
# Shahjalal University of Science &Technology, Sylhet

Department of Computer Science and Engineering

Course No: STA-202D

Assignment No: 01

**Statistics and Probability**

Submitted To

Dr S M Khurshid Alam

*Professor*

*Department of Statistics*

Submitted By

Name : Jakir Hasan

Registration no : 2018331057

Section : A

Session : 2018-19

Submission date: 09/12/2020

# Solution

## Group-01

**Question: 01**

**a.**

A scale is a device on an object used to measure or quantify any event on another object.

There are four different scales of measurement.

Nominal scale:

A nominal scales is usually deals with the non-numeric variables on the numbers that do not have any value.

Ordinal scale:

It reports the ordering and ranking of data without establishing the degree of variation between them.

Interval scale:

It is defined as a quantitative measurement scale in which difference between two variables is meaningful.

Ratio scale:

It allows researchers to compare the differences on intervals.

b)

i.

| Stem | Leaf |
|---|---|
| 0 | 2, 3, 5, 5, 6, 6, 7, 7, 7, 8, 8, 8, 8, 9, 9, 9, 9, 9, 9 |
| 1 | 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 4, 4, 4, 4, 4, 5, 5, 5, 5, 6, 6, 8, 9 |
| 2 | 0, 0, 2, 3 |

ii.

| Stem | Leaf |
|---|---|
| 0 | 2, 3 |
| 0 | 5, 5, 6, 6, 7, 7, 7, 8, 8, 8, 8, 9, 9, 9, 9, 9, 9 |
| 1 | 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 4, 4, 4, 4, 4 |
| 1 | 5, 5, 5, 5, 6, 6, 8, 9 |
| 2 | 0, 0, 2, 3 |
| 2 | |

III. Given,

Total number of observation, $N = 56$

Lowest observation $= 2$

Highest observation $= 23$

Range $= 23 - 2$

$= 21$

Number of classes, $K = 1 + 3.322 \log N$

$= 1 + 3.322 \log 56$

$= 6.80 \approx 6$

Class interval, $h = \frac{21}{6}$

$= 3.5 \approx 4$

The frequency table is:

| Class Interval | Tally marks | Frequency |
|---|---|---|
| 2-5 | IIII | 4 |
| 6-9 | 𝍸 𝍸 𝍸 | 15 |
| 10-13 | 𝍸 𝍸 𝍸 𝍸 | 20 |
| 14-17 | 𝍸 𝍸 I | 11 |
| 18-21 | IIII | 4 |
| 22-25 | II | 2 |

# Question - 02

a.

- Central tendency is a descriptive summary of a dataset through a single value that reflects the centre of data distribution.

  The different measures of central tendency are:

  i. Mean
  ii. Median
  iii. Mode.

- The arithmetic mean is the sum of all the numbers in a data set divided by the quantity of numbers in the set.

  Arithmetic mean is so popular because-

  i. It is easy to calculate and understand arithmetic mean.

  ii. Arithmetic mean is based on all the values of the series.

• The circumstances to use median and mode as a suitable measure of central tendency are:

The median is usually prefferred to other measures of central tendency when the data set is skewed or the data is ordinal.

The mode is the least used measure of central tendency and can only be used when dealing with nominal data.

(b)

| Class Interval | Mid values($x_i$) | Frequency($f_i$) | $f_i x_i$ |
|---|---|---|---|
| 4-6 | 5 | 75 | 375 |
| 6-8 | 7 | 115 | 805 |
| 8-10 | 9 | 80 | 720 |
| 10-12 | 11 | 50 | 550 |
| 12-14 | 13 | 25 | 325 |
| | | $\Sigma f_i = 345$ | $\Sigma f_i x_i = 2775$ |

$$\text{Mean} = \frac{\Sigma f_i x_i}{\Sigma f_i}$$

$$= \frac{2775}{345}$$

$$= 8.04$$

$$\text{Mode} = L + \frac{f_1 - f_0}{(f_1 - f_0) + (f_1 - f_2)} \times c$$

$$= 6 + \frac{115 - 75}{(115 - 75) + (115 - 80)} \times 2$$

$$= 7.067$$

$L = 6$
$f_1 = 115$
$f_2 = 80$
$f_0 = 75$
$c = 2$

## Question-03

(a)

The standard deviation is a statistic that measures the dispersion of a data set relative to its mean and is calculated as the square root of variance.

The mean deviation is measure as a statistical measure which is used to calculate the average deviation from the mean value of the given dataset.

