldots x_n$ be a random sample of size n from the distribution with PDF $f(x; \theta)$ where θ_0 and θ_1 are two possible values of θ. let the joint PDF of $x_1, x_2 \ldots x_n$ be denoted by the likely hood function $L(\theta) = L(\theta; x_1, x_2 \ldots x_n) = f(x_1, \theta)$

$f(x_1, \theta) \ldots f(x_n, \theta)$ if and only if (IFF) there exist a positive constant k and critical region c of the sample space such that

(i) $P(x_1, x_2 \cdots x_n) \in C ; \theta_0) = \alpha$

(ii) $\frac{L(\theta_0)}{L(\theta_1)} \leq k$ for $(x_1, x_2 \cdots x_n) \in C$

(iii) $\frac{L(\theta_0)}{L(\theta_1)} \geq k$ for $(x_1, x_2 \cdots x_n) \in C'$

Then C is a best critical region of size α for testing the simple hypothesis ($H_0 : \theta = \theta_0$) against an alternative simple hypothesis ($H_1 : \theta =$

Note that a best critical region can be obtained by using the Neyman pearson Lemma if and only if

$$\frac{L(\theta_0)}{L(\theta_1)} \leq k \in \bar{x} \geq C$$

or

$$\frac{L(\theta_1)}{L(\theta_0)} \geq k \in \bar{x} \leq C$$

example. Let $x_1, x_2 \cdots x_n$ be a random sample from a normal distribution with $N(\mu, 36)$. find the best critical region for testing the $H_0 : \mu = 50$

against

$H_1 : \mu = 55$

Solution.

$$f(x, \theta) = (2\pi\sigma^2)^{-n/2} \; e^{-\frac{1}{2\sigma^2}\sum(x-\mu)^2}$$

$$L(x; \theta) = (2\pi\sigma^2)^{-n/2} \; e^{-\frac{1}{2\sigma^2}\sum(x-\mu)^2}$$

$$\frac{L_0}{L_1} \leq k$$

Date: / / 1/20....

$$\frac{L(50)}{L(55)} = \frac{(72\pi)^{-n/2}}{(72\pi)^{-n/2}} \exp\left[-\frac{1}{72}\sum(x_i - 50)^2 \right.$$
$$\left. \exp\left(-\frac{1}{72}\right)\sum(x_i - 55)^2\right.$$

$$\frac{L(50)}{L(55)} = \exp\left(\frac{-1}{72}\right) \frac{\sum(x_i^2 - 100x_i + 2500)}{\sum(x_i^2 - 110x_i + 3025)}$$

$$= \exp\left(-\frac{1}{72}\right) \frac{\sum x_i^2 - 100\sum x_i + n2500}{\sum x_i^2 - 110\sum x_i + n3025}$$

$$= \exp\left[-\frac{1}{72}\right]\left[\sum x_i^2 - 100\sum x_i + n2500 - \sum x_i^2 + 110\sum x_i - n3025\right] \leq k.$$

$$= \exp\left[-\frac{1}{72}\right]\left[10\sum x_i - n525\right] \leq k$$

If we take the natural Log of each member the inequality we final out $\frac{-1}{72}(10\sum x_i - n525 \leq \ln k$.

now multiplying through by 72

$-(10\sum x_i - n525) \leq 72 \ln k$

to make $\sum x_i$ the subject we have

$-10\sum x_i \leq 72\ln k - n525$

$\sum x_i \geq n52.5 - 72\ln k$ by dividing both sides by n

$\bar{x} \geq 52.5 - \frac{72\ln k}{n}$

$\bar{x} \geq C$

where $C = 52.5 - 7.2\frac{\ln k}{n}$ is the best critical region according to NPL.

P.T.O

9/9/2020FS

Date: 1/1 1/20....

Exercise.

(1) Use Person Lemma (NPL) to obtain the best critical region for testing $H_0: \theta = \theta_0$ vs $H_1: \theta = \theta_1$ for a normal distribution given as $X \sim N(\mu, \sigma^2)$

~~Examples: determine whether the following vect~~ are -

(2) show that $C = x_1, x_2 \ldots x_n) \bar{X} < c$ is a best critical region for testing $H_0: \mu = 50$ vs $H_1: \mu = 76$

$C = \{(x_1, x_2 \ldots x_n) \bar{x} < c\}$ for $N(\mu, 64)$.

Date: / /20....

Topic:

UNBIASED TEST.

In statistical hypothesis testing, a test is said to be unbiased if the probability of committing type I error is equal to the significant level.

