**STAT 6210 Homework 1 Report**

Xue Ming Wang (Vivian)

G20580112

**Part A**

The dataset, hw1a, is a simulated data set that includes 1000 records and 3 variables. The three variables are X, Y, and Z, where X is a numerical value, and Y and Z are categorical values.

1. Summary statistics of variable X

1.1 Basic statistical summary table of variable X

|  |  |
| --- | --- |
| Moments for Variable X | |
| Mean | 23.183 |
| Variance | 16.4019 |
| Standard Error | 4.0499 |
| CV | 17.4693 |
| Skewness | 0.0954 |
| Kurtosis | −0.1229 |
| Table 1. Basic statistical summary table of variable X | |

1.2 Quantiles of variable X

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Minimum | Q1 | Median | Q3 | Maximum |
| 12 | 20 | 23 | 26 | 37 |
| Table 2. Quantiles of variable X | | | | |

* 1. Test for Location: μ0=23

|  |  |  |  |
| --- | --- | --- | --- |
| Test | Statistic | P Value | |
| Student’s t | 1.4289 | pr > |t| | 0.1533 |
| Sign | 4 | pr >= |M| | 0.8155 |
| Signed Rank | 9550.5 | pr >= |S| | 0.2195 |
| Table 3. Test for Location: μ0=23 | | | |

Table 3 shows that the p-value for all three tests is greater than 0.05. It indicates that these tests are not significant and that the assumption of μ0 = 23 cannot be rejected. Therefore, the mean of variable X is approximately 23 (μ0 = 23).

* 1. Distribution of variable X

Chart, line chart, histogram

Description automatically generated

Figure 1. Histogram of variable X

Figure 1 displays the histogram and the distribution of variable x. The shape of the histogram (μ = 23.183 and σ= 4.0499) demonstrates that the data's distribution is following the normal curve, which assumes that the data is normally distributed. However, it still needs more testing to be verified. To do so, the Q-Q plot will be able to analyze the normality of the data in further detail.

Chart, scatter chart

Description automatically generated

Figure 2. Normal Quantile-Quantile Plot for Variable X

The distribution shown in figure 2 is mostly linear, except for a little heavy tail. In general, it can be assumed that the data is following the normal distribution. Meanwhile, the Shapiro-Wilk test (W = 0.8089, p-value < 0.0001) and the Kolmogorov-Smirnov test (D = 0.2426, p-value < 0.0100) show that the p-values are all significant (< 0.05). Therefore, we can say that the data are normally distributed.

2. Variable X in different groups

When considering the variables Y and Z, which are categorical variables, they would influence the other variables. To test, the dataset is divided into three groups (0, 1, and 2) to examine the effects of variables Y and Z.

2.1 Different Y Groups

Chart, histogram

Description automatically generated

Figure 3. Histogram of variable X with different Y values

Chart, box and whisker chart

Description automatically generated

Figure 4. Boxplot of variable X with different Y values

The histogram in figure 3 and box plot in figure 4 depict the distribution of variable X with different Y values. In figure 3, the mean and standard deviation for variable X (on the top right of each histogram) are different in each Y group, but the difference is small. Yet, looking at the shape of the histogram and the normal curve, the Y = 0 group seems normally distributed, while the Y = 1 group seems left skewed, and the Y = 2 group seems right skewed. In figure 4, the boxplot shows that the quantile distributions of variable X have similar means and medians. However, the interquartile range is different for the three groups, where the interquartile range for the Y = 0 group is greater than that of the Y = 1 group. In addition, the maximum and minimum values of each group are not the same as well. Thus, we can say that variable X does not have the same distribution in these Y groups.

2.2 Different Z Groups

Chart, histogram

Description automatically generated

Figure 5. Histogram of variable x with different Z values

Chart, box and whisker chart

Description automatically generated

Figure 6. Boxplot of variable x with different Z values

The histogram in figure 5 and box plot in figure 6 depict the distribution of variable X with different Z values. In figure 3, the mean and standard deviation for variable X (on the top right of each histogram) are different in each Z group, and the difference is bigger than Y values. I think that the influence of group Z on variable X is greater than that of group Y, from the means’ and standard deviations’ perspective. On the other hand, looking through the shape of the histogram and normal curve, the three Z groups (Z = 0, Z = 1, and Z = 2) seem to be all normally distributed, even though the 3 histograms are not too smooth as big gaps appear in the middle of the histogram (especially when Z = 0 and Z = 1). In figure 4, the boxplot shows that the quantile distributions of variable X have different means and medians, as well as the interquartile range. It is obvious that Z = 2 is lower than other groups. Hence, we can say that variable X does not have the same distribution in these Z groups.

**Part B**

The dataset, hw1b, is a simulated data set with three variables. The three variables are X, Y, and Z, where X and Y are categorical values, and Z is the frequency.

1. The Observed Frequencies Z with X as grouping variable

Chart, bar chart

Description automatically generated

Figure 7. Bar chat of Z with X as groping variable

In figure 7, the bar chart of z with different x values is shown with the sum of the variable Z on the top of each bar. The maximum value is when X = 1 and Z = 327, while the minimum value is when X = 2 and Z = 290. The difference between each X group is not too large as the range of each X group is 37.

1. The Observed Frequencies Z with Y as major grouping variable

Chart, bar chart

Description automatically generated

Figure 8. Bar chart of Z with Y as major grouping variable and X as sub-grouping variable

Figure 8 presents the bar chart of variable Z with a major group of Y and a sub-group of X. The sum of variable Z appears at the top of each bar. The maximum value is when Y = 2 and Z = 736, while the minimum value is when X = 1 and Z = 711. The difference between each Y group is not very large at all since the range of the X group is only 25. The percentage shown in each block is the percentage of each X sub-group for the frequency of Z.  The maximum percentage is 16.58% when Y = 2 and X = 2, while the minimum percentage is 11.25% when Y = 1 and X = 2.