Syllabus for Statistics (B.Sc. CSIT) program

Tribhuvan University
Institute of Science & Technology(IOST)

Level: B.Sc. | Full Marks: 60 + 20 + 20
Course Title: Statistics II | Pass Marks: 24 + 8 + 8
Course Code: STA 210 | Credit Hrs : 3
Nature of the Course: Theory and Practical

Course objectives:

To impart the theoretical as well as practical knowledge of estimation, testing of hypothesis, application of parametric and non-parametric statistical tests, design of experiments, multiple regression analysis, and basic concept of stochastic process with special focus to data/problems related with computer science and information technology.

1. Sampling Distribution and Estimation **[6]**

Sampling distribution of mean and proportion; Concept of Central Limit Theorem; Concept of inferential Statistics; point estimation; Properties of a "Good" estimator: unbiasedness, consistency, efficiency and sufficiency; Methods of estimation: Maximum likelihood estimation, Method of moments; Interval estimation: Confidence interval and confidence coefficient, confidence limits, confidence interval of mean for normal population. Confidence interval for proportion; Determination of sample size, relationship of sample size with desired level of error.

Problems and illustrative examples related to computer Science and IT

2. Testing of hypothesis **[8]**

Types of statistical hypotheses – null and alternative hypothesis, type I and type II errors, level of significance, critical value and critical region, power of the test, concept of p-value and use of p - value in decision making, steps used in testing of hypothesis, one sample tests for mean of normal population (for known and unknown variance), test for single proportion, test for difference between two means and two proportions, paired sample t-test; Linkage between confidence interval and testing of hypothesis; Assumptions for applying independent t-test, paired t-test; Test of equality of two variances

Problems and illustrative examples related to computer Science and IT

3. Non parametric test **[8]**

Parametric vs. non-parametric test; Needs of applying non-parametric tests; One-sample test: Run test, Binomial test**,** Kolmogorov–Smirnov test; Two independent sample test: Median test, Kolmogorov-Smirnov test, Wilcoxon Mann Whitney test, Chi-square test; Paired-sample test:

Wilcoxon signed rank test; Cochran's Q test; Friedman two way analysis of variance test; Kruskal Wallis test.

Problems and illustrative examples related to computer Science and IT

4. Multiple correlation and regression **[6]**

Multiple and partial correlation; Introduction of multiple linear regression; Least square estimation of parameters; Properties of least square estimators; Matrix approach to multiple linear regression; Hypothesis testing of multiple regression(upto two independent variables): Test of significance of regression, test of individual regression coefficient; Model adequacy test: Residual analysis, influential observation, multicollinearity; Coefficient of determination, Adjusted R^2 , and their interpretations.

Problems and illustrative examples related to computer Science and IT

5. **Design of experiment** **[10]**

Basic terminologies of experimental design; Basic principles of experimental designs; Completely Randomized Design (CRD): Statistical analysis of CRD, ANOVA table, Advantages and disadvantages, concept of multiple comparisons; Randomized Block Design (RBD): Statistical analysis of RBD for one observation per experimental unit, ANOVA table, Efficiency of RBD relative to CRD, Estimations of missing value (one observation only), Advantages and disadvantages; Latin Square Design (LSD): Statistical analysis of m x m LSD for one observation per experimental unit, ANOVA table, Estimation of missing value in LSD (one observation only), Efficiency of LSD relative to RBD, Advantage and disadvantages.

Problems and illustrative examples related to computer Science and IT

6. Stochastic Process **[7]**

Definition and classification; Markov Process: Markov chain, Matrix approach, Steady- State distribution; Counting process: Binomial process, Poisson process; Simulation of stochastic process; Queuing system: Main component of queuing system, Little's law; Bernoulli single server queuing process: system with limited capacity; M/M/1 system: Evaluating the system performance.

Practical (Computational Statistics): **[15]**

Practical problems to be covered in the Computerized Statistics laboratory:

Practical problems

S. No.	Title of the practical problems (Using any statistical software such as SPSS, STATA etc. whichever convenient).	No. of practical problems
1	Sampling distribution, random number generation, and computation of sample size	1
2	Methods of estimation(including interval estimation)	1
3	Parametric tests (covering most of the tests)	3
4	Non-parametric test(covering most of the tests)	3
5	Partial correlation	1
6	Multiple regression	1
7	Design of Experiments	3
	Stochastic process	2
	Total number of practical problems	**15**

Text Books:

1. Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers, & Keying Ye(2012). Probability & Statistics for Engineers & Scientists. 9th Ed., Printice Hall.

2. Michael Baron (2013). Probability and Statistics for Computer Scientists. 2nd Ed., CRC Press, Taylor & Francis Group, A Chapman & Hall Book.

