DETAIL SYLLABUS

Course Title: Basic Statistics **Full Marks:** 60 + 20 + 20
Course No: STA154 **Pass Marks:** 24 + 8 + 8
Nature of the Course: Theory + Lab **Credit Hrs. :** 3
Semester: II

Course Description:

The course familiarizes students with the basic concepts of statistics including introduction, diagrammatical and graphical representation, descriptive statistics, probability, random variables, sampling, and correlation and regression.

Course Objective:

To impart the knowledge of descriptive statistics, correlation, regression, concept of sampling and sampling distribution, theoretical as well as applied knowledge of probability and some probability distributions.

Course Contents:

Unit 1: Introduction (5 Hrs.)

Basic concept of statistics(definitions, concept of descriptive and inferential statistics), application of Statistics in different fields including information technology, limitations of statistics; scales of measurement(nominal, ordinal, interval and ratio), variables(Discrete, continuous and categorical), types of data(cross-sectional and longitudinal) and data source(primary and secondary), data preparation- editing, coding, and transcribing.

Unit 2: Diagrammatical and Graphical Presentation of Data (3 Hrs.)

Bar diagrams; Pie diagrams; Pareto chart; Graph of frequency distribution (Histogram, frequency polygon, frequency curve, less than ogive and more than ogive, stem and leaf display) and their interpretations

Unit 3: Descriptive Statistics (7 Hrs)

Measures of central tendency: Definition of measures of central tendency, arithmetic mean and its mathematical properties, weighted mean, median, mode, empirical relations between mean , median and mode; choice of measure of central tendency, interpretations

Measures of dispersion: Need of measures of dispersion, absolute and relative measures, range, quartile deviation, mean deviation, standard deviation and their relative measures including coefficient of variation, choice of appropriate measure of dispersion, interpretations

Measures of skewness: Concept of skewness, types of skewness, Pearson’s coefficient of skewness, Bowley’s coefficient of skewness; Exploratory Data Analysis (EDA): Five number summary, box and whisker plots, outliers, use of five number summary and boxplots to assess the skewness of data distribution
Measures of kurtosis: Concept of kurtosis, types of kurtosis, measure of kurtosis based on percentiles; overall assessment of nature of data distribution
Moments: Introduction of moments, central moments and raw moments, relations between central moments and raw moments, measures of skewness and kurtosis based on moments
Problems and illustrative examples related to IT

Unit 4: Introduction to Probability (7 Hrs.)

Concepts of probability, definitions of probability (mathematical, statistical and subjective approach), terminologies used in probability, laws of probability (additive and multiplicative), conditional probabilities, Bayes theorem: prior and posterior probabilities
Problems and illustrative examples related to IT

Unit 5: Random Variables and Mathematical Expectation (3 Hrs.)

Concept of a random variable and its types, probability distribution of a random variable, mathematical expectation of a discrete random variable, standard deviation and variance of discrete random variable, addition and multiplication theorems of expectation and variance(without proof).
Problems and illustrative examples related to IT

Unit 6: Probability Distributions (6 Hrs.)

Probability distribution function**,** Binomial distribution, Poisson distribution, Normal distribution and their characteristic features; applications of these distributions in IT related data problems
Problems and illustrative examples related to computer Science and IT

Unit 7: Sampling and Sampling Distribution (7 Hrs.)

Definitions of population, sample survey vs. census survey, sampling error and non-sampling error, concept of parameter and statistic, types of sampling(concept of simple random, stratified, cluster and systematic sampling, concept of non-probability sampling), standard error of mean, standard error of proportion, sampling distribution of mean and proportion, need of inferential statistics, concept of central limit theorem, concept of estimation(point and interval), confidence interval estimation for mean & proportion, problem specific interpretations of confidence interval
Problems and illustrative examples related to IT

Unit 8: Correlation and Linear Regression (7 Hrs.)

Bivariate data, bivariate frequency distribution, correlation between two variables, Karl Pearson's coefficient of correlation(r), assumptions of Pearson's correlation coefficient, properties of correlation coefficient, Spearman's rank correlation including repeated ranks, interpretation of correlation coefficient, need of regression analysis, fitting of lines of regression by the least squares method, interpretation of regression coefficients, coefficient of determination(R^2) and its interpretation, residual plots for assessing the goodness of fit of the model
Problems and illustrative examples related to IT

Laboratory Works:

Practical (Computational Statistics):

Practical problems to be covered in the Computerized Statistics laboratory

Practical Problems

S. No.	**Title of Practical Problems** (Using any statistical software such as Microsoft Excel, SPSS, STATA, etc. whichever convenient).	**No. of practical problems**
1	Diagrammatical and graphical presentation of data	1
2	Computation of measures of central tendency (ungrouped and grouped data), use of an appropriate measure and interpretation of results and computation of partition values	1
3	Computation measures of dispersion (ungrouped and grouped data) and computation of coefficient of variation.	1
4	Measures of skewness and kurtosis using method of moments, measures of skewness using box and whisker plot	2
5	Scatter diagram, correlation coefficient (ungrouped data) and interpretation. Compute manually and check with computer output	1
6	Fitting of simple linear regression model (results to be verified with computer output), residuals plot	1
7	Conditional probability and Bayes theorem	3
8	Problems related to Binomial, Poisson and Normal probability distributions	2
9	Problems related sampling and sampling distribution of mean and proportion, confidence interval estimation for mean and proportion	3
	Total number of practical problems	**15**

