

**Group Report**

**Course Title: MSc in Data Analytics**

**Lecturer Name: Dr Shahram Azizi**

**Module Title: Statistics for Data Analytics**

**Assignment Title: CA One**

**Submitted by:**

**10634205 – Sandeep Gopu**

**10634610 - Anuradha Anuradha**

**10634920 – Anjali Mary Philip**

Table of Contents

[**Question 1** 3](#_Toc130852407)

[**Question 2** 9](#_Toc130852408)

[**Question 3** 22](#_Toc130852409)

[**MINUTES OF GROUP MEETING** 26](#_Toc130852410)

[**REFERENCES:** 28](#_Toc130852411)

# **Question 1**

1. **Describe the dataset using appropriate plots/curves/charts.**

After reviewing many datasets, we chose “Sleep in Mammals” dataset as it contains categorical and continuous variables.

First, we are reading the CSV file which includes our above dataset into the variable ‘data’ and when we print ‘data’ we can see the result which shows our dataset which includes 11 columns and 62 rows.

A screenshot of a computer

Description automatically generated with medium confidence

Describing the dataset using plots, curves, and charts:

Code:

Text

Description automatically generated

Plot:

Chart, histogram

Description automatically generated

Curve:

Histogram

Description automatically generated with low confidence

PieChart:

Chart, pie chart

Description automatically generated

BarPlot:

Chart, bar chart

Description automatically generated

1. **Consider one of the continuous attributes, and computer central and variational measures.**

Central measures:

The central measure is a scalar quantity that attempts to describe the whole data set using a single value representing the middle or center of its distribution.

The three main central measures are:

* Mean
* Median
* Mode

Mean: Mean is the average of a collection of data values.

Here, we want to check the Mean for the variable ‘body\_wt’. Hence, we are allocating that into a variable ‘X’ for calculating the mean. It delivers the mean of the variable.

Graphical user interface, text, application

Description automatically generated

Median: The middle value in a distribution is the median. In a distribution with odd values, the middle value is the median. In a distribution with even values, the mean of the two middle values is the median.

Graphical user interface, text

Description automatically generated

Mode: In a distribution, the value which has the highest frequency is the mode.

In the code, we are allocating the variable ‘exposure’ into ‘n’. First, we are finding the frequency of the values using the table() function. Then, using the mode function we found out that 1 has the highest frequency.

Text

Description automatically generated

Variational Measures

Variational measures represent how the data is spread out or dispersed.

The main variational measures are:

* Range
* Interquartile range
* Variance
* Standard deviation

Range: The difference between the smallest data and the largest data in a dataset is known as range.

Graphical user interface, text, application

Description automatically generated

Interquartile range: It is a variational measure which shows how data is spread around the mean.

Formula: IQR = Q3-Q1

Chart

Description automatically generated with medium confidence

Graphical user interface, application

Description automatically generated

Variance: It is the average of the squared differences from the mean.

Formula:

Text

Description automatically generated with medium confidence

Where,

S2 = variance

X = value of one observation

x = mean of all observation

n = number of values in the sample

Graphical user interface, text, application

Description automatically generated

Standard deviation: Standard deviation is a measure of how far the data is from the mean. It is the squared root of the variance.

Formula:

Diagram

Description automatically generated with low confidence

Graphical user interface, text, application

Description automatically generated

# **Question 2**

1. **Select four variables of the dataset, and propose an appropriate probability model to quantify the uncertainty of each variable.**

2. a) In this question, we have taken the insurance dataset, and below mentioned are the four variables that we will use in this question.

Moreover, we will implement a discrete and continuous probability model based on the type of variables.

* + 1. Bmi
    2. Children
    3. Region
    4. Charges

Also, below mentioned are the steps involved in this section.

Step 1: Read the insurance dataset using the read.csv(file.choose()) method as the file is available in .csv format i.e “insurance.csv”.

Table

Description automatically generated

Step 2: We are assigning all the rows of the selected four columns in four individual variables i.e Bi\_model, Mul\_model, Nor\_model, Exp\_model

Graphical user interface, text

Description automatically generated

Step 3: In this step, we will decide the probability model of the variables mentioned above. Below mentioned is the definition of the discrete and continuous probability model.

1. **Discrete probability model**: It is a probability distribution to quantify the uncertainty of a discrete random variable, which is a finite number of values. The corresponding probability distribution is called probability mass function i.e pmf
2. **Bernoulli model**: It is a model, in which a binary variable can be modeled and quantified
3. **Binomial model**: It is an extension of the Bernoulli model to the case where we have more than one binary experiment. i.e. the outcomes of a binomial random variable can be 0,1,2,...,n where n is the number of binary experiments.
4. **Multinomial model**: It is an extension of the Binomial model to the case where we have a multinary experiment. i.e the possible outcomes of the random variable can be more than two choices.
5. **Poisson model**: This model is an appropriate distribution to quantify any count variable
6. **Continuous probability model**: It is a probability distribution to quantify the uncertainty of a continuous random variable, which can take on any value in a given interval. The corresponding probability distribution is called the probability density function i.e pdf
7. **Uniform model**: It is a model, in which a continuous variable can be evenly modeled and quantified.
8. **Normal model**: This model can quantify any real-value variable.
9. **Exponential model**: This model can quantify any positive real-value variable.
10. **Gamma model**: The Gamma model is used to quantify a positive real value. This model is the extension of the exponential model into two parameters.

