STATISTICAL ANALYSIS

COVID 19 Data set

MARCH 2, 2022

SACHIN JARE
(M.SC- STATISTICS)

Assessment 1

1.Obtain the number of patients by taluka's, occupation, education respectively.

The following table shows the number of patients by taluka:

Taluka	Taluka Name	Number of Patients
1	Gadchiroli	105
2	Chamorshi	109
3	Aheri	100
4	Armori	103
5	Kurkheda	105
6	Desaiganj	95
7	Dhanora	78
8	Etapalli	99
9	Sironcha	108
10	Mulchera	117
11	Korchi	107
12	Bhamragad	93

The following table shows the number of patients by occupation:

Occupation	Occupation Name	Number of Patients
1	Farming	153
2	Home Maker/Housewife	185
3	Buisness	166
4	Service	180
5	Student	178
6	Labour	188
7	Nill	169

The following table shows the number of patients by education:

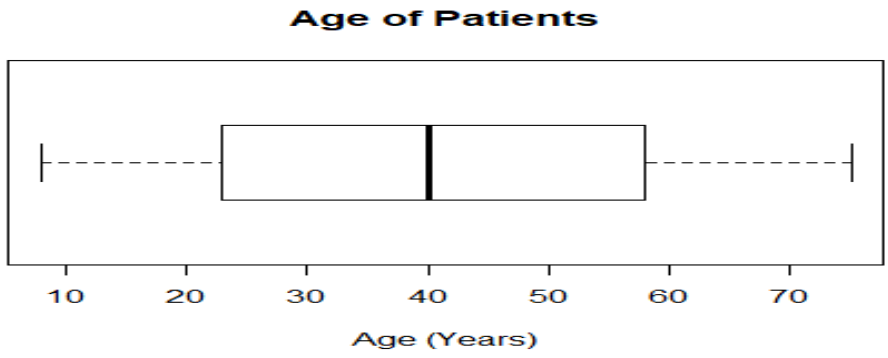
Education	Education Level	Number of Patients
1	Illiterate	0
2	Literate	1219
3	1 to 4	0
4	5 to 7	0
5	8 to 10	0
6	11 to 12	0
7	Graduate	0
8	Post Graduate	0
9	Above Post graduate	0
10	Not available	0

2. Obtain the mean age of patients in each taluka. What is the distribution and shape of age? Plot appropriate graph and interpret.

The following table shows the mean age of patients by taluka:

Taluka	Taluka Name	Mean Age
1	Gadchiroli	37.06
2	Chamorshi	39.72
3	Aheri	41.3
4	Armori	40.04
5	Kurkheda	41.18
6	Desaiganj	43.13
7	Dhanora	40.56
8	Etapalli	42.67
9	Sironcha	41.96
10	Mulchera	41.68
11	Korchi	40.58
12	Bhamragad	39.37

The **distribution** of a data set is the shape of the graph when all possible values are plotted on a frequency graph. The shape of patient's age is symmetrically distributed. It can be observed through the boxplot of age distribution.



The five number summary of age is given in following table:

Minimum	First Quartile	Median	Third Quartile	Maximum
8	23	40.00	58	75

The youngest and oldest patients are 8 years old and 75 years old, respectively. Twenty-five percent of patients are under the age of 23. Half of the patients are under the age of 40. There are also 25% of patients who are older than 58 years old.

3. classification obtain the taluka-wise prevalence of Mild, Moderate, and Severe COVID-19 cases.

The following table shows the stage of COVID-19 by taluka:

Taluka	Taluka Name	Mild COVID-19	Moderate COVID-19	Severe COVID-19
1	Gadchiroli	42	56	7
2	Chamorshi	40	57	12
3	Aheri	38	49	13
4	Armori	44	42	17
5	Kurkheda	46	48	11
6	Desaiganj	36	51	8
7	Dhanora	26	38	14
8	Etapalli	31	53	15
9	Sironcha	57	39	12
10	Mulchera	44	59	14
11	Korchi	40	50	17
12	Bhamragad	42	44	7

4. **The number of patients of appropriate age group-wise. Classify stages of COVID-19 according to age group.**

The age group classification is considering; Babies are those under the age of three. Children are those aged 3 to 16 years old. Young adults are those between the ages of 17 and 30. Adults between the ages of 31 and 45 are considered middle-aged. Old adults are those over the age of 45.