Difference between Mean deviation and standard deviation:

| Mean Deviation | Standard deviation |
|---|---|
| 1. In calculating mean deviation algebraic is signs are ignored. | 1. In calculating standard deviation, algebraic signs are taken into account. |
| 2. Mean or median is used in calculating mean deviation. | 2. Only mean is used in calculating standard deviation. |

(c)

| class interval | Mid value ($x_i$) | Frequency ($f_i$) | $f_i x_i$ | $f_i x_i^2$ |
|---|---|---|---|---|
| 0-5 | 2.5 | 7 | 17.5 | 43.75 |
| 5-10 | 7.5 | 26 | 195 | 1462.5 |
| 10-15 | 12.5 | 59 | 737.5 | 9218.75 |
| 15-20 | 17.5 | 71 | 1242.5 | 21743.75 |
| 20-25 | 22.5 | 62 | 1395 | 31387.5 |
| | | $\Sigma f_i = 225$ | $\Sigma f_i x_i = 3587.5$ | $\Sigma f_i x_i^2 = 63856.25$ |

$$\text{Variance} = \frac{\Sigma f_i x_i^2}{\Sigma f_i} - \left(\frac{\Sigma f_i x_i}{\Sigma f_i}\right)^2$$

$$= \frac{63856.25}{225} - \left(\frac{3587.5}{225}\right)^2$$

$$= 29.5802$$

$$\text{Standard deviation} = \sqrt{\text{variance}} = \sqrt{29.5802}$$

$$= 5.4387$$

$$\text{Coefficient of variation} = \frac{SD}{AM} \times 100$$

$$= \frac{5.4387 \times 225}{3587.5} \times 100$$

$$= 34.11$$

## Question-04

### (a)

- Correlation may be defined as the degree of relationship existing between two or more variables

- The correlation co-efficient is a statistical measure of the strength of the relationship between the relative movements of two variables.

- The difference between correlation and regression are:

  i. Correlation is a single statistic, or data point, whereas regression is the entire equation with all of the data points that are represented with a line.

  ii. Correlation shows the relationship between two variables, while regression allows us to see how one affects the other.

(b)

Positive Correlation:

Two variables are said to be positively correlate if they tend to change in the same direction.

Negative Correlation:

Two variables are said to be negatively correlated if then tend to change in the opposite direction.

Interpretation of correlation co-efficient:

i. $r = +1$. This implies that two variables are perfectly correlated. The relation two variables is positive.

ii. $r = -0.75$. This implies that there exists a strong relationship between two variables and the relationship is negative.

iii. $r = -1$. This implies that two variables is perfectly correlated negatively.

iv. $r = 0$. This indicates there is no relationship between two variables.

(c)

| $X_i$ | $Y_i$ | $X_iY_i$ | $X_i^2$ | $Y_i^2$ |
|---|---|---|---|---|
| 1.5 | 1.7 | 2.55 | 2.25 | 2.89 |
| 2.0 | 1.32 | 2.64 | 4 | 1.7424 |
| 2.5 | 1.48 | 3.7 | 6.25 | 2.1904 |
| 3.0 | 1.0 | 3 | 9 | 1 |
| 3.5 | 1.41 | 4.935 | 12.25 | 1.9881 |
| 4.0 | 1.19 | 4.76 | 16 | 1.4161 |
| 4.5 | 1.23 | 5.535 | 20.25 | 1.5129 |
| 5.0 | 0.8 | 4 | 25 | 0.64 |
| 5.5 | 0.59 | 3.245 | 30.25 | 0.3481 |
| 6.0 | 0.35 | 2.1 | 36 | 0.1225 |
| 6.5 | 0.61 | 3.965 | 42.25 | 0.3721 |
| $\Sigma X_i = 44$ | $\Sigma Y_i = 11.68$ | $\Sigma X_iY_i = 40.43$ | $\Sigma X_i^2 = 203.5$ | $\Sigma Y_i^2 = 14.2226$ |

So, $\bar{X} = 4$ , $\bar{Y} = 1.0618$

Correlation coefficient $r = \dfrac{\Sigma X_iY_i - n\bar{X}\bar{Y}}{\sqrt{(\Sigma X_i^2 - n\bar{X}^2)(\Sigma Y_i^2 - n\bar{Y}^2)}}$

$= \dfrac{40.43 - 11 \times 4 \times 1.0618}{\sqrt{(203.5 - 11 \times 4^2)\{14.2226 - 11 \times (1.0618)^2\}}}$

$= -0.1256$

## Question-05

### (a)

A linear regression model describes the relationship between a dependent variable, $y$ and one or more independent variable, $x$.

