Micro Syllabus for Statistics (B.Sc. CSIT) program

Tribhuvan University
Institute of Science & Technology(IOST)

Level: B.Sc.	
Course Title: Statistics I	Full Marks: 60 + 20 + 20
Course Code: STA 164	Pass Marks: 24 + 8 + 8
Nature of the Course: Theory and Practical	Credit Hrs : 3

Course objectives:

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

1. Introduction [4]

Basic concept of statistics: Definitions, descriptive and inferential statistics, limitations of statistics; Application of Statistics in the field of Computer Science & Information technology; Scales of measurement: nominal, ordinal, interval and ratio scale; Variables: qualitative, quantitative, discrete and continuous variables; Types of Data: primary data, secondary data and their sources, cross-sectional data, time series data, failure time data, panel data; Notion of a statistical population: finite population, infinite population, homogeneous population and heterogeneous population.

2. Descriptive statistics [6]

Measures of central tendency: mean, combined arithmetic mean, weighted mean, median, mode, partition values; Measures of dispersion: absolute and relative measures of dispersion, range, quartile deviation, mean deviation, standard deviation, combined standard deviation, coefficient of variation; Measure of skewness: Bowley's coefficient of skewness, Karl Pearson's coefficient of skewness; Measure of kurtosis: leptokurtic, mesokurtic and platykurtic frequency distributions, measures of kurtosis using partition values; Moments: raw and central moments, measures of skewness and kurtosis based on moments; steam and leaf display; five number summary; box plot; normal probability plot

Problems and illustrative examples related to computer Science and IT

3. Introduction to Probability [8]

Concepts of probability; Deterministic and random experiments; Basic terminology: trial and event, outcome, sample space, equally likely, mutually exclusive, exhaustive and favorable cases, sure and impossible events, independent and dependent events; Definitions of probability: mathematical (classical), statistical (relative frequency) and subjective definition; Laws of probability: Additive and multiplicative theorems; Conditional probability; Bayes theorem; prior and posterior probabilities.

Problems and illustrative examples related to computer Science and IT

4. Sampling [3]

Definitions of population, sample survey vs. census survey, sample, sampling unit, sampling frame, sampling error and non sampling error; Steps in sampling; Fundamental characteristics, advantages and disadvantages of sampling; Types of sampling: Probability (simple, stratified, systematic, cluster and multistage); non probability sampling (convenience, purposive, quota, snowball).

5. Random Variables and Mathematical Expectation [5]

Concept of a random variable; Types of random variables: Discrete and continuous random variables; Probability distribution of a random variable: probability mass function and probability density function, distribution function and its properties; Mathematical expectation of a random variable (discrete and continuous); Properties of mathematical expectation of random variables; Addition and multiplicative theorems of expectation; Concept of conditional expectations; variance, standard deviation and covariance; Properties of covariance and variance(without derivation)

Problems and illustrative examples related to computer Science and IT

6. **Probability Distributions** [12]

Concept of Probability distribution function**,** Joint probability distribution of two random variables: Joint probability mass function and density functions, marginal probability mass and density function; Conditional distribution functions; statistical independence; Discrete distributions: Bernoulli trial, Binomial and Poisson distributions, their mass functions, mean and standard deviation of their distribution, fitting of binomial and Poisson distribution(without derivation); Continuous distribution: Normal distributions, it’s probability density functions, measurement of area under the normal curve; Standardization of normal distribution; Normal distribution as an approximation of Binomial and Poisson distribution(without derivation); Exponential, Gamma distribution, measure characteristics(without derivation) and their application in relevant areas.

Problems and illustrative examples related to computer Science and IT

7. Correlation and Linear Regression [7]

Bivariate data; Bivariate frequency distribution; Correlation between two variables; Positive correlation; Negative correlation; Scatter diagram to explore the type of correlation; Karl Pearson’s coefficient of correlation (r): Definition, computation for grouped and ungrouped data and interpretation; Properties of correlation coefficient; Spearman’s rank correlation including tied cases; Regression Analysis: Concept of regression, lines of regression, fitting of lines of regression by the least squares method, interpretation of slope and intercept, concept of linearity; properties of regression coefficient; explained and unexplained variation, residual analysis; coefficient of determination.

Problems and illustrative examples related to computer Science and IT

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 Microsoft Excel, SPSS, STATA etc. whichever convenient).	**No. of practical problems**
1	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
2	Computation measures of dispersion (ungrouped and grouped data) and computation of coefficient of variation.	1
3	Measures of skewness and kurtosis using method of moments, Measures of Skewness using Box and whisker plot, normal probability plot	2
4	Scatter diagram, correlation coefficient (ungrouped data) and interpretation. Compute manually and check with computer output.	1
5	Fitting of lines of regression (Results to be verified with computer output)	1
6	Fitting of lines of regression and computation of correlation coefficient, Mean residual sum of squares, residual plots.	1
7	Conditional probability and Bayes theorem	3
8	Obtaining descriptive statistics of probability distributions	2
9	Fitting probability distributions in real data (Binomial, Poisson and Normal)	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/First Year/Second Semester/Science — Full Marks: 60

Computer Science and Information Technology STA 164 — Pass Marks: 24

(Statistics I) — 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. A new computer program consists of two modules. The first module contains an error with probability 0.2. The second module is more complex; it has a probability of 0.4 to contain an error, independently of the first module. An error in the first module alone causes the program to crash with probability 0.5. For the second module, this probability is 0.8. If there are errors in both modules, the program crashes with probability 0.9. Suppose the program crashed. What is the probability of errors in both modules?