The following table shows the stage of COVID-19 by age group:

Age Group	Mild COVID-19	Moderate COVID-19	Severe COVID-19
Babies	0	0	0
Children	75	92	30
Young Adults	114	111	33
Middle Aged Adults	90	135	34
Old Adults	207	248	50

From above table, it can observe that there are no cases of COVID-19 in babies, and the majority of cases are in elderly people with mild to moderate COVID19. The majority of COVID19 cases in middle-aged adults are mild.

5. Is it possible to do a simple linear regression of the prevalence of COVID-19 on the mean age of patients of taluka. Explain

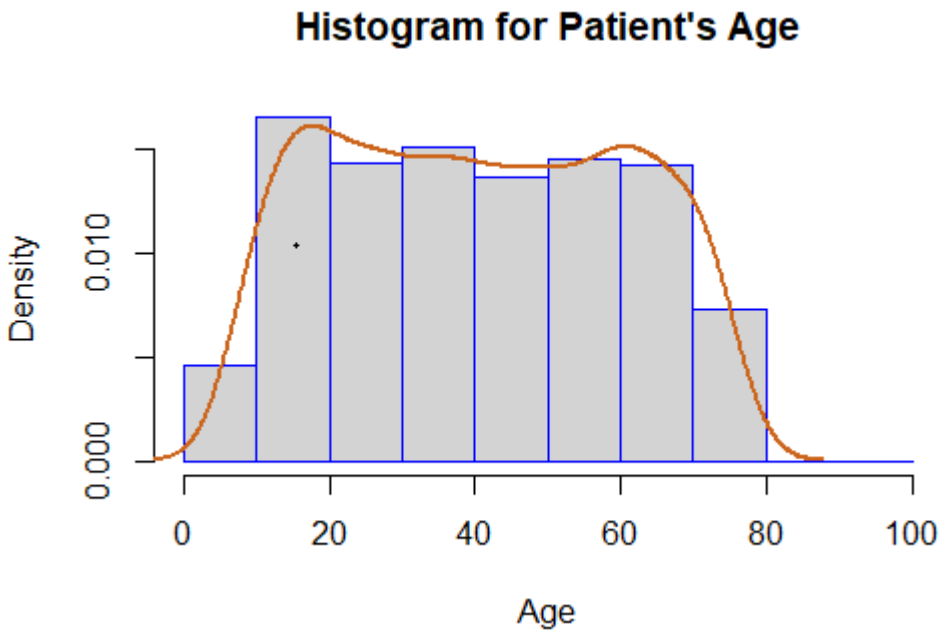
No, because the regression assumptions seem to be invalid in this case. There should be a linear relationship between age and COVID-19 prevalence, and observations should be independent of each other, but in the case of COVID-19 virus spread, observations are not independent for each patient. Hence every patient of COVID19 in this data set may not be independent.

6. Visually and numerically determine the shape of the age, occupation, respiratory rate. Interpret.

The Pearson coefficient of skewness used to decide the shape of distribution. A value of zero means no skewness at all. A large negative value means the distribution is negatively skewed. A large positive value means the distribution is positively skewed.

Age:

The person coefficient of skewness of age is **0.04**. It approximately equal to 0, it indicates that the distribution of age is no skew (**symmetrica**l).



From the above histogram of age, it can observe that the data is approximately symmetrically distributed.

