Probability and Statistics
Csc. 103-2066

Group A

1. Define the following three measures of locations – mean, median and mode – and clearly state their properties. Write down a situation where mode is preferred to mean. Score obtained by 14 Students in a test are given below. Compute mean, median and mode.

42	39	45	55	38	35	60	55	55	65	40	43	35	37

Mean: - The arithmetic mean (A.M) is more popular than other means. It is helpful simply to analyze the data end in other further analysis of statistics.

Generally, arithmetic mean Is defined as the sum of observations divided by the no. of observations. The A.M can be divided into two parts. It is denoted by $(\overline{X})$

1) Simple arithmetic mean
2) Weighted arithmetic mean

Median: - Median is defines as the middle value of ordered data set which divided the distribution into two equal parts. The number of observations below the median and number of observations above the median is equal i.e, 50% observations lies on both side of the median. It is denoted by Md.

Mode:- Mode is the value having maximum frequency i.e, the value which is repeated more than other value. It is denoted by (M_0)

Here, given data are,

42 39 45 55 38 35 60 55 55 65 40 43 35 37

Arranging the data in ascending order, we get

35 35 37 38 39 40 42 43 45 55 55 55
60 65

Mean $(\overline{X}) = \frac{\sum fx}{N}$

Now, table for calculation:

x	f	fx
35	2	70
37	1	37
38	1	38
39	1	39
40	1	40

Source: www.csitnepal.com

42	1	42
43	1	43
45	1	45
55	3	165
60	1	60
65	1	65
	N=14	∑fx=644

Mean ($\overline{X}$) = $\frac{\sum fx}{N}$ = $\frac{644}{14}$ = 46

Median (Md) = $(N+1)/2^{th}$ value

= 14/2

= 7.5 th value

Here, 7.5the value must be average of 43 and 45, so

= (43+45)/2

= 44

Mode (M_0) = the value of maximum frequency =55

2. Explain the terms- sample space and events of a random experiment. State the classical and the statistical definition of probability. Which of the two definitions is most useful in statistics and why? A survey of 300 families was conducted to study income level versus brand preference. The data are summarized below:

	Brand			
Income Level	Brand 1	Brand 2	Brand 3	Total
High	55	45	20	120
Medium	45	25	25	95
Low	25	35	25	85
Total	125	105	70	300

If a family is selected at random, then compute the probability that (a) the family belongs to high income group (b) the family prefers brand 3, and (c) the family belongs to the low income group and prefers Brand 3.

Sample Space:-

The set of possible outcomes of a random experiment is called sample space. It is denoted by capital letter S.

Example in tossing a coin once, the sample space is S = {H.T}

Event:-

The result of an experiment is called events. In coin-toss experiment getting a head is one event and getting a tail is another event. Similarly, in throwing a dice, getting one of the face 1, 2… 6 is an event.

Classical Approach or Mathematical Approach:-

If there are ‘n’ exhaustive, mutually exclusive and equally likely event out of which ‘m’ event are the occurrence of event ‘A’ (say) then the probability of occurrence of event A is denoted by P(A) and is defines as P(A)= m/n , m≤n

$= \frac{Favourable\ no.of\ cases}{Exhaustive\ of\ tatal\ no.of\ cases}$

Statistical Approach:-

If an experiment (trial) is replaced under similar condition for a large number of times then the limiting value of ratio of number of time the occurrence of event A(say) to the total number trials, provided the trails are indefinitely large and limit is unique and finite is known as the probability if occurrence of event A.

Mathematically,

$P\ (A) = \lim_{n\to\infty} \frac{m}{n}$

a) Here,

Total families (n) =300

Family of high income group (m) =120

Probability of a family belonging to high income group

$P\ (n) = \frac{m}{n} = \frac{120}{300} = 2/5$

b) Total family (n) = 300

Family prefer Brand 3 (m) =70

$P\ (B_3) = \frac{m}{n} = \frac{70}{300} = \frac{7}{30}$

c) Total family (n) = 300

Family preferred Brand 3 (m) =25 (low Income)

$P\ (B_3) = \frac{m}{n} = \frac{25}{300} = \frac{1}{12}$

3. Make a clear distinction between correlation and co-efficient and slope regression co-efficient. A school teacher believes that there is a linear relationship between the verbal test

score (Y) for eighth graders and the number of library books checked out (X). Following are the data collected on 10 students:

X	12	15	3	7	10	5	22	9	13	7
Y	77	85	48	59	75	41	94	65	79	70

The above data reveal the following statistics:

$\sum XY = 7881\ (\sum X) = 103\ (\sum Y)$=693 $\sum X^2 = 1335\ (\sum Y)^2 = 50447$

a) Compute the correlation co-efficient r between X and Y. Interpret the meaning of r^2.

b) Fit a simple linear regression model of Y on X using the least square method. Interpret the estimated slope regression coefficient.

