Ahmedabad University

MAT 502 : Advanced Statistics

Winter Semester 2022

G

Project Report

Submitted to Faculty : Prof. Shashi Prabh

Teaching Assistants : Dhruvil Dave and Arpit Vaghela

Group Members	Enrollment No
Malav Doshi	AU1940017
Sameep Vani	AU1940049
Kavya Patel	AU1940144
Kashvi Gandhi	AU1940175

Code and Database

Motivation

As the use of sanitizers has highly increased after the pandemic, we decided to conduct a survey for the same. We have gathered information regarding the change in people's willingness to pay for sanitisers and the quantity they use before and after pandemic. Also, analysis is done for which brand and what type of sanitizer has been used the most.

Objective

Our main objective is to study the statistics and come up with a conclusion whether the variation is due to chance error or pandemic on various factors. Multiple hypotheses have been tested on the dataset, to draw several conclusions.

Sampling Methods and Experiment design

The sampling method we have used is **Voluntary response sampling** as it was circulated, and the users willing to participate in online surveys participated in the survey.

The questions are designed in such a way to draw conclusions on various aspects of usage of sanitizers. Factors such as age group, gender, brand and type preferences, and willingness to pay were taken into account.

Based on these variables and data collected using the aforementioned sampling method, we are going to test two hypotheses. Below are the statements for the same.

1. Hypothesis to test whether the maximum willingness to pay has actually increased due to the pandemic or is it just a chance variation. (Using z-Test for the same)
    a. **Null Hypothesis (H_0):** The increase in average maximum willingness to pay for a bottle of sanitizer is only due to chance error and not a direct implication of the pandemic.
    b. **Alternate Hypothesis (H_1):** The increase in average maximum willingness to pay for a bottle of sanitizer is not due to chance error but an implication of the pandemic.

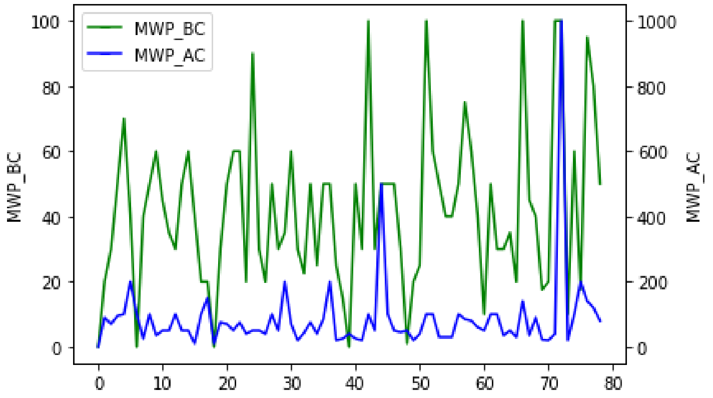
2. Hypothesis to test whether the preference of chemical from which sanitizer is made is independent of gender or not.
    a. **Null Hypothesis(H_0)**: The chemical preferences of a person is independent of gender.
    b. **Alternative Hypothesis(H_1)**: There is some association between chemical preferences and gender.

Data Collection

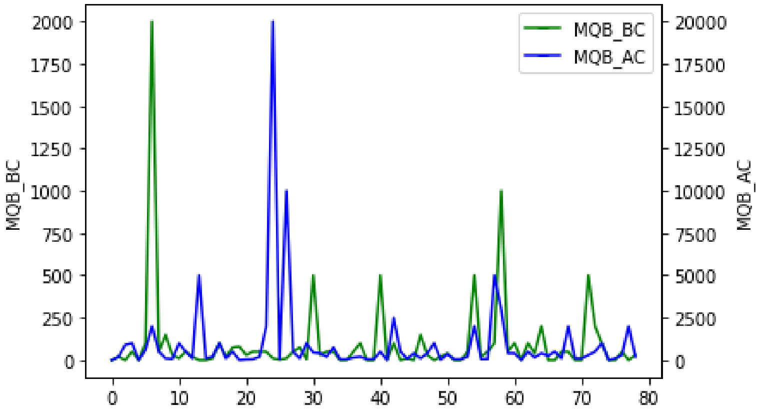
A survey about the use of sanitizer before and after the pandemic is taken to know about the variation in its use. The **data collection method** hence used is **'Survey'**. We have collected around 80 responses. However, there is a '**selection bias'** because the majority of the respondents fall in the age group of 16-20. This is because the survey was majorly circulated amongst the college students. This may involve the possibility of 'Non Response Bias' which might vary different variables such as brand preferred or the willingness to pay. Further, this survey was voluntary (as mentioned above), in which respondents can choose not to answer at all. However, if a respondent chooses to answer, then for sure, all the questions are answered. Furthermore, the survey doesn't consider the fact that respondents can change their maximum willingness as per the seriousness of covid.