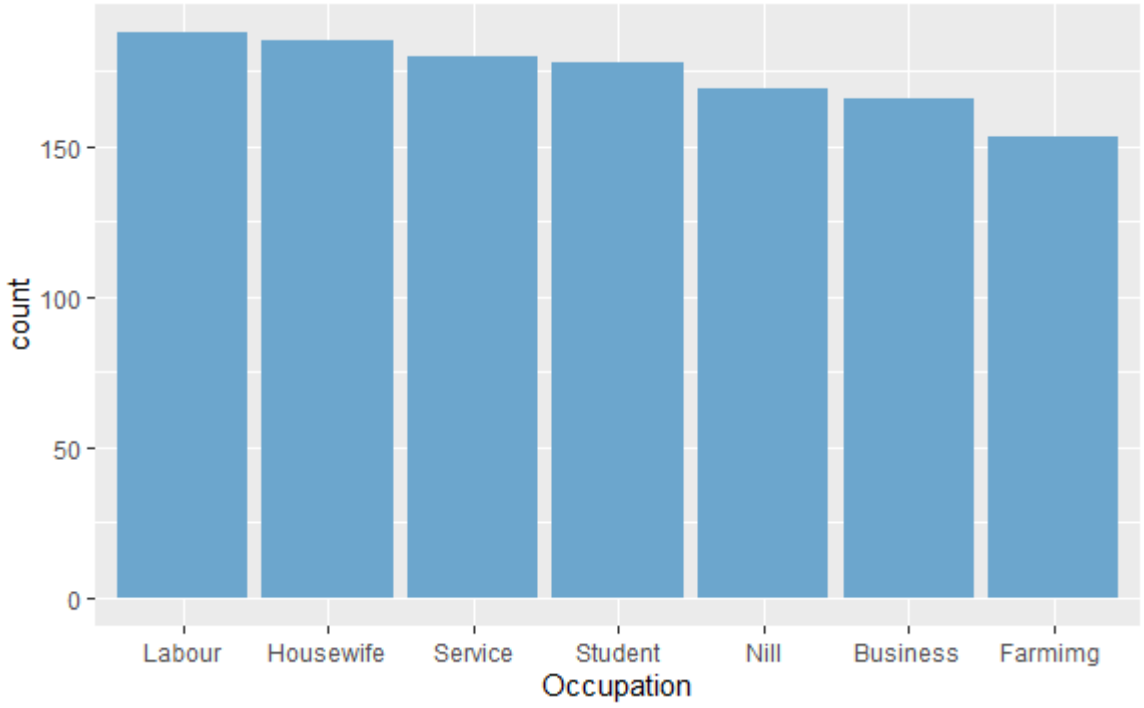
Occupation:

Occupations is categorical variable (Nominal data) so skewness is not appropriate measure.

The following table shows the patients occupation counts:

Occupation	Counts
Labour	188
Housewife	185
Service	180
Student	178
Nill	169
Business	166
Farmimg	153

Visualization: Bar chart

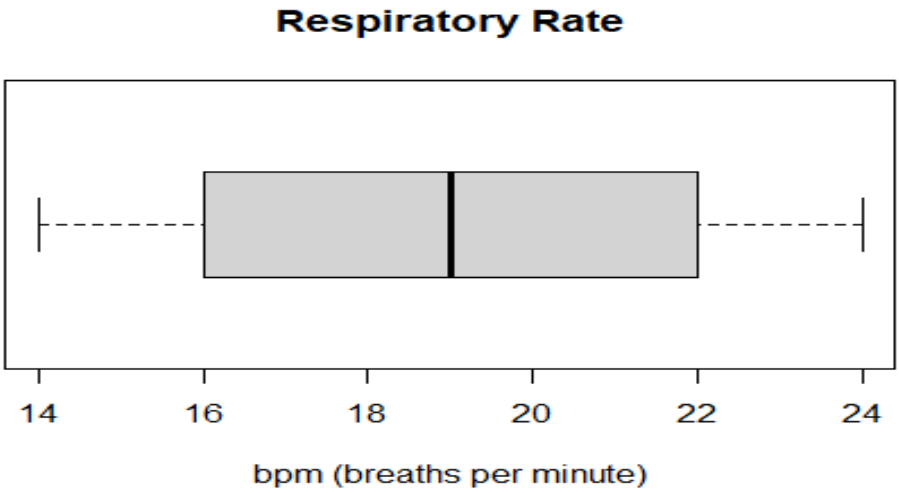


It can be seen that the majority of people work as labourers, with very few covid19 patients farming.

Respiratory rate:

The person coefficient of skewness of respiratory rate is **-0.0423**. It approximately equal to 0, it indicates that the distribution of age is no skew (**symmetrica**l).

Boxplot



Five number summary of Respiratory rate (bpm)

Minimum	First Quartile	Median	Third Quartile	Maximum
14.00	16.00	19.00	22.00	24.00

From above chart, it can observe that the distribution of data from left and right of median is same. Therefore, the Respiratory rate is not skew distribution.

7. Apply any three appropriate transformations to the data.

The following table shows the patients percentage according to work in public:

Work in Public	Percentage
Often	36.83
Never	22.81
NA	15.91
Sometimes	12.55
Rarely	11.89

People who work in public are at a high chance of infection covid19.