Main point that directly differentiates between correlation coefficient and regression coefficient is that correlation coefficient is the geometric mean between two regression co-efficients.

Regression coefficient gives the slope of regression line while correlation coefficient gives the relationship. The arithmetic mean of regression coefficient is greater than the correlation coefficient.

$\left(\frac{b_{xy}+b_{yx}}{2}\right) \geq$ r

Where,

b_{yx} and b_{xy} are regression coefficient and r is correlation coefficient.

a) Correlation coefficient

r = $\frac{\text{n}\sum\text{XY}-\ (\sum\text{X})\ (\sum\text{Y})}{\sqrt{n\sum x^2-(\sum x)^2}\ -\sqrt{n\sum y^2-(\sum y)^2}}$

= $\frac{10\times7881-(103)\ (693)}{\sqrt{10\times1335-(103)^2}\ -\sqrt{10\times50447-(693)^2}}$

= $\frac{7431}{8147.23}$

= 0.912

Here, coefficient of determination,

r^2. = 0.912^2 = 0.831744

This means 83.1 % of the score is due to the books and remaining is due to other factors.

b) Here,

b_{yx} =$\frac{\text{n}\sum\text{XY}-\ (\sum\text{X})\ (\sum\text{Y})}{n\sum\text{X}^2-\ (\sum\text{X})^2}$

= $\frac{10\times7881-(103)\ (693)}{10\times1335-\ (103)^2}$

$= \frac{7431}{2741}$

= 2.71

Now,

$a = \frac{\sum Y - b_{yx}\sum X}{n}$

$= \frac{693 - 2.71 \times 103}{10}$

= 41.387

Therefore, required average

Y= 41.387 + 2.71 X

b_{yx}=2.71 is the estimates slope regression as it is positive it is moving upward.

Group B

4. State with suitable examples the role played by computer technology in applied statistics and also the role of statistics in Information technology.

5. Define discrete and continuous random variables with suitable examples. A continuous random variable X has the following density function.

   $f(x) = \begin{cases} kx(1-x) & \textit{for } 0 < x < 1 \\ 0 & \textit{elsewhere} \end{cases}$

   Find the value of k so that the total probability would be 1. Also find E(X).

   Discrete random variable

   If a random variable can take exact or whole number of numerical values then it is called discrete random variable. For example: the number of students in a class, the number of books in a library.

   X=2, 5, 9, 15, 22

   Continuous random variable

   If a variable can take all the possible values with a certain range or interval is known as continuous random variable. For e.g. height, weight, length, breadth etc.

   20<x<25

   We know,

$\int_{-\infty}^{\infty} f(x)dx = 1$

$\int_0^1 kx(1-x)dx = 1$

$\int_0^1 kx - kx^2 dx = 0$

$k\left[\int_0^1 x - \int_0^1 x^2\right]dx$

$\left[\frac{kx^2}{2}\right]_0^1 - \left[\frac{kx^3}{3}\right]_0^1 = 1$

$\frac{k}{2} - \frac{k}{3} = 1$

$\frac{3k-2k}{6} = 1$

Therefore,

k= 6

Now, we know

$E(X) = \int_0^1 xf(x)dx$

$= \int_0^1 xkx(1-x)dx$

$= k\int_0^1 x^2 - x^3 dx$

$= 6\left(\int_0^1 x^2 dx - \int_0^1 x^3 dx\right) \quad [\because k = 6]$

$= 6\left(\left[\frac{x^3}{3}\right]_0^1 - \left[\frac{x^4}{4}\right]_0^1\right)$

$= 6\left(\frac{1}{3} - \frac{1}{4}\right)$

$= 6\left(\frac{4-3}{4}\right)$

$= \frac{3}{2}$

6. Assume that the two continuous random variables X and Y have the following density function

$$f(x) = \begin{cases} \frac{6-x-y}{8}, & 0 < x < 2, 2 < y < 4 \\ 0, & elsewhere \end{cases}$$