2. Explain how box-plot is helpful to know the shape of the data distribution. The following data set represents the number of new computer accounts registered during ten consecutive days.

43	**37**	**50**	**51**	**58**	**105**	**52**	**45**	**45**	**10**

a) Compute the mean, median, quartiles, and sample standard deviation.

b) Check whether there are outliers or not.

c) If outliers are present, then delete the detected outliers and compute the mean, median, quartiles, and sample standard deviation again.

d) Make your conclusion about the effect of outliers on descriptive statistical analysis.

3. A computer manager interested to know how efficiency of his/her new computer program which depends on the size of incoming data. Efficiency will be measured by the number of processed requests per hour. In general, larger data sets require more computer time, and therefore, fewer requests are processed within 1 hour. Applying the program to data sets of different sizes, the following data were gathered.

Data size(gigabytes)	**6**	**7**	**7**	**8**	**10**	**10**	**15**
Processed requests	40	55	50	41	17	26	16

a) Identify which one response variable, and fit a simple regression line, assuming that the relationship between them is linear.

b) Interpret the regression coefficient with reference to your problem.

c) Obtain coefficient of determination, and interpret this.

d) Based on the fitted model in (a), predict the efficiency of new computer for data size 12(gigabytes). Does it possible to predict efficiency for data size of 30 gigabytes? Discuss.

Group B

Attempt any Eight questions (8 × 5 = 40)

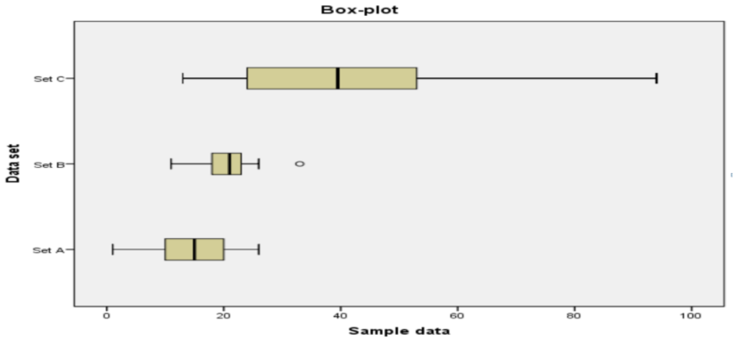
4. Explain the role of statistics in computer science and information technology.

5. Following table presents some descriptive statistics computed from three different independent sample dataset(X).

Data	Sample size (n)	$\sum_{i=1}^{30} X_i$	Minimum	Q1	Median	Q3	Maximum	$\sum_{i=1}^{30}(X_i - \bar{X})^2$
Data set A	30	439	1	10	15	20	26	1348.97
Data set B	30	625	11	18	21	23	33	540.17
Data set C	30	1239	13	24	39.5	53	94	12836.3

a) Compare sample mean and median, and explain about the shape of the data distribution for each dataset. Compare the variability of the three set of dataset.

Box-plots have been generated through SPSS for each dataset as follows.



b) Do these box-plots support your findings obtained in a) about the shape of the distribution? Explain.

6. A large chain retailer purchases a certain kind of electronic device from a manufacturer. The manufacturer indicates that the defective rate of the device is 3%.

a) The inspector randomly picks 20 items from a shipment. What is the probability that there will be at least one defective item among these 20?

b) Suppose that the retailer receives 10 shipments in a month and the inspector randomly tests 20 devices per shipment. What is the probability that there will be exactly 3 shipments each containing at least one defective device among the 20 that are selected and tested from the shipment?

7. Messages arrive at an electronic message center at random times, with an average of 9 messages per hour.

a) What is the probability of receiving at least five messages during the next hour?

b) What is the probability of receiving exactly seven messages during the next hour?

8. The time, in minutes, it takes to reboot a certain system is a continuous variable with the density function:

$$f(x) = \begin{cases} C(10-x)^2,\ 0 < \text{x} < 10 \\ 0\text{ , Otherwise} \end{cases}$$

Compute C, and then compute the probability that it takes between 1 and 2 minutes to reboot.

9. Following data represent the *preference of 10 students studying B.Sc .(CSIT) towards two brands of computers namely DELL and HP.*

Computer	**Student preference**									
DELL	5	2	9	8	1	10	3	4	6	7
HP	10	5	1	3	8	6	2	7	9	4

Apply appropriate statistical tool to measure whether the brand preference is correlated. Also interpret your result.

10. Define exponential distribution with parameter λ . The time required to reach to the printer after ordering in the computer follows exponential distribution at an average rate of 3 jobs per hour.

a) What is the expected time between jobs?

b) What is the probability that the next job is sent within 5 minutes?

11. The lifetime of a certain electronic component is a normal random variate with the expectation of 5000 hours and a standard deviation of 100 hours. Compute the probabilities under the following conditions

a) Lifetime of components is less than 5012 hours

b) Lifetime of components between 4000 to 6000 hours

c) Lifetime of components more than 7000 hours

12. Write short notes on the following.

a) Sampling error and non-sampling error
b) Conditional probability