IMPORTANT CONCEPTS Part-II

1. **Normal Distribution:** It is the limiting case of binominal distribution when n is very large and neither p or q is close to zero. The distribution is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\Lambda}} \quad e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2} \qquad -\infty < X < +\infty$$

μ = A. M
σ = S. D
Λ = 3.14159.....
e = 2.71828.....

2. **Properties of Normal Distribution:**
   - i) It is symmetric about x = U, the mean
   - ii) Its maximum ordinate is
   - iii) Total area under the normal curve is unity
   - iv) Its mean – Median – Mode and lie at the center.
   - v) The points of inflection lie at and
   - vi) The normal curve is asymptotic to the x – axis.
   - vii) All odd ordered mean moments are
   - viii) Its M. D – 4/5 (S. D)
   - ix) Its Q. D – 2/3 (S. D)
   - x) Lower and upper quartiles are given by the relations.
3. **Normal Frequency Distribution:** when Normal distribution is multiplied by N, the number of the sets or experiments, then normal frequency distribution is obtained as

4. **Relationship Between Binominal and Normal Distribution:** if n, is large and if neiter p nor q is too close to zero, the binominal distribution can be closely approximated by a Normal distribution with standard variable given by

$$Z = \frac{X - np}{\sqrt{npq}}$$

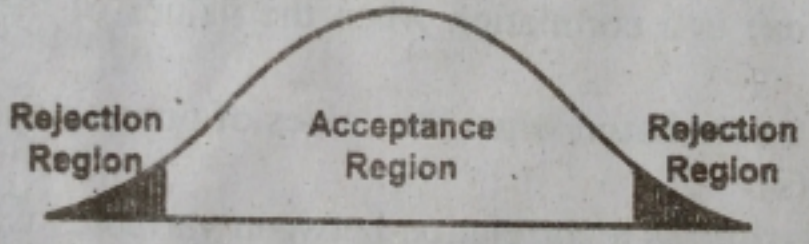
The approximation becomes better with increasing n, and in the limiting case it is exact, this approximation is very good if both np and nq are greater than 5.

5. **Population or Universe:** Is the whole aggregate about which some information is needed
6. **Sample:** is any part of population and is selected from the population with the belief that the part selected will show all the characteristics of the population.
7. **Sampling:** is a process of drawing sample from the population.
8. **Sample Units or Sample elements:** are the members, which make a sample
9. **Individual:** is the single member unit of the sample
10. **Cluster:** is a sample unit consisting of more than one members/elements
11. **Finite Population:** is a population consisting of finite number of elements e.g human population No of teachers, No of chairs, etc.
12. **Infinite Population:** is a population consisting of infinite number of elements e.g No of points on a line segment, No of stars in the sky.
13. **Existent Population:** is a population that is physically tested.
14. **Target Population:** is a population about which we want to get some information.
15. **Sampled Population:** is a population from which a sample is drawn
16. **Basic aims of sampling are:**
    - i) To get information about the Population on the basis of sample.

ii) To get reliability of estimates extracted from the sample.