Text Books:

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

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

Reference Books:

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

2. Richard A. Johnson (2001). Probability and Statistics for Engineers. 6th Ed., Pearson Education, India

Tribhuvan University
Institute of Science and Technology

Model Question

Bachelor Level/I Year/II Semester/Science — Full Marks: 60
Bachelor of Information Technology (BIT) — Pass Marks: 24
Basic Statistics (STA154) — 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. **[2×10 = 20]**

1. There are two popular Cyber Café's (namely Café A and Café B) in Kathmandu located at *Thamel* area. Each of them has good number of computers in the cyber and maintains peace, comfort and good working environment. Each costumer, before entering into the café, asks to the café owner about the average time to download an image file since the internet in Kathmandu is a bit of sporadic. Each of them replied that the time to download an image file is on an average no more than 70 seconds. The following is the data of downloading time for an image file experienced by 8 customers in each café in some random 8 different days.

Download time in café A(in seconds)	Download time in café B(in seconds)
70	70
55	75
55	72
45	73
40	73
80	80
75	74
69	40

a) Explain whether the owner's response to the costumers are satisfied in each café? Even having the average waiting time within the owner's limit, is there any further comments on the data with reference to measures of central tendency you have computed? Discuss.

b) What similarities and differences are observed based on the average download time in Café A and in café B?

c) Compare the download time between two café's with respect to variability and shape of the data distribution.

d) On the basis of your statistical analysis, which café would be suggested to the customers so that one can download an image file in a lesser time?

2. A big computer supplier in Kathmandu used to sell large number of computers in each year. His interest is to increase his sales volume in each year for which the supplier has started to take the help of advertisement and allocated some advertisement expenditure in his annual budget each year. The supplier wants to quantify the effect of advertise expenditure on sales of computers. The advertisement expenditure (in lakhs rupees) and their corresponding sales from the computers (in crores rupees) is tabulated as follows.

Advertisement expenditure	**40**	**50**	**38**	**60**	**65**	**50**	**35**
Sales	38	60	55	70	60	48	30

   Assuming that the relationship between the advertisement expenditure and sales is linear. Response the following questions.

   a) Perform appropriate statistical analysis and quantify the effect of advertisement on sales obtained from the computer selling. Also interpret the results.
   b) Estimate the sales corresponding to advertising expenditure of Rs. 45 lakhs.

3. The following data represent the number of days absent of IT faculty per semester in a population of 4 faculties in an academic institute.

1,	**3,**	**6,**	**7**

   a) Select all possible samples of size n = 2 with replacement, and construct the sampling distribution of mean.
   b) Compare the population mean and mean of all sample means. Are they equal?
   c) Compare the shape of the population data and shape of the sampling distribution. Do you find any differences? Comment.
   d) Compare the population standard deviation and standard deviation of sample means and explain your observation.

Group B

Attempt any eight questions. **[8×5 = 40]**

4. Explain the differences between ordinal and interval scales of measurement with suitable examples.

5. The following are the two regression lines:
$3X+2Y=26$ & $6X+3Y=31$
Compute the correlation coefficient between them and interpret the result.

6. Following table represents the probability distribution for the number of computers crashes monthly in a reputed software company in Biratanagar.

Number of computer crashes	**0**	**1**	**2**	**3**	**4**	**5**
Prob(X=x)	0.10	0.20	0.45	0.15	0.05	0.05

Compute mean and standard deviation of number of computer crashes and interpret them.

7. A set of final examination grades in Basic Statistics, is following normal distribution with mean of 73 and standard deviation of 8.
   a) What is the probability of a student secured less than 93 marks?
   b) What is the probability of a student secured marks between 65 and 89?

8. The rate of denying to take vaccine for COVID19 in a rural population of India is reported to be 0.45 per 10,000 people. If the distribution of denying follows Poisson distribution, what is the probability that in the next 10,000 people, there will be:
   a) No one will deny to take vaccine?
   b) At least two persons will deny to take vaccine?

9. Suppose Rajesh receives 50 messages and Harish receives 90 messages in the personal emails respectively. Please note that the email address of Rajesh and Harish is different. Rajesh receives 1% junk emails, and Harish receives 2% junk emails. A person is chosen at random at the end of a day and found the email message is junk. What is the probability that this junk email found in the Harish's email inbox?

10. The standard deviation of a symmetric distribution is 7. Compute the possible value of fourth central moment for the distribution to be (i) mesokurtic (ii) platykurtic, and (iii) leptokurtic.

11. Assuming that the population is normally distributed, construct 95% confidence interval for the population mean using the following sample data.

1, 2, 3, 4, 5, 20

Again in the same data set, replace the value of 20 by 6, then compute the confidence interval for the mean. Explain why there is considerable difference in the confidence interval?

12. Write short notes on the following:
    a) Box and whisker plot
    b) Choice of appropriate measure of central tendency