F(x)=$\frac{6-x-y}{8}$ for 0<x<2, 2<y<4

$=\frac{1}{8}\int_0^2\int_2^4 (6-x-y)\,dy\,dx$

$=\frac{1}{8}\int_0^2 \left[(6y-xy-\frac{y^2}{2}\right]_2^4\, dx$

$=\frac{1}{8}\int_0^2 \left[\left(6.4-x.4-\frac{4^2}{2}\right)-\left(6.2-x.2-\frac{2^2}{2}\right)\right]dx$

$=\frac{1}{8}\int_0^2 (24-2x-8-12+2x+2)dx$

$=\frac{1}{8}\int_0^2 (6-2x)\,dx$

$=\frac{1}{8}[6x-\frac{2x^2}{2}]_0^2$

$=\frac{1}{8}\left[6\,.\,2-2\,.\frac{2^2}{2}\right]$

$=\frac{1}{8}[8]$

=1

7. In a binomial distribution with parameters n and p, prove that mean and variance in binomial distribution are correspondingly np and npq, where q = 1-p.

Now, for mean:

Mean = E(x)

=∑x.P(X=x)
$= \sum_{x=0}^{n} x\ {}_{x}^{n}C\ p^x q^{n-x}$
$= \sum_{x=0}^{n} x\ \frac{n!}{(n-x)!x!}\, p^x q^{n-x}$
$=\sum_{x=1}^{n} x\ \frac{n!}{(n-x)!x(x-1)!}\, p^x q^{n-x}$
$=\sum_{x=1}^{n} x\ \frac{n(n-1)!}{(n-x)!(x-1)!}\, p^x q^{n-x}$
$= n\sum_{x=1}^{n} x\ \frac{(n-1)!}{(n-x)!(x-1)!}\, p^x q^{n-x}$
=n $\sum_{x=1}^{n} {}_{x-1}^{n-1}C\ p^x q^{n-x}$
=np $\sum_{x=1}^{n} {}_{x-1}^{n-1}C\ p^{x-1} q^{n-x}$
=np$(p+q)^{n-1}$

$(a+b)^n$ = a^n + ${}_{1}^{n}c\ a^{n-1}\ \ b^1$ + ${}_{2}^{n}c\ a^{n-2}\ b^2$ + ……..
= n $\sum_{x=1}^{n} {}_{x-1}^{n-1}C\ p^{x-1} q^{n-x}$
=${}_{0}^{n-1}C\ p^0 q^{n-1}$ + ${}_{1}^{n-1}C\ p^1 q^{n-2}$ + ${}_{2}^{n-1}C\ p^2 q^{n-3}$ + …….. +${}_{n-1}^{n-1}C\ p^{n-1} q^{n-n}$
= q^{n-1} + ${}_{1}^{n-1}C\ p^1 q^{n-2}$ + ${}_{2}^{n-1}C\ p^2 q^{n-3}$+ p^{n-1}
= $(q+p)^{n-1}$

Therefore, E(x) = np [since, p+q =1]

For variance,

E(X^2) = $\sum x^2$p (X=x)

= $\sum_{x=0}^{n} x^2 \; {}^{n}_{x}C \; p^x q^{n-x}$

= $\sum_{x=0}^{n}\{x(x-1)+x\}. {}^{n}_{x}C \; p^x q^{n-x}$

=$\sum_{x=0}^{n} x(x-1). {}^{n}_{x}C \; p^x q^{n-x}$ + $\sum_{x=0}^{n} x. {}^{n}_{x}C \; p^x q^{n-x}$

= $\sum_{x=0}^{n} x(x-1). \frac{n!}{(n-x)!x!} \; p^x q^{n-x}$ + E(x)

=$\sum_{x=2}^{n} x(x-1). \frac{n!}{(n-x)!x(x-1)(x-2)!} \; p^x q^{n-x}$ + np

= $\sum_{x=2}^{n} \frac{n(n-1)(n-2!}{(n-x)!(x-2)!} \; p^x q^{n-x}$ + np

=$n(n-1)\sum_{x=2}^{n} \frac{(n-2!}{(n-x)!(x-2)!} \; p^x q^{n-x}$ + np

= $n(n-1)p^2 \sum_{x=2}^{n} \; {}^{n-2}_{x-2}C \; p^{x-2} q^{n-x}$ + np

= $n(n-1)p^2 \; (p+q)^{n-2}$ + np

= $n(n-p\,)^2 \times 1$ + np [since, p+q=1]

E(X^2) = $n(n-1)p^2$ + np

Now,

Variance,

V(X) = $\sum(x^2) - [E(x)]^2$

= $n(n-1)p^2$ + np – $(np)^2$

= n^2p^2 - p^2n + np -n^2p^2

= np - p^2n

= np(1-p)