Unbiased estimator: an estimator $\hat{\theta}$ of parameter θ ~~and estimator~~ is unbiased if the expected value of θ is θ $E(\hat{\theta}) = \theta$

some distribution

(i) Bernoulli = $f(x/\theta) = \theta^x (1-\theta)^{1-x}$

$E(x) = \theta$ $\quad var = \theta(1-\theta)$.

(ii) poisson = $f(x/\theta) = \frac{\theta^x e^{-\theta}}{x!}$ $\quad E(x) = \theta$, $var = \theta$

(iii) Binomial $B(x; n, \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$

$E(x) = n\theta$, $var(x) = n\theta(1-\theta)$

(iv) Exponential

$f(x/\theta) = \frac{1}{\theta} e^{-x/\theta}$

$E(x) = \frac{1}{\theta}$ $\quad var(x) = \frac{1}{\theta^2}$

(v) Normal

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

Example: Let $X_1, X_2, \ldots, X_n$ be a random sample from an exponential distribution with parameter θ

(i) Show that $\hat{\theta}$ is an unbiased estimator of θ i.e $E(\hat{\theta}) = \theta$.

(ii) Show that its variance $Var(\hat{\theta}) = \frac{\theta^2}{n}$

Solution:

$$f(x/\theta) = \begin{cases} \frac{1}{\theta} e^{-x/\theta} & 0 \le x \le \infty \\ 0, & \text{elsewhere} \end{cases}$$

(i) Mean of exponential distribution.

$$E(x) = \hat{\theta}$$

$$\therefore \hat{\theta} = \bar{x}$$

$$E(\hat{\theta}) = E(\bar{x})$$

$$E(\bar{x}) = E\left(\frac{\Sigma x}{n}\right)$$

$$= \frac{1}{n}\left[E(x)\right]$$

Since $Ex = \theta$

$$= \frac{1}{n} E(\theta).$$

$$= \frac{1}{n} \cdot n\theta.$$

$$E(\hat{\theta}) = \theta.$$

(ii) Variance of exponential distribution

$$V(\bar{x}) = V\left(\frac{\Sigma x}{n}\right)$$

$$= \frac{1}{n}\left[V(x)\right]$$

Date | | 1/20...

$$= \frac{1}{n^2} \sum (\text{Var}(x_i))$$

$$\text{Var}(x) = \theta^2$$

$$= \frac{1}{n^2} \sum (\theta^2)$$

$$= \frac{1}{n^2} \cdot n\theta^2 = \frac{\theta^2}{n}$$

$$\therefore \text{Var}(\theta) = \frac{\theta^2}{n}$$

Exercise:

(1) Let $x_1, x_2, \ldots, x_n$ be a random sample from binomial distribution with parameter n and θ i.e. $b(x; n, \theta)$. find the ~~MLE~~ MM and show that $\bar{x}$ is an UMVUE estimator of $n\theta$

LIKELIHOOD RATIO TEST FOR UNIVARIATE

Definition: Let $L(\theta)$ equal to $L(x_1, x_2 \ldots \theta)$ be a likelihood function for the random variables $x, x_1, x_2, \ldots, x_n$. If $\hat{\theta}$ is the value of θ which maximizing likelihood function then $\hat{\theta}$ is called the maximum likelihood estimator (MLE) of θ for the sample $x_1, x_2, \ldots x_n$

Steps/condition

(1) Many likelihood function satisfied regularity condition. the MLE is the solution to the equation

$$\frac{\partial L(\hat{\theta})}{\partial \theta} = 0$$

Date: / /20...

② $L(\theta)$ and natural log $\log_e L(\theta)$ has been found to have their maximum at the same value θ and it's sometimes easier to find the maximum of natural log i.e $\log_e L(\theta)$.

Note if $x_1, x_2, \ldots x_n$ is a random sample the likelyhood function is the joint PDF $f(x_1, x_2, \ldots, x_n / \theta)$ which is given as

$$L(x_1, x_2, \ldots x_n / \theta) = \prod_{i=1}^{n} f(x_i / \theta).$$

MAXIMUM LIKELYHOOD ESTIMATOR (MLE)

an estimator $\hat{\theta}$ is defined as a maximum likelyhood estimator if

$$L(x_1, x_2, \ldots x_n / \hat{\theta}) = \text{Max}\{L(x_1, x_2 \ldots x_n)\}_{\theta}$$

example ①: Given a random sample $x_1, x_2, \ldots x_n$ of a binomial θ random sample. find the MLE of θ

Solution.

$$f(x; \theta) = \theta^x (1-\theta)^{1-x}$$

$$L(x; \theta) = \prod^{n} \left[\theta^x (1-\theta)^{1-x}\right].$$

$$= \theta^{\sum x - 1} (1-\theta)^{n - \sum x}$$

where $t = \sum x$.

$$= \theta^{t} (1-\theta)^{n-t}.$$

$$\log_e L(\hat{\theta}, \theta) = t \log \theta + n - t \log (1-\theta)$$

Date: / / 1/20....