17. **Sample Design:** is a definite statistical plan used in taking a sample and in estimation process.
18. **Sampling Frame:** is the complete list of all the elements of the population
19. **Advantages of Sampling are:**
    i) Sampling is cheaper
    ii) Skilled labour with better supervision can be employed
    iii) Investigation of already obtained results is easy
    iv) Sampling is a useful check on the accuracy of complete count
20. **Parameter:** is any result obtained from the population units. It is denoted by Greek letter.
21. **Statistic:** is any result obtained from the sample
22. **Probability Sampling:** is a type of sampling in which each and every element of the population has a known probability of being included in the sample i.e. Random sampling, stratified sampling, systematic sampling
23. **Non-Probability Sampling:** is a type of sampling in which selection of sample elements is not based on the probability theory, but personal judgement plays an important role e.g Quota sampling, Purposive sampling
24. **Sampling with Replacement:** is the method of selecting a sample in which unit selected once is returned to the population before drawing the next unit.
25. **Sampling without replacement:** is a method of selecting a sample in which unit once chosen is not replaced in the population before next unit is drawn. In this method a unit cannot be chosen more than once in the sample.
26. **Strata:** The non homogenous population is divided in different groups containing similar units. These groups are called strata. For each group (stratam) is selected in random manner.
27. **True Value:** We mean the value that would be obtained if no errors were made in any way, computing the characteristics of the population.
28. **Accuracy:** It is the difference between the sample result and true value. The smaller the difference the greater will be the accuracy.
    i. Elimination of technical errors.
    ii. Increasing the sample size. iii. Stratification.
29. **Precision:** By precision we refer to how closely, we can reproduce, from a sample, the results which would be obtained in a complete count was taken using the same method of measurement.
30. **Error:** It is the difference between the estimated value and the population true value.
31. **Sampling Error:** is the difference between the actual value of the parameter and the value of the statistic concerned.
32. **Non-Sampling Error:** are the errors which are not due to sampling techniques but they creep in during the collection of actual data. The main reasons are:-
    i) Defects in sample frame.
    ii) Negligence or indifference on the part of investigator.
    iii) Non-response to questionnaire.
33. **Bias:** is a systematic component of error which deprives a result of its representativeness.
34. **Sampling distribution:** is a probability distribution of the values of any statistic.
35. **Standard Error:** is the standard deviation of the sampling distribution of any statistic.
36. **Estimation:** is a process by which we get information about the unknown value of the population parameter by using sample values.
37. **Estimate:** is a particular value obtained from sample value as an estimate of the unknown population parameter.
38. **Estimator:** is a particular rule or formula used to calculate an estimate.
39. **Point estimate:** is an estimate of population parameter in the form of a single value.
40. **Interval estimate:** is an estimate of population parameter in the form of an interval.

41. **Unbiased estimator:** is an estimator whose expected value equals the population parameter concerned. An estimator of parameter θ is said to be unbiased. If $E(\theta) = \theta$
42. **Hypothesis:** is any statement which may or may not be true.
43. **Tests of Hypotheses/Tests of significance/Decision Procedures:** are the procedures used to make decisions about the population on the basis of information available from the sample.
44. **Null Hypothesis:** is any hypothesis about the population parameter under the assumption that it is true and further tested for possible rejection or acceptance. It is denoted by Ho.
45. **Alternative Hypothesis:** is any other hypothesis which is accepted when Null hypothesis is rejected. It is denoted by H_1 or Ha.
46. **Simple Hypothesis:** is a hypothesis which specifies all the values of all the parameters of a population. For example, a random variable X is normally distributed with mean 5 and variance 3.
47. **Composite Hypothesis:** is a hypothesis which does not specify all the values of all the parameters. For example, a random variable x is normally distributed with mean μ and variance 2.
48. **Level of Significance:** is the probability of rejecting the null hypothesis Ho when it is assumed to be true. It is denoted by α.
49. **Region of Rejection:** is the portion of sample distribution consisting of sampling results which lead to reject the true null hypothesis.
50. **Region of Acceptance:** is the portion of sampling distribution consisting of sample result which lead to accept the null hypothesis.



51. **Type-1 Error:** is the error committed by rejecting Ho when it should be accepted. Its probability is α.
52. **Type-II Error:** is the error committed by accepting Ho when it should be rejected. Its probability is β.

Decision	Accept H_o	Reject H_o
H_o is true	Correct decision	Wrong decision (Type-1 error)
H_o is false	Wrong decision (Type-II error)	Correct decision

53. **Test Statistics:** is a statistic following a particular well-known probability distribution, on which the decision whether to accept or reject H_o is based. The commonly used probability distributions to which test-statistics relate are:-
(i) Binominal Distribution (ii) Normal Distribution (iii) Chi-Square Distribution
(iv) Student's t-distribution (v) F-distribution
54. **One – tailed test:** is a test in which rejection region falls at only one end of the sampling distribution.
55. **Two – tailed test:** is a test in which rejection region falls equally at both ends of the sampling distribution.

Critical Values of Z for one-tailed and two-tailed tests

Level of Significance α	0.10	0.05	0.01	0.005	0.002
Critical Values of Z for One-tailed Tests	-1.28 or 1.28	-1.645 or 1.645	-2.33 or 2.33	-2.58 or 2.58	-2.88 or 2.88
Critical Values of Z for Two-tailed Tests	-1.645 and 1.645	-1.96 and 1.96	-2.58 and 2.58	-2.81 and 2.81	-3.08 and 3.08

Note: Critical values of other levels of significance can be determined by using the table of area under the normal curve.