Data Visualization and Data Analysis

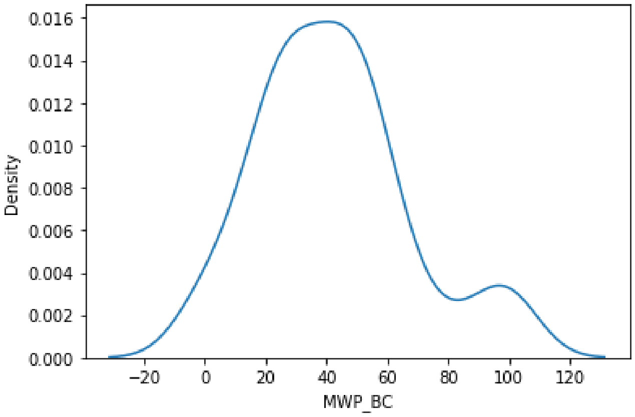
1. The difference between the willingness to pay for sanitizer before and after pandemic for every respondent.



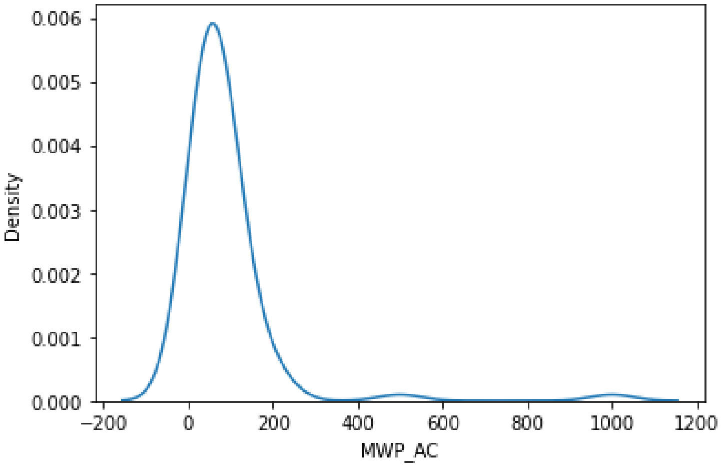
2. The difference in quantity of sanitizer bought before and after pandemic by every respondent.



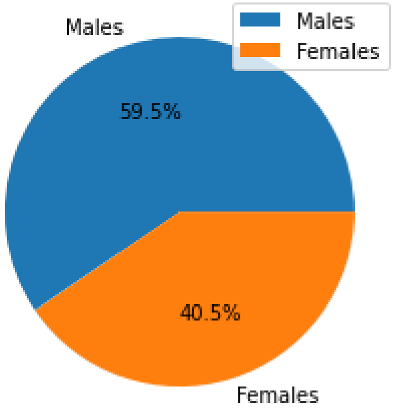
3. The distribution of willingness to pay money before pandemic by the respondents