The following table shows the patients treatment percentage by occupation:

	Treatment (count)		Percentage	
Occupation	**No**	**Yes**	**No**	**Yes**
Business	77	89	46.39	53.61
Farming	65	88	42.48	57.52
Housewife	94	91	50.81	49.19
Labour	111	77	59.04	40.96
Nill	74	95	43.79	56.21
Service	87	93	48.33	51.67
Student	81	97	45.51	54.49

The majority of labour do not receive covid19 treatment.

8. Choose four appropriate variables and do multivariate analysis. Justify the technique used.

Different techniques of multivariate analysis:

- Principle component analysis (pca)
- Hierarchical clustering

Assessment 2

1. **What are the types of variables?**

A variable is a property that can take on any value. There two types of variables: Discrete variable, and Continuous variable.

A discrete variable is variable whose value is obtained by counting. Example: Number of students present in class, Number of children in family, Respiratory Rate (bpm)

A continuous variable is a variable whose value is obtained by measuring, i.e. on which can take on an uncountable set of values. Example: Height, Weight, Temperature.

2. **Why is McNemar test?**

McNemar's test is a statistical test used on paired nominal data. It is used to determine whether the row and column marginal frequencies are equal in 2 x 2 contingency tables with a dichotomous attribute and matched pairs of subjects (that is, whether there is "marginal homogeneity").

3. **When you use median, mean and mode?**

Mean, Median, Mode are the measures of central tendency of data. Mean is average of all data points. Median is middle observation of sorted data points (ascending or descanting). Mode is most repeated observation in data.

Mean is the most frequency used measure of central tendency. It is used for non-categorical data set. It is highly affected by outlier values.

Median is mostly used in case of skewed data set. It is unaffected by the outlier values.

Mode is the preferred measure when data are measured in a nominal or ordinal scale. In other words, mode is used for categorical variables.

4. **Why p values are used in any test?**

The P-value is used to take a decision about the null hypothesis. The p-value represents the likelihood of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis.

5. **What are the assumptions of ANOVA?**

There are Three basic assumptions used in ANOVA:

- The populations from which the samples were taken are normally distributed.
- Homogeneity of variance - that the variance of data in the different groups should be the same.
- Random sampling - each sample has been drawn independently of the other samples

6. **When you use non- parametric test?**

When the assumptions of a parametric test do not meet, we can use a non-parametric test.

For example, in the following situation, non- parametric can be used:

- Data are not normally distributed.
- Samples are not selected randomly.
- If there are less samples.
- It can be used on ordinal and nominal scale data.

7. **What is the non-parametric alternative for T-test?**
   - Sign test, Wilcoxon signed ranked test are non-parametric alternative to T-test.
   - Mann-Whitney test is a non-parametric test for dependent T –test (paired t test).

8. **What is the difference between association and correlation?**

The term association refers to the general relationship (whether dependent or independent) between two random variables, whereas correlation refers to a more or less linear relationship between the random variables. Correlation reflects the magnitude and direction of a linear relationship between two numerical variables.

9. **To perform linear regression what needs to be satisfied?**

Following requirements must be satisfied in order to perform linear regression:

Linearity: The relationship between X and the mean of Y is linear.

Homoscedasticity: The variance of residual is the same for any value of X.

Independence: Observations are independent of each other.

Normality: For any fixed value of X, Y is normally distributed.

No Multicollinearity: One independent variable is not correlated with another independent variable.

10. **What are the different techniques that be can used for categorical regression.?**

In the regression it is necessary the data should be in numerical form. if some categorical variables are present. In this case, the categorical variable can be converted to numerical values by using Label encoding or one hot encoding techniques.

11. **What is meta analysis? What are the requirements for performing meta analysis and explain the analysis steps?**

Meta-analysis is the statistical combination of results from two or more separate studies. Potential advantages of meta-analyses include an improvement in precision, the ability to answer questions not addressed by individual studies

In meta-analysis, it is crucial to evaluate the direction of effect, size of effect, homogeneity of effects among studies, and strength of evidence

Steps of Meta- analysis as follows:

1. Define the Research Question
2. Perform the literature search
3. Select the studies
4. Extract the data
5. Analyze the data
6. Report the results.