Therefore,

V(X) = npq ……(ii) [since, p+q=1]

8. The systolic blood pressure of 18 years old women(X) is normally dist ributed with a mean of 120 mm Hg and a standard deviation of 12 mm Hg randomly selected 18 years old women. Compute the following probabilities:
(a) P(X>150), (b) P(X<115) (c) P(110<X<130)

Solution,

Given, let x be the normal variate that follows normal distribution with mean(μ) = 120 & standard deviation(σ) = 12

(a)For P(X>150)

When x=180

$$z = \frac{x - \mu}{\sigma}$$

$$= \frac{150 - 120}{12}$$

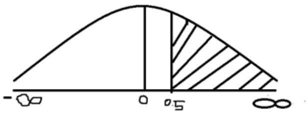
$$= \frac{30}{12}$$

$$= 2.5$$

P(x>150)=p(z>2.5)

=p(0<z<∞)-P(0<z<2.5)

=0.50-0.4938



=0.006200

(b)P(x<115)

Here,

x=115

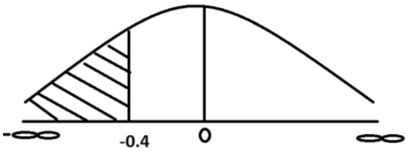
z=$\frac{x-\mu}{\sigma}$

=$\frac{115-120}{12}$

=$\frac{-5}{12}$

=-0.4166667

Now,



P(x<115)=P(z<-0.4166667)

=p(-∞<z<0)-(0<z<0.4166667)

=0.50-0.159

=0.340900

(c)P(110<x<130)

Here,

X=110

z=$\frac{x-\mu}{\sigma}$

=$\frac{110-120}{12}$

$=\frac{10}{12}$

=0.833

Again,

X=130

$z=\frac{x-\mu}{\sigma}$

$=\frac{130-120}{12}$

$=\frac{10}{12}$

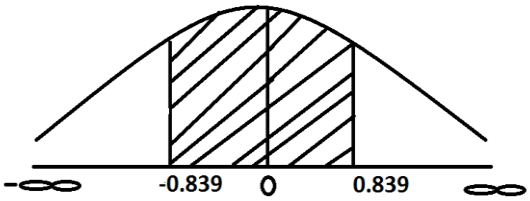
=0.833

Now,

P(110<x<130)=p(-0.833<z<0.833)

=p(-0.833<z<0)+p(0<z<0.833)

=p(0<z<0.833)+p(0<z<0.833)



=0.2967+0.2967

=0.59340

9. If X_1, …..,X_n are n independent random variables each is distributed as normal with mean μ and variance σ^2, then derive the distribution of $\sum_{i=1}^{n} x_i$.
10. Write the density function of negative exponential distribution, and derive its mean and variance.
11. Obtain the maximum likelihood function of n independent random sample drawn from the normal population with unknown mean μ and unknown variance $\sigma^2$and, using the principle of maximum likelihood method of estimation derive the estimators of μ and σ^2.
12. A survey of 100 percents of first and second grade children revealed that the number of hours per week their children watch television (X) had an average of 25.8 hours and standard deviation of 4.0 hours. The problem is to determine whether there is statistical evidence to conclude that μ(population mean of X) exceeds 25 hours. Set up appropriate null and alternative hypothesis and carry out appropriate test at 5% level of significance.

13. A standardized psychology exam has mean of 70. A research psychologist wished to see whether a particular drug had a effect on performance on the exam. He administered exam to 18 volunteers who had taken the drug and obtained the following scores:

    68, 71,71,65,64,70,64,71,73,62,78,70,69,76,67,69,72 which yielded $\overline{X}$= 69.4444 and s^2= 16.8497. The problem is to determine whether there is statistical evidence suggestion that taking drug reduces one's score on the exam. Set up appropriate null and alternative hypothesis and carry out the test at 5% level.

For t-test

H_0= μ1 = μ2

Vs H_1 = μ1≠ μ2

Test statistics is ,

$$T= \frac{\overline{X}-\mu}{\mathrm{s}/\sqrt{\mathrm{n}-1}},$$

$$= \frac{70-69.4444}{\sqrt{16.8497}/\sqrt{18-1}}$$

= 0.5581

T_{tab}= $T_{17,0.05}$ = 1.740

T_{cal} < T_{tab}we should accept H_0 and reject H_1 with the conclusion that there is significant difference.