56. **Regression:** is a term used to know the dependence of one variable on the other variable(s).
57. **Linear Regression:** is the linear (in the form of straight line) dependence of one variable upon the other(s).
58. **Curvi-linear Regression:** is the other than linear dependence of one variable upon the other(s).
59. **Correlation:** is the inter-dependence or inter-relationship between two or more variables.
60. **Positive Correlation:** is a Correlation when the values of the variables involved, move in the same direction.
61. **Negative Correlation:** is a correlation when the values of the variables involved, more in the opposite direction.
62. **No Correlation:** is a correlation when the values of one variables change and there is no change in the other variable(s).
63. **Curvilinear Correlation:** is a correlation between variables which is represented by some other than a straight line curve.
64. **Perfect Correlation:** is a correlation between variables when there is an increase or decrease in one, the value of the other variable increassessor decreases in a fixed proportions.
65. **Perfect Positive Correlation:** is the Correlation between two variables when both the variables move in the same direction but in the same proportions.
66. **Perfect Negative Correlation:** is the Correlation between two variables when both the variables move in the opposite direction but in the same proportions.
67. **Properties of a Correlation Co-efficient r:**
    i) It is a symmetric w.r.t. variables.
    ii) It is independent of change of origin and scale.
    iii) It lies between –1 and +1 both inclusive.
    iv) It is the Geometric Mean of two regression Co-efficients.
68. **Attribute:** is a characteristics of qualitative nature. It is denoted by capital letters A,B,C etc.
69. **Positive attributes:** are those attributes which are represented by letter A,B, AB, AC etc. etc.
70. **Negative attributes:** are those attributes which are represented by α,β, $\alpha\beta$, $\alpha\gamma$ etc. etc.
71. **Independence of attributes:** is defined as if $(AB) = \frac{(A)(B)}{N}$ Similarly $(A\beta) = \frac{(A)(\beta)}{N}$. Then attributes A and B are said to be independent.
72. **Association of attributes:** is defined as when they are not independent.
73. **Positive association of attributes:** is defined as if $(AB) \quad \frac{(A)(B)}{N}$

74. **Negative association of attributes:** is defined as if (AB)

75. **Degrees of freedom:** are defined (in Chi-Square Goodness of fit test) as the number of cells minus the number of quantities obtained from the observed data.

76. **Time Series:** is the arrangement of statistical data with respect to time.
77. **Components of time series:** are:-
    i) Secular Trend or long-term movements.
    ii) Seasonal variations or Short-term movements.
    iii) Cyclical movements.
    iv) Accidental or Irregular or Random movements.
78. **Secular Trend:** is the gradual and smooth long term movements. Its period is between 10 to 20 years. It is used to measures sales, population, industrial production, business progress, development etc.
79. **Seasonal variations:** are due to seasons. These are short-term and regular and can occur within a day, a week, a month or a quarter of a year.
80. **Cyclical movements:** are due to different cycles in the statistical data. A well-known example is of business cycle, which has four phases, namely
    (i) Prosperity (ii) Recession (iii) Depression (iv) Recovery
81. **Accidental movements:** are irregular in nature and are due to random causes. For example wars, floods, heavy rains, earth quakes, famines etc.
82. **Methods of measuring secular trend:** are
    i) Free hand drawing Method.
    ii) Semi-Average Method.
    iii) Moving Average Method.
    iv) Method of least square.
83. **Analysis of time series:** It is the de composition of time-series into various components.
84. **Computer (Def):** A computer is an electronic device, which under the direction of a program, process data, alters its own program instructions, and performs computations and logical operations without human intervention.
85. **Computer Program:** The term program refers to a specific set of instructions given to the computer to accomplish a specific task.
86. **Types of computer:** There are two main types of computer.
    1. Analog Computer
    2. Digital Computer

1. **Analog Computers:** An analog device is one that measures or processes data in a continuous form. A traditional watch or clock, fuel gauges are analog devices.
2. **Digital Computer:** A digital device is one that measure and represent quantities as discrete digits. Calculators, digital timepieces are examples of digital computers.

87. **Classification of computer:** Basically modern digital computer system falls into one of the following categories.