Set

$$\frac{\partial \log L(x;\theta)}{\partial \theta} \Longrightarrow 0$$

$$\therefore \frac{t}{\theta} - \frac{n-t}{1-\theta} = 0$$

$$\frac{t}{\theta} = \frac{n-t}{1-\theta}$$

$$t(1-\hat{\theta}) = \hat{\theta}(n-t)$$

$$t - t\hat{\theta} = \hat{\theta}n - \hat{\theta}t$$

$$t - \hat{\theta}n = t = \hat{\theta}n$$

$$\hat{\theta} = \frac{t}{n}, \quad \hat{\theta} = \frac{\sum x}{n} = \bar{x}$$

Example 2: find d maximum likelyhood estimate of poisson distribution with parameter θ

Solution:

$$f(x_i;\theta) = \frac{\theta^x e^{-\theta}}{x!}$$

$$L(x;\theta) = \prod_{i=1} \frac{(\theta^x e^{-\theta})}{x!}$$

$$= \frac{\theta^{\sum x} e^{-n\theta}}{\prod x!}$$

where $\prod_{i=1}^{n} x_i$ = error $t = \sum x$

$$= \theta^t e^{-n\theta}$$

$$\log_e L(x;\theta) = t\log\theta - n\theta.$$

$$\frac{\partial \log_e L(x;\theta)}{\partial\theta} = 0$$

$$\frac{t}{\hat{\theta}} - n = 0$$

$$\frac{t}{\hat{\theta}} = n, \quad t = n\hat{\theta}, \quad \hat{\theta} = \frac{t}{n}$$

$$\hat{\theta} = \frac{\sum x}{n}, \quad \hat{\theta} = \bar{x}.$$

Exercise

(8) Given that a random sample $x_1, x_2, \ldots x_n$ of a normal distribution $X \sim N(\mu, \sigma^2)$ find.

(i) the MLE of μ and σ^2

(ii) show that $\bar{X}$ is an Unbiased estimator of μ

(iii) show that $\hat{\sigma}^2$ is not an Unbiased estimator of σ^2

Date 1/1 1/20...

TOPIC: Exercise,

Bernoulli

$(x_1, \dots x_n) = \theta^{x_i}(1-\theta)^{1-x_i}$

$= \theta^{x}(1-\theta)^{1-x}$

$= \theta^{\Sigma x}(1-\theta)^{n-\Sigma x} = 0$

$= \log \theta^{\Sigma x} + \log(1-\theta)^{n-\Sigma x} =$

$= \Sigma x \log\theta + (n-\Sigma x)\log(1-\theta)$

Differentiate $= \frac{\Sigma x}{\theta} - \frac{n-\Sigma x}{1-\theta} = 0$

$\frac{\Sigma x}{\theta} = \frac{n-\Sigma x}{1-\theta}$

$\theta n - \Sigma x\theta = \Sigma x - \Sigma x\theta$

$\theta n = \Sigma x$

$\theta = \frac{\Sigma x}{n}$

Binomial $= \theta^{x}(1-\theta)^{n-x}$

$L = \theta^{\Sigma x}(1-\theta)^{n-\Sigma x}$

$= \Sigma x \log\theta + (n-\Sigma x)\log(1-$

Poisson Distribution.

$P(X/\theta) = \frac{\theta^x e^{-\theta}}{x!}$

Normal distribution

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

Date: / /20....

Topic:

Youden Squares.

These are incomplete latin square designs in which the number of columns does not equal the number of rows and treatments. e.g Consider the design shown in the table below (Youden Squares for five treatments (A, B, C, D, E).

Row

(In working it's not necessary to that every Latin Square with more than one Column

The known model of Youden-square is

$$Y_{ijk} = \mu + \alpha_i + \tau_j + \beta_k + \varepsilon_{ijk}$$

where

μ is the overall mean

α_i is the i^{th} block effects to j^{th} treatment effect, β_k is the k^{th} position since position occur exactly once and ε_{ijk} is the $NID(0, \sigma^2)$ error since position occur exactly once position are orthogonal to treatment and blocks.

Example:

Day (Block)	1	2	3	4	y_i
1	A = 3	B = 1			
2	B = 0	C = 0			
3	C = -1	D = 0			
4	D = 1	E = 6			
5	E = 5				