So, as per the above model definition and values of the columns shown in the snapshot attached in Step 1, we can conclude that

* Children -> Discrete Binomial variable
* Region -> Discrete Multinomial variable
* Charges -> Continuous Normal variable
* Bmi -> Continuous Exponential variable

2. b) In this section, after identifying the probability model we are moving ahead to estimate the parameters of each model, and below mentioned are the steps performed.

**Variable 1: Children:** As children is the discrete binomial variable, below mentioned are the formulas to calculate the R code, pmf and cdf of it.

* R code = rbinom(s,n,p)
* pmf = dbinom(j,n,p)
* cdf i.e P(x<=j) = pbinom(j,n,p)
* Expectation is E(x|p) = np
* Var(x|p) = np(1-p)

where s = sample size

j = possible outcome

n = number of binary trials

p = the probability of getting one

So, in this step we are assigning all the values of children column into ‘N’ variable and to calculate the number of binary trials we are using max() function and assigning the value to variable ‘n’. In addition, as per the above formula, probability is the result of expectation/n i.e mean/n and below attached are the result snapshot of this step.

Graphical user interface, application

Description automatically generated

Chart, line chart

Description automatically generated

Graphical user interface, chart

Description automatically generated

Chart, line chart

Description automatically generated

Graphical user interface, application

Description automatically generated

**Variable 2: Region:** As region is the discrete multinomial variable, below mentioned are the formulas to calculate the R code, pmf and cdf of it.

* R code = rmultinom(s,n,(p1,…,pk))
* pmf = dmultinom(j,n,(p1,…,pk)) If j=(j1,...,jk)
* Expectation is E(xi|pi) = npi
* Var(xi|pi) = npi(1-pi)
* Cov(xi,xj|pi,pj) = -npipj

where s = sample size

j = the value of the multinomial random variable come

n = number of binary trials

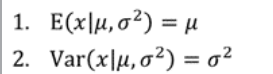
p = the probability of class i, i=1,…,k where k is the number of classes.

So, in this step, we are assigning all the values of the region column into ‘X’ variable and to calculate the frequency we are using table(X) function and assigning the value to variable ‘f’. In addition, as per the above formula, the probability is the result of expectation/n i.e f/sum(f), and below attached are the resulting snapshot of this step.

Graphical user interface, text, application, email

Description automatically generated

**Variable 3: Charges:** As charges is the continuous normal variable, below mentioned are the formulas to calculate the R code, pmf and cdf of it. In the below formula, the parameter mean (µ) can take any real values and the parameter variance (σ2) can take only a positive real value.



* R code = rnorm(s, µ, σ )
* pdf = dnorm(j, µ, σ )
* cdf i.e P(x<=j) = pnorm(j, µ, σ )

where s = sample size

j = random variable

µ = the parameter mean

p = the parameter standard deviation

So, in this step we are assigning all the values of charges column into ‘J’ variable and to calculate the mean and sigma we are using mean() and sd() function and assigning the value to variable ‘mu’ and ‘sigma’. In addition, we are also computing rnorm, dnorm and pnorm using the above mean and sigma value and below attached are the result snapshot of this step.

Graphical user interface, text, application

Description automatically generated

Table

Description automatically generated

**Text, table

Description automatically generated**

**Table

Description automatically generated**

**Text, table

Description automatically generated**

**Graphical user interface, application

Description automatically generated**

**A picture containing table

Description automatically generated**

**Chart, line chart

Description automatically generated**

**A picture containing table

Description automatically generated**

**Chart, line chart

Description automatically generated Graphical user interface, text, application

Description automatically generated**

**Variable 4: Bmi:** As charges is the continuous exponential variable, below mentioned are the formulas to calculate the R code, pmf and cdf of it. Below attached are the resulting snapshot of this step.

Diagram

Description automatically generated

Diagram

Description automatically generated

* R code = rexp(s, λ)
* pdf = dexp(j, λ)
* cdf i.e P(x<=j) = pexp(j, λ)
* alpha = (xbar^2)/var(X)
* λ = xbar/var(X)

where s = sample size

j = random variable

λ = the parameter rate

So, in this step we are assigning all the values of bmi column into ‘X’ variable and to calculate the mean and variance we are using mean() and var() function and assigning the value to variable ‘xbar’ and ‘V’. In addition, we are also computing rnorm, dnorm and pnorm using the above mean, sigma, alpha and lambda value and below attached are the result snapshot of this step.

**Table

Description automatically generated**

**A picture containing text

Description automatically generated**

**Table

Description automatically generated with medium confidence**

**Graphical user interface, text, application, email

Description automatically generated**

**Chart

Description automatically generated**

**Chart

Description automatically generated**

**A picture containing rectangle

Description automatically generated**

# **Question 3**

1. **Consider two categorical variables of the dataset, develop a binary decision-making strategy to check whether two variables are independent at the significant level alpha=0.01. To do so,**
2. **State the hypotheses.**
3. **Find the statistic and critical values.**
4. **Explain your decision and Interpret results.**

Hypothesistesting is used to assess the quality of the evidence from the sample and offers a framework for conclusions about the population.