1. Mainframe Computers 2. Mini Computers 3. Super Computers
4. Micro Computers 5. Workstation

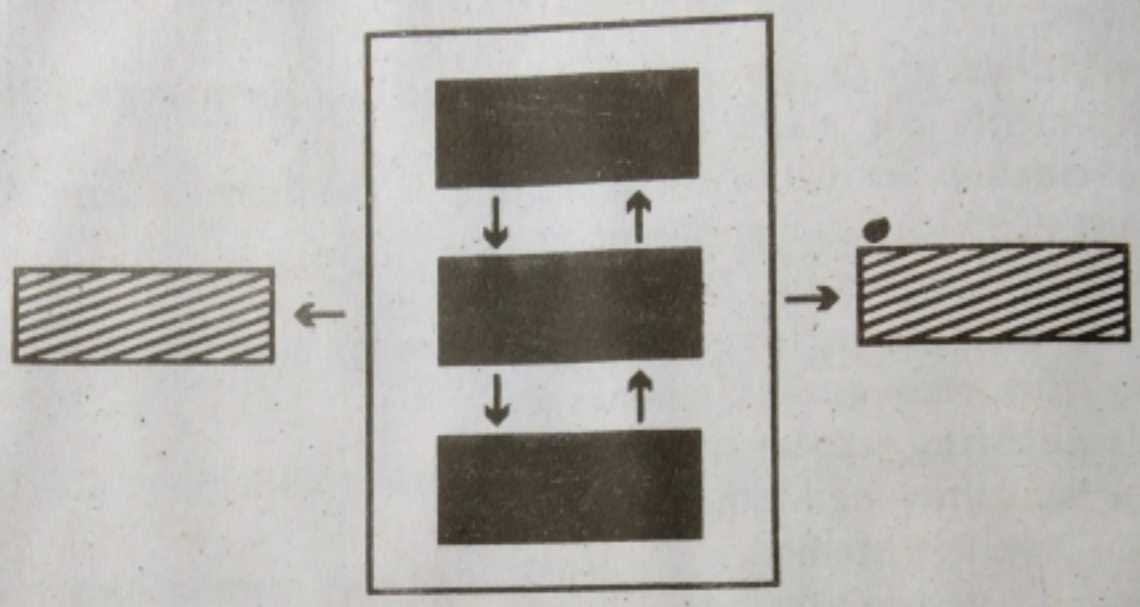
1. **Mainframe Computers:** The largest type of computer in common use is the mainframe. They are used in large organizations like insurance companies and banks where many people need frequent access to the same data, which are organized into one or more huge database. A mainframe system can house an enormous volume of data, containing literally billions of records.

2. **Mini Computers:** Mini computers were introduced in the 1960s. The capabilities of a mini computer are some where between mainframe and personal computers. Hundreds of personal computers can be connected to a mini computer.
3. **Workstation:** Workstations are specialized, single user computer with many of the features of a personal computer but with the processing power of minicomputer.
4. **Microcomputers:** The term microcomputer and personal computer are interchange able. A PC stands for personal computer. IBM introduced it in 1981.
5. **Super computers:** Super computers are the most powerful computers. These systems are built to process huge amounts of data, and fastest super computer can perform more than one trillion calculations per second.

Overview of computers and Languages

88. Computer units: A computer consist of five major components.

1. Input unit
2. Output unit
3. Memory unit
4. Control Unit
5. ALU



Block Diagram of Computer

1. **Input Unit:** Input unit consists of devices that permit a computer to receive information e.g. keyboard, mouse and scanners etc.
2. **Output Unit:** Output unit consists of devices that permit a computer to exhibit Information e.g.
3. **Memory Unit:** Memory unit is the component where all data and results are stored. The memory unit consists of many cells, each capable of storing a unit of information. These cells are also called storage location.
4. **Control Unit:** Control unit coordinates all the activities of the various components of the computer. It sends out command and control signals and determines the sequence of the various instruction.
5. **Arithmetic and logic unit:** ALU consists of electric circuits that perform the various arithmetic and logic operations.

Memory Unit, Control Unit and arithmetic unit collectively make CPU
(Central Processing Unit)

