Homework

Jiaying Liu
G45268292

1. Part A

Part A data have 1000 records with 3 variables which x is numerical value and y and z are categorical values. And the following step is to show the statistical result

1.1. Summary statistics of variable x

Tabla 1. Basic statistical summary of variable x.

Moment			
Mean	23,183	CV	17,4693
Variance	16,4019	Skewness	0,0954
Standard error	4,0499	Kurtosis	−0,1229

Tabla 2. Quantiles of variable x.

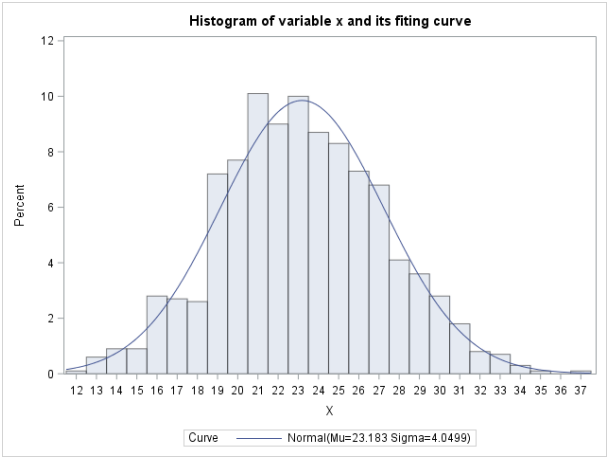
Minimum	Q1	Median	Q3	Maximum
12	20	23	26	37

Tabla 3. Test for Location:μ_0=23.

Test	Statistic	P Value	
Student's t	1,4289	$p > \|t\|$	0,1533
Sign	4	$p > \|M\|$	0,8155
Signed Rank	9550,5	$p > \|S\|$	0,2195

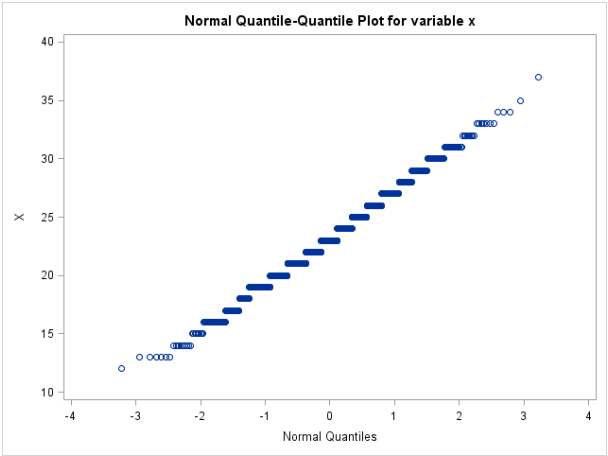
From above table 3, it is obviously that no matter which test method, the p value both greater than 0,05 which mean the test is not significant and we can not reject the original assumption. It can conclude that the mean of variable is near to 23.

Figura 1. Histogram of variable x and its fitting curve



From above figure 1, it shows the histogram and the distribution of variable x. From the shape of plot, with the $\mu = 23{,}183$ and $\sigma = 4{,}0499$,it can assume that the data is follow by the normal distribution but it still need exact test to verify its normality. So next step, I use Q-Q plot to text normality in further detail.

Figura 2. Normal QQ plot of data



From the above figure 2, it is obviously the line is linear which can represent that the data is a normal distribution data. To further verify the normality, I use the Shapiro-will test. according to the SAS result W =0.9929 and the $p - value$ =0.0001. It is a significant $p$value so we can say that the data follow normal distribution.

1.2. Statistical character of variable x with different class values

In the data, there are two categorical variable which may affect the statistical character of variable x. So the next step is to check the influences of variable y and z.

Figura 3. Histogram of variable x with different y values

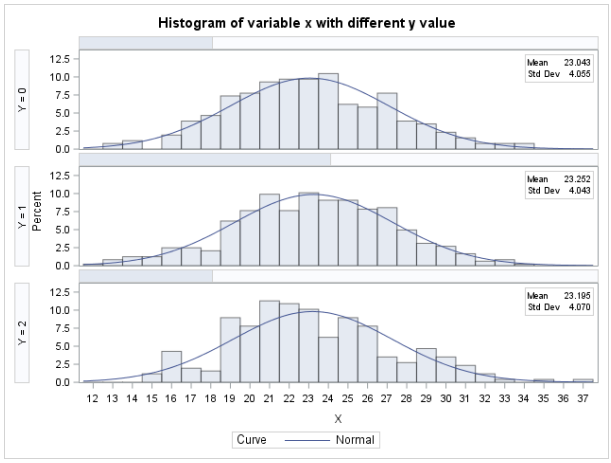
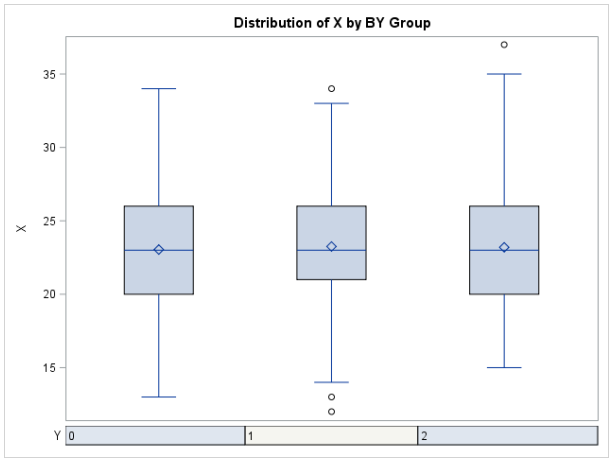


Figura 4. Boxplot of variable x with different y values



From the figure 3 and figure 4, it show the distribution and boxplot of variable x with class y. Focus on figure 3, on the right side of plot, it show the mean and standard deviation of x. When $y = 0$, the mean is 23.043 and the std Dev if 4.055. when