There are four assumptions associated with linear regression model:

1. Linearity: The relationship between X and mean of Y is linear.
2. Homoscedasticity: The variance of residual is the same for any value of X.
3. Independence: Observations are independent of each other.
4. Normality: For any value of X, Y is normally distributed.

A bivariate distribution, put simply, is the probability that a certain event will occur when there is two independent random variables. For examples, having two bowls each filled with two different types of candies, and pulling one candy from each bowl gives you two independent random variables, the two different candies.

Since you are pulling one candy from each bowl at the same time, you have a bivariate distribution when calculating your probability of ending up with particular kinds of candies.

(c)

| $X_i$ | $Y_i$ | $X_iY_i$ | $X_i^2$ |
|---|---|---|---|
| 4.5 | 53 | 238.5 | 20.25 |
| 6.7 | 82 | 549.4 | 44.89 |
| 8.2 | 102 | 836.4 | 67.24 |
| 5.0 | 60 | 300 | 25 |
| 4.6 | 39 | 179.4 | 21.16 |
| 6.1 | 42 | 256.2 | 37.21 |
| 3.0 | 27 | 81 | 9 |
| $\Sigma X_i = 38.1$ | $\Sigma Y_i = 405$ | $\Sigma X_iY_i = 2440.9$ | $\Sigma X_i^2 = 224.75$ |

$\bar{X} = 5.44$ , $\bar{Y} = 57.857$

So, $b = \dfrac{\Sigma X_iY_i - n\bar{X}\bar{Y}}{\Sigma X_i^2 - n\bar{X}^2}$

$= \dfrac{2440.9 - 7 \times 5.44 \times 57.857}{224.75 - 7 \times (5.44)^2}$

$= 13.509$

$a = \bar{Y} - b\bar{X}$

$= 57.857 - 13.509 \times 5.44$

$= -15.63$

## Question-06

### (a)

**Random experiment:**

A random experiment is a process by which we observe something uncertain.

**Favorable outcome:**

The result which is desired is called favorable outcome

**Mutually exclusive outcome:**

If the occurance of one outcome supersedes the other, that is called mutually exclusive outcome.

**Equally likely outcomes:**

The outcomes of a sample space are called equally likely if all of them have the same chance of occurring.

(b)

- In an experiment, the probability of an event is the likelyhood of that event occuring.

- Addition Law:

When two events A and B are mutually exclusive, the probability that A or B will occur is the sum of the probability of each event.

$$P(A \text{ or } B) = P(A) + P(B)$$

Proof:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$= P(A) + P(B) \qquad [P(A \cap B) = \phi]$$

When three coins are tossed simultaneously,

Sample space = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}

Total numbers of outcome are = $2^3$
= 8

i.

The probability of no head occurs = $\frac{1}{8}$

ii.

The probability of two or more head occurs,

$= \frac{4}{8}$

$= \frac{1}{2}$

## Question-07

### (a)

- The binomial distribution is a probability distribution that summarizes the likelihood that a value will take one of two independent values. under a given set of parameters or assumptions.

The conditions for binomial distribution are:

1. Fixed number of trials
2. Independent trials
3. Two different classification
4. The probability of success stays the same for all trials.

(b)

Normal distribution is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.

The chief characteristics of normal distribution are:

i. Normal distributions are symmetric, unimodal and asymptotic.

ii. The mean, median and mode are equal.

iii. A normal distribution is perfectly symmetric around its centre.

iv. There is only one mode, or peak in a normal distribution.

## Question-08

### (a)

In statistics, a poisson distribution is a statistical distribution that shows how many times an event is likely to occur within a specified period on time.

### (b)

Queuing theory is the mathematical study of the congestion and delays of waiting in line. Queuing theory examines every components of waiting in line to be served, including the arrival process, service process etc.

The assumptions of queuing theory are:

i. The source population has infinite size.

ii. The inter-arrival time has an exponential

probability distribution.

iii. There is no unusual customer behavior.

iv. The service discipline is FIFO.

Poisson arrival rate:

The probability that one arrival occurs between t and t+deltat is $\lambda t + o(t)$, where $\lambda$ is a constant, independent of time t, and e independent of arrivals in either intervals. $\lambda$ is called the poisson arrival rate.

(c)

The difference between Markov chain and Markov process are:

Markov process is a random process in which the future is independent of the past, given the present. Thus, Markov processes are the natural stochastic analogs of the deterministic processes described by differential and difference equations.

A markov chain is a mathematical system that experiences transitions from one state to another according to certain probabilistic rules. The defining characteristics of Markov

chain is that no matter how the process arrived at its present state, the possible future states are fixed. In other word, the probability of transitioning to any particular state is dependent solely on the current state and time elapsed.