Reference Books:

1. Douglas C. Montgomery & George C. Runger(2003). Applied Statistics and Probability for Engineers. 3rd Ed., John Willey and Sons, Inc.

2. Sidney Siegel, & N. John Castellan, Jr. Nonparametric Statistics for the Behavioral Sciences, 2nd Ed., McGraw Hill International Editions.

Tribhuvan University

Institute of Science and Technology

Model Question

Bachelor Level/Second Year/Third Semester/Science — Full Marks: 60

Computer Science and Information Technology STA 210 — Pass Marks: 24

(Statistics II) — Time : 3 Hours

Candidates are required to give their answers in their own words as far as practicable.
All notations have the usual meanings.

Group A

Attempt any Two questions $(2 \times 10 = 20)$

1. Suppose a population of 4 computers with their lifetimes 3, 5, 7 & 9 years. Comment on the population distribution. Assuming that you sample with replacement, select all possible samples of n = 2, and construct sampling distribution of mean and compare the population distribution and sampling distribution of mean. Compare population mean versus mean of all sample means, and population variance versus variance of sample means and comment on them with the support of theoretical consideration if any.

2. A computer manager is keenly interested to know how efficiency of her new computer program depends on the size of incoming data and data structure. Efficiency will be measured by the number of processed requests per hour. Data structure may be measured on how many tables were used to arrange each data set. All the information was put together as follows.

Data size(gigabytes)	**6**	**7**	**7**	**8**	**10**	**10**	**15**
Number of tables	4	20	20	10	10	2	1
Processed requests	40	55	50	41	17	26	16

Identify which one is dependent variable? Fit the appropriate multiple regression model and provide problem specific interpretations of the fitted regression coefficients.

3. State and explain the mathematical model for randomized complete block design. Explain all the steps to be adopted to carry out the analysis, and finally prepare the ANOVA table.

Group B

Attempt any Eight questions $(8 \times 5 = 40)$

4. In order to ensure efficient usage of a server, it is necessary to estimate the mean number of concurrent users. According to records, the average number of concurrent users at 100 randomly selected times is 37.7, with a sample standard deviation of 9.2. At the 1% level of significance, do these data provide considerable evidence that the mean number of concurrent users is greater than 35? Draw your conclusion based on your result.

5. A sample of 250 items from lot A contains 10 defective items, and a sample of 300 items from lot B is found to contain 18 defective items.
At a significance level $\alpha = 0.05$, is there a significant difference between the quality of the two lots?

6. Modern email servers and anti-spam filters attempt to identify spam emails and direct them to a junk folder. There are various ways to detect spam, and research still continues. In this regard, an information security officer tries to confirm that the chance for an email to be spam depends on whether it contains images or not. The following data were collected on n = 1000 random email messages.

	Image containing status		
Spam status	With images	No images	**Total**
Spam	160	240	400
No spam	140	460	600
Total	300	700	1000

Assess whether being spam and containing images are independent factors at 1% level of significance.

7. Two computer makers, A and B, compete for a certain market. Their users rank the quality of computers on a 4-point scale as "Not satisfied", "Satisfied", "Good quality", and "Excellent quality, will recommend to others. The following counts were observed:

Computer maker	Not satisfied	Satisfied	Good quality	Excellent quality
A	20	40	70	20
B	10	30	40	20

Is there a significant difference in customer satisfaction of the computers produced by A and by B using Mann-Whitney U test at 5% level of significance.

8. Define queuing systems with suitable examples. Also explain the main components of queuing systems in brief.

9. In some town, each day is either sunny or rainy. A sunny day is followed by another sunny day with probability 0.7, whereas a rainy day is followed by a sunny day with probability 0.4. Weather conditions in this problem represent a homogeneous Markov chain with 2 states: state 1 = "sunny" and state 2 = "rainy." Transition probability matrix of sunny and rainy days is given below.

$$P = \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}$$

Compute the probability of sunny days and rainy days using the steady-state equation for this Markov chain.

10. Consider a completely randomized design with 4 treatments with 7 observations in each. For the ANOVA summary table below, fill in all the missing results. Also indicate your statistical decision.

Source	Degrees of freedom	Sum of Squares	Mean Sum of Squares	F-ratio
Treatments	?	SSA =?	70	F =?
Error	?	SSE = 590	?	
Total	?	SST =?		

11. Following are the scores obtained by 10 university staffs on the computer proficiency skills before training and after training. It was assumed that the proficiency of computer skills is expected to be increased after training.

Staffs	Score	
1	Before training	After training
2	50	55
3	30	40
4	15	30
5	22	30
6	34	36
7	45	45
8	40	41
9	10	30
10	26	40

Test at 5% level of significance whether the training is effective to improve the computer proficiency skills applying appropriate statistical test. Assume that the given score follows normal distribution.

12. Write short notes on the following.

a) Concept of Latin Square Design
b) Multiple correlation