Hypothesis testing of independence for two categorical variables is used to test whether the two categorical variables we selected are independent or dependent.

In question 3 we are using the dataset ‘Sleep in Mammals’. As mentioned, we took two categorical variables ‘exposure’ and ‘danger’ where Hypotheses H0 state that both the variables are independent and H1 state that they are dependent. We need to combine both variables using the matrix and find the expectation.

Then we find out the test value and the result is ’82.0440’. The critical value is found, and the result is ’31.9999’. As the test value is greater than the critical value, H0 is rejected, and we conclude that our variables X1 and X2 are dependent.

Code:

Graphical user interface, text, application

Description automatically generated

Output:

Scatter chart

Description automatically generated

1. **Consider one categorical variable, apply goodness of fit test to evaluate whether a candidate set of probabilities can be appropriate to quantify the uncertainty of class frequency at the significant level alpha=0.05.**

Goodness of fit indicates that sample data describes the data that we would predict to observe the actual population. As mentioned, we took one categorical variable ‘predation’ where Hypotheses H0 state that p1=p2=p3=p4=p5=1/5 and Hypotheses H1 state that it is not H0. To find the frequency of the data we evaluate the table function. The probability is found by dividing the frequency by the sum of the frequency. The expected value is calculated by multiplying the number of rows by the hypothetical probability. Then find out the test value, the result is ’3.32258’. The critical value is found, and the result is ’9.48772’. As the critical value is greater than the test value, H0 is accepted, and we conclude that p1=p2=p3=p4=p5=1/5.

Code:

Text

Description automatically generated

Output:

Chart, scatter chart

Description automatically generated

1. **Consider one continuous variable in the dataset, and apply test of mean for a proposed candidate of at the significant level alpha=0.05.**

In the test of mean, we have chosen the upper one-sided test. As mentioned, we took one continuous variable ‘body\_wt’ where Hypotheses H0 state that mu is less than 3 and Hypotheses H1 state that mu is greater than equal to 3. We calculate the mean, standard deviation, and size of our variable. Then find out the test value, the result is ’1.71455’. The critical value is found, and the result is ’1.64485’. As the test value is greater than the critical value, H0 is rejected, and we conclude that mu is greater than equal to 3.

Code:

**Text

Description automatically generated**

Output:

Text

Description automatically generated

# **MINUTES OF GROUP MEETING**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **DATE & TIME** | **TOPICS DISCUSSED** | **MEETING NOTES** | **PARTICIPANTS** |
| 1. | 16/02/2023  15:00 PM | Dataset selection | We focused on selecting the dataset from Kaggle. | Anuradha  Anjali  Sandeep |
| 2. | 17/02/2023  09:00 AM | Allocation of topics | After the confirmation of the dataset. We discussed the topics to be included and divided them among us. |
| 3. | 20/02/2023  12:00 PM | Follow-up meeting about the progress. | Here, we discussed about data preparation. |
| 4. | 27/02/2023  11:00 AM | Follow-up meeting about the progress | Completed the coding part. |
| 5. | 15/03/2023  11:00 AM | Preparation of draft | Collected all the data from team members and combined it to prepare the draft report. |
| 6. | 27/03/2022  2:00 PM | Preparation and finalizing of the report. | This was the final meeting where we all reviewed and finalized. | Anuradha  Anjali  Sandeep |

# **REFERENCES:**

[1] [https://www.abs.gov.au/statistics/understanding-statistics/statistical-terms-and-concepts/measures-central tendency#:~:text=A%20measure%20of%20central%20tendency,median](https://www.abs.gov.au/statistics/understanding-statistics/statistical-terms-and-concepts/measures-central%20tendency#:~:text=A%20measure%20of%20central%20tendency,median)

[2] <https://www.statisticshowto.com/measures-variation/>

[3] <https://study.dbs.ie/2122/msc-data/B9DA101/u2/index.html#/lessons/9vHVWZcy1KF6sHp9SZz1Rk9UKufuzRkY>

[4] <https://study.dbs.ie/2122/msc-data/B9DA101/u3/index.html#/lessons/x_flwgRvj5cK5Iyly6YyN68p4IW5a2R3>

[5] <https://elearning.dbs.ie/pluginfile.php/1771813/mod_resource/content/1/decision%20making%20and%20hypothesis%20testing%201.pdf>

[6] <https://www.probabilitycourse.com/chapter8/8_4_3_hypothesis_testing_for_mean.php>

[7] <https://www.simplilearn.com/tutorials/statistics-tutorial/hypothesis-testing-in-statistics>

[8] <https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_hypothesistest-means-proportions/bs704_hypothesistest-means-proportions_print.html#:~:text=In%20hypothesis%20testing%2C%20we%20assume,or%20under%20the%20null%20hypothesis>).