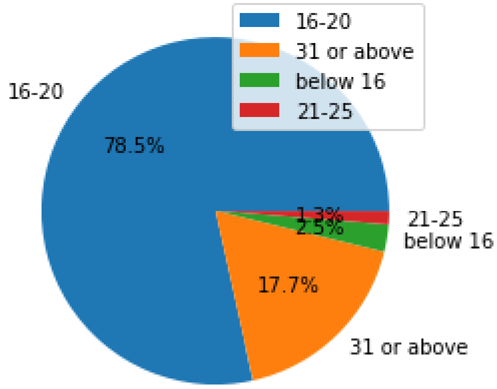
4. The distribution of willingness to pay money after pandemic by the respondents



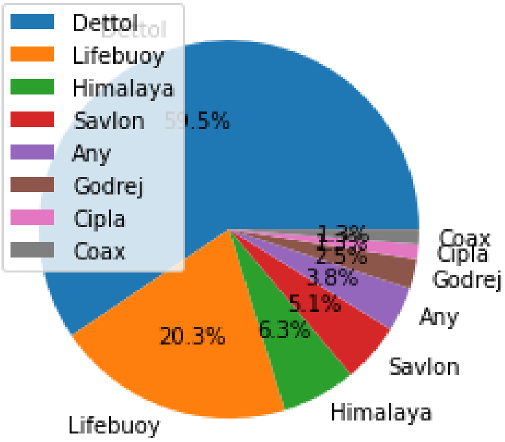
5. The gender ratio distribution



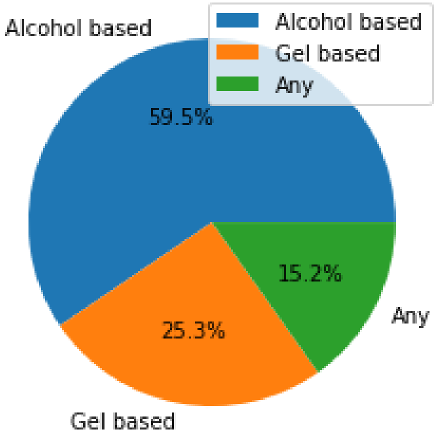
6. The age distribution



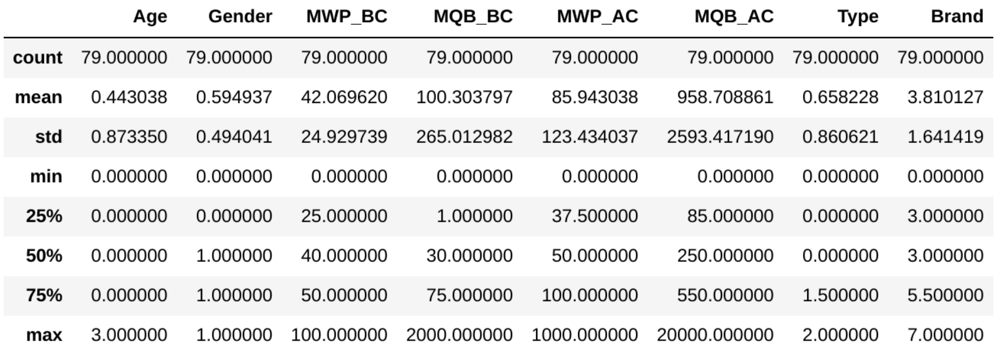
7. Preferred brand distribution



8. The type of sanitizer preferred



Summary Statistics



	Age	Gender	MWP_BC	MQB_BC	MWP_AC	MQB_AC	Type	Brand
count	79.000000	79.000000	79.000000	79.000000	79.000000	79.000000	79.000000	79.000000
mean	0.443038	0.594937	42.069620	100.303797	85.943038	958.708861	0.658228	3.810127
std	0.873350	0.494041	24.929739	265.012982	123.434037	2593.417190	0.860621	1.641419
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	25.000000	1.000000	37.500000	85.000000	0.000000	3.000000
50%	0.000000	1.000000	40.000000	30.000000	50.000000	250.000000	0.000000	3.000000
75%	0.000000	1.000000	50.000000	75.000000	100.000000	550.000000	1.500000	5.500000
max	3.000000	1.000000	100.000000	2000.000000	1000.000000	20000.000000	2.000000	7.000000

In the above image, there are multiple features like Age, Gender, Type and Brand which are categorical features and have been encoded here to contain numerical inputs. Thus, the summary statistics of these features is not of prime importance. However, MWP_BC (willingness to pay before pandemic), MQB_BC (quantity of sanitizer bought before pandemic), MWP_AC (willingness to pay after pandemic) and MQB_AC (quantity of sanitizer bought after pandemic) are the features for which these statistics would make sense.

Hypotheses

1. **First Hypothesis :**

   Null Hypothesis (H_0): The increase in average maximum willingness to pay for a bottle of sanitizer is only due to chance error and not a direct implication of the pandemic.

   Alternate Hypothesis (H_1): The increase in average maximum willingness to pay for a bottle of sanitizer is not due to chance error but an implication of the pandemic.

   We used z-statistics using python to know whether the difference is because of chance variation or not.

   The results are shown below:

   - Z value= 15.742131127205779
   - p_value= 3.888970538024388e-56

   The z-value is very high which tells us that the observed value is very far from the expected value and as a result we get a low p-value. Hence, we reject the null hypothesis that says that the variation is NOT due to chance error.

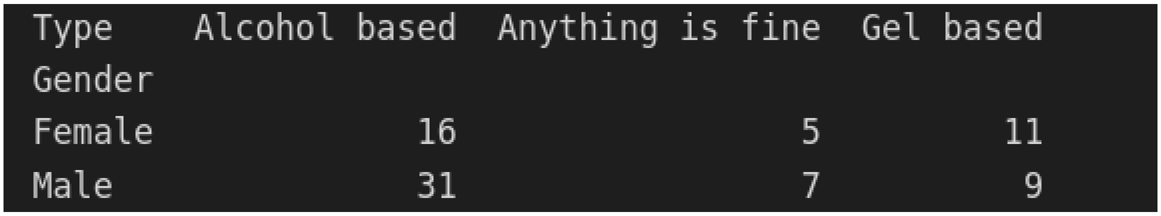
2. **Second Hypothesis :**

   Null Hypothesis(H_0): The chemical preferences of a person is independent of gender.

   Alternative Hypothesis(H_1): There is some association between chemical preferences and gender.

   We used the Chi-square method for independence using Python.

   Following Results have been obtained:



```
Type    Alcohol based  Anything is fine  Gel based
Gender
Female             16                 5         11
Male               31                 7          9
```

- Degree of Freedom = 2
- The Chi-square Statistic comes out to be : 0.82155
- P-value : 0.336

  The p-value is high, hence there is not enough evidence to reject Null Hypothesis, hence the two variables are independent.

Conclusions

- On carrying out the hypothesis tests, we got to know that the difference in the maximum willingness to pay before and after the pandemic was not due to any chance variation. This is also true as due to covid, there was an increase in the use of sanitizers. Hence, individuals' willingness to pay increased. Thus, our hypothesis supports the evident data.
- The Second Hypothesis was to oversee the relationship between variables 'Gender' and 'Sanitizers Type Preferences', where the outcome comes out to be that both these variables are independent. Now this test was conducted on different variables, but the outcome was the same. This might be due to a small sample size which might contain 'bias'.

Improvements to get Better Results

- Different sampling methods, such as probabilistic methods such as Cluster Sampling, unlike the current non -probabilistic sampling method which we have selected due to some constraints. Cluster or Simple Random Sampling would be a better choice, to get better samples which represent less skewed data than we have obtained.
- Larger sample size would make a significant difference since the current dataset does not merely represent the population.
- There are some outliers in the data. One can build on this project by removing such outliers in order to avoid any distortion caused during analysis.

How can someone use it to build

- Based on the dataset provided, one can formulate multiple hypotheses to test other observations and get a more general idea about the population.
- Someone can perform a better sampling method and recalculate the test scores and compare them with ours to find the caveats in our methods and ways to fix them.