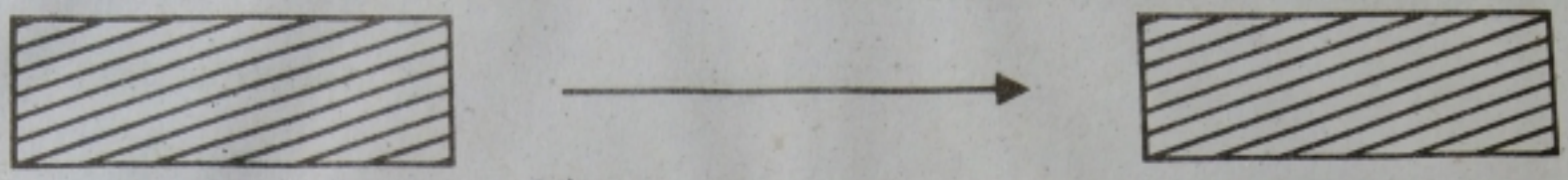
89. **(Central Processing Unit)**

Computer Languages

1. **Machine Language:** Every computer has its own machine language. Instruction to the computer must be given in that language since this is the only language that is understood by the computer. Program written in machine language are machine dependent. Instruction is usually represented by 1's and 0's format.
2. **Assembly Language:** In this language 0's and 1's are replaced by symbols, so that instruction can be given in symbolic code called mnemonics. An assembler program is then required to translate assembly language program into machine language program.

3. **High-level Language:** Today we can write computer program in an almost English language such as FORTRAN, BASIC, ALGOL, COBOL, C, C++ etc. programs written in this language are portable. A compiler is needed by the computer to translate programs into machine language. The original programe is called ***Source Program*** and translated program is called

90. Object Program



91. Computer Hardware Software

Hardware: Mechanical devices that make up computer are called hardware. Hardware is any part of the computer that you can touch. e.g. keyboard, mouse, hard disk, c.d. printer, monitor etc.

Software: Software is a set of electronic instructions consisting of complex codes that make computer perform task. In other words software tells computer what to do. e.g. MS Word, MS Windows, MS Paint etc.

92. Memory Units

1 bit	0 or 1
1 nibbles	4 bits
1 byte	8 bits
1-Kilo byte	10^3 byte
1 mega byte	10^6 bytes
1 gega byte	10^9 bytes
1 tera byte	10^{12} bytes

93. DOS: stand for Disk Operating System.

94. Application software: programs that help people accomplish specific tasks are refer to as application software, e.g. word processing software, databases etc.

IMPORTANT FORMULAE Part-II

NORMAL DISTRIBUTION

The normal distribution is the limiting form of the binominal distribution, when 'n' is large, neither "P" nor "q" are very small. The equation of the normal curve is given.

Where μ and σ are called parameters of the normal distribution also Mean, and standard deviation of population.

$\Pi = 3.1416$ $e = 2.7183$ constants

Normal frequency distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\Lambda}} \; e^{\,2\left(-\frac{\mu}{\sigma}\right)} \qquad -\infty < X < +\infty \text{ and } \sigma > 0$$

Where 'N' is No, of experiments.

$$y = N\left[\frac{1}{\sigma\sqrt{211}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\right] \qquad -\infty < X < +\infty \text{ and } \sigma > 0$$

3. SAMPLING AND SAMPLING DISTRIBUTION

Notations and Explainations:

"n" sample size 'N' population size

No, of possible samples that can be drawn are from

(i) With replacement (W.R) = $(N)^n$

(ii) Without replacement (W.O.R) = ${}^{N}_{n}c = \frac{N!}{(N-n)\,!n!}$

$\bar{X} = \frac{\Sigma X}{n}$ sample Mean, $\mu = \frac{\Sigma X}{N}$, Population Mean

$\hat{p} = \frac{x}{n}$ sample proportion of success P $\frac{x}{N}$ population proportion of success

"X" No. of success (one category)

$S^2 = \frac{\Sigma(x - \bar{x})^2}{n}$ biased sample variance.

$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$ unbiased sample variance.

$\sigma^2 = \frac{\Sigma(x - \mu)^2}{N} = \frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2$, Population variance

If $n = 2$, then $S^2 = \left(\frac{x_1 - x_2}{2}\right)^2$ and $s^2 = \frac{(x_1 - x_2)^2}{2}$

where X_1, X_2 are the ist and 2nd value of each sample

Sampling distribution of Sample Means ($\bar{x}$)

$$\mu_{\bar{x}} = \mu$$

$$\sigma^2_{\bar{x}} = \text{var}(\bar{x}) = \frac{\sigma^2}{n} \quad \text{(W.R.)}$$

$$\sigma^2_{\bar{x}} = \text{var}(\bar{x}) = \frac{\sigma^2}{n}\,\frac{N-n}{N-1} \quad \text{(W.O.R.)}$$

$\sigma_{\bar{x}}$ is called standard error of Mean i.e. (S.E. of $\bar{x}$)

Sampling distribution of difference between two sample Means ($\bar{x}_1 - \bar{x}_2$)

$$\mu_{(\bar{x}_1 - \bar{x}_2)} = \mu_1 - \mu_2$$

$$\sigma^2(\bar{x}_1 - \bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \quad \text{(W.R.)}$$

$$\sigma^2(\bar{x}_1 - \bar{x}_2) = \frac{\sigma_1^2}{n_1}\left(\frac{N_1 - n_1}{N_1 - 1}\right) + \frac{\sigma_2^2}{n_2}\left(\frac{N_2 - n_2}{N_2 - 1}\right) \quad \text{(W.O.R)}$$

Sampling distribution of sample proportion of success (p)

$$\mu_{\hat{p}} = P$$

$$\text{var}(\hat{p}) = \sigma^2_{\hat{p}} = \frac{PQ}{n}$$

$$\text{var}(\hat{p}) = \sigma^2_{\hat{p}} = \frac{PQ}{n}\left(\frac{N-n}{N-1}\right) \quad \text{(W.O.R.)}$$

Sampling distribution of difference between two sample proportions ($p_1 - p_2$)

$$\mu(\hat{p}_1 - \hat{p}_2) = P_1 - P_2$$

$$\text{var}(\hat{p}_1 - \hat{p}_2) = \sigma^2_{(\hat{p}_1 - \hat{p}_2)} = \frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2} \quad \text{(W.R)}$$

$$\text{var}(\hat{p}_1 - \hat{p}_2) = \sigma^2_{(\hat{p}_1 - \hat{p}_2)} = \frac{P_1Q_1}{n_1}\left(\frac{N_1 - n_1}{N_1 - 1}\right) + \frac{P_2Q_2}{n_2}\left(\frac{N_2 - n_2}{N_2 - 1}\right) \quad \text{(W.O.R)}$$

Where

$$\mu_{\bar{x}} = \Sigma \bar{x} f(\bar{x}) \qquad \mu_{\hat{p}} = \Sigma p f(p)$$

$$\text{var}(\bar{x}_1 - \bar{x}_2) = \Sigma(\bar{x}_1 - \bar{x}_2)^2 f(\bar{x}_1 - \bar{x}_2) - \left(\Sigma(\bar{x}_1 - \bar{x}_2) f(\bar{x}_1 - \bar{x}_2)\right)^2$$

$$\text{var}(\hat{p}_1 - \hat{p}_2) = \Sigma(\hat{p}_1 - \hat{p}_2)^2 f(\hat{p}_1 - \hat{p}_2) - \left(\Sigma(\hat{p}_1 - \hat{p}_2) f(\hat{p}_1 - \hat{p}_2)\right)^2$$

STATISTICAL INTERENCE

CONFIDENCE INTERVAL (C.I)

C.I estimating Population mean 'μ'

$\bar{X} \mp Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ When σ known

$\bar{X} \mp Z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$ When σ Not Known and 'n' > 30

$\bar{X} \mp t_{\frac{\alpha}{2}}(v) \frac{s}{\sqrt{n}}$ When σ Not Known and 'n' < 30

Where v = (n - 1) d.f.

C.I. estimating difference between two population means '$\mu_1 - \mu_2$'

(i) $(\bar{X}_1 - \bar{X}_2) \mp Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ when variances are known

(ii) $(\bar{X}_1 - \bar{X}_2) \mp Z_{\frac{\alpha}{2}} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$ when variances are not known and n_1, n_2 are large.

$(\bar{X}_1 - \bar{X}_2) \mp t_{\frac{\alpha}{2}}(v)\ Sp \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ When variances of normal populatios are not known but assumed to be equal and $n_1, n_2 < 30$

where $v = n_1 + n_2 - 2$, d.f.

$$S^2p = \frac{1}{v}\left[\Sigma (X_{1i} - \bar{X}_2)^2 + \Sigma (X_{2i} - \bar{X}_2)^2\right]$$

$$S^2p = \frac{1}{v}\left[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2\right]$$

C.I. estimating population proportion of success (P) n > 30

$\hat{P} \mp Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}}$ $\hat{q} = 1 - \hat{p}$

C.I. estimating difference between two population proportions when n1, n2 are large (P1 - P2)

$$(\hat{P}_1 - \hat{P}_2) \mp Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

Useful table values for C.I. In case of large samples

$1 - \alpha$	85%	90%	92%	95%	98%	99%
$Z_{\frac{\alpha}{2}}$	1.44	1.645	1.75	1.96	2.33	2.58

TESTING HYPOTHESIS

Testing hypothesis about the population mean (μ)

Null and Alternative hypothesis

Ho: $\mu = \mu_o$ against H_1: $\mu \neq \mu_o$ or $\mu < \mu_o$ or $\mu > \mu_o$ (two tailed test)

Ho: $\mu \geq \mu_o$ against H_1: $\mu < \mu_o$ (one tailed test)

Ho: $\mu \leq \mu_o$ against H_1: $\mu > \mu_o$ (one tailed test)

Test Statistic

$Z = \dfrac{\bar{x} - \mu_o}{\sigma/\sqrt{n}}$ is exactly N (0,1) When σ is known

$Z = \dfrac{\bar{x} - \mu_o}{S/\sqrt{n}}$ is approximately N (0,1) When σ is Not Known and 'n' > 30

$t = \dfrac{\bar{x} - \mu_o}{s/\sqrt{n}}$ with ν df. When σ Not Known and 'n' < 30
$\nu = n - 1$ population to be normal

Testing hypothesis about difference between two population means

Null and Alternative hypothesis

Ho: $\mu_1 - \mu_2 = \Delta_0$ against H_1: $\mu_1 - \mu_2 \neq \Delta_0$ (two tailed test)

Ho: $\mu_1 - \mu_2 \geq \Delta_0$ against H_1: $\mu_1 - \mu_2 < \Delta_0$ (two tailed test)

Ho: $\mu_1 - \mu_2 \leq \Delta_0$ against H_1: $\mu_1 - \mu_2 > \Delta_0$ (two tailed test)

Test Statistic

$Z = \dfrac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$, is exactly N (0,1) When variances are known

$Z = \dfrac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$, is exactly N (0,1) When variances are equal

$Z = \dfrac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$, is app. N (0,1) When variances are not known and $n_1, n_2 > 30$

$t = \dfrac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{Sp\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ with $\nu = n_1 + n_2 - 2$ When variances are not known but assumed to be equal and $n_1, n_2 < 30$

Testing hypothesis about the population proportion 'P' (n>30)

Ho: $P = P_o$ against H_1: $P \neq P_o$ (two tailed test)

Ho: $P \geq P_o$ against H_1: $P < P_o$ (one tailed test)

Ho: $P \leq P_o$ against H_1: $P > P_o$ (one tailed test)

Test Statistic

$$Z = \frac{P - P_o}{\sqrt{\frac{P_o q_o}{n}}}$$ is approximately N (0,1) $q_o = 1 - P_o$

Testing hypothesis about the difference between population proportion when (n_1, n_2) > 30)

Ho: $P_1 - P_2 = \Delta_1$ against H_1: $P_1 - P_2 \neq \Delta_1$ (two tailed test)

Ho: $P_1 - P_2 \geq \Delta_1$ against H_1: $P_1 - P_2 < \Delta_1$ (one tailed test)

Ho: $P_1 - P_2 \leq \Delta_1$ against H_1: $P_1 - P_2 > \Delta_1$ (one tailed test)

Test Statistic

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (\Delta_1)}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}}$$ is app. N (0,1) When P_1, P_2 are known

$$Z = \frac{(P_1 - P_2) - \Delta_0}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}}$$ is app. N (0,1) When P_1, P_2 are not known and not equal) (Δ_0 any given value)

$$Z = \frac{(P_1 - P_2)}{\sqrt{\hat{p}_c\hat{q}_c\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$ is app. N (0,1) When P_1, P_2 are unknown but equal

where $\hat{P}c = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$ $\hat{q}c = 1 - \hat{P}c$

REGRESSION AND CORRELATION

Regression

Regression lines i) Y on X $(y - \bar{y}) = byx\,(x - \bar{x})$

ii) X on Y $(x - \bar{x}) = bxy\,(y - \bar{y})$

Where 'byx' and 'bxy' are called the regression coefficients which measures the slope of the line. (Either both positive or both negative) where

$$byx = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2} \qquad bxy = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(y - \bar{y})^2}$$

$$byx = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma x^2 - n(\bar{x})^2} \qquad bxy = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma y^2 - n(\bar{y})^2}$$

$$byx = \frac{Sxy}{nS_x^2} \qquad bxy = \frac{Sxy}{nS_y^2}$$

$$byx = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma(x - \bar{x})^2} \qquad bxy = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma(y - \bar{y})^2}$$

$$byx = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{nS_x^2} \qquad bxy = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{nS_y^2}$$

$$byx = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} \qquad bxy = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma y^2 - (\Sigma y)^2}$$

$$byx = \frac{n\Sigma DxDy - (\Sigma Dx)(\Sigma Dy)}{n\Sigma Dx^2 - (\Sigma Dx)^2} \qquad bxy = \frac{n\Sigma DxDy - (\Sigma Dx)(\Sigma Dy)}{n\Sigma Dy^2 - (\Sigma Dy)^2}$$

$$byx = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2} \qquad bxy = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma v^2 - (\Sigma v)^2}$$

Where $Dx = X - P.M.(x)$ $\qquad Dy = Y - P.M.(y)$

$$u = \frac{X - P.M.(x)}{h} \qquad v = \frac{Y - P.M.(y)}{h}$$

h, k are common or equal intervals

Sx, Sy S.D.(x) and S.D.(y) r : coefficient of correlation

$$byx = r\left(\frac{Sy}{Sx}\right) \qquad bxy = r\left(\frac{Sx}{Sy}\right)$$

Correlation (r)

(1) $$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2 \Sigma(y - \bar{y})^2}}$$

(2) $$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{nSxSy}$$

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

(4) $$r = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{(n\Sigma u^2 - (\Sigma u)^2)(n\Sigma v^2 - (\Sigma v)^2)}}$$

(5) $$r = \pm\sqrt{(bxy)(byx)}$$

(6) $$r = \frac{n\Sigma DxDy - (\Sigma Dx)(\Sigma Dy)}{\sqrt{(n\Sigma D^2x - (\Sigma Dx)^2)(n\Sigma D_y^2 - (\Sigma Dy)^2)}}$$

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

ASSOCIATION

Chi-Square test for independence in contigency table (X^2)

H_o: Attributes are independent, Against H_1: They are associated.

Test Statistic

$$X^2 = \Sigma \frac{(fo - fe)^2}{fe}$$ with (r - 1) (c - 1) d.f.

where, r = No, of rows, and c = No, of columns

fo = observed frequencies, fe = Expected frequencies

Co-efficient of association (Q)

$$Q = \frac{(AB)(\alpha\beta) - (\alpha B)(A\beta)}{(AB)(\alpha\beta) + (\alpha B)(A\beta)}$$

If
- Q = o there is no association (independence)
- Q = +ve there is positive association
- Q = -ve there is negative association

Attributes A and B are independent

$$\text{If } (AB) = \frac{(A)(B)}{N}$$

TIME SERIES

Components of time series

Secular Trend (T) — Seasonal Variation (S)

Cyclical fluctuation (C) — Irregular variation (I)

Time Series 'Y' follows (i) Additive Model (ii) Multiplicative Model

In Additive Model $Y = T + S + C + I$

In Multiplicative Model $Y = T\ S\ C\ I$

(i) Straight line

$Y = a + bx$

Normal equations $\Sigma y = na + b\Sigma x$

$\Sigma xy = a\Sigma x + b\Sigma x^2$

(ii) Second degree parabola (Quadratic curve)

$Y = a + bx + cx^2$

Normal equations $\Sigma y = na + b\Sigma x + c\Sigma x^2$

$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3$

$\Sigma x^2y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4$

Usually, the coded "x" obtained, from the time series data as

$x = \frac{1}{h}$ (years - Middle year) When No. of year is odd

$x = \frac{2}{h}$ (years - Mean of two Middle years) When No. of year is even

In both cases $\Sigma x = \Sigma x^3 = 0$

Then the normal equations are for

(i) $a = \frac{\Sigma y}{n} = \bar{y}$, $b = \frac{\Sigma xy}{\Sigma x^2}$ then $\hat{y} = a + bx$

(ii) $\Sigma y = na + c\Sigma x^2$ $\Sigma xy = b\Sigma x^2$ $\Sigma x^2y = a\Sigma x^2 + c\Sigma x^4$

when $b = \frac{\Sigma xy}{\Sigma x^2}$ after solving we get a and c

they fitted, $\hat{y} = a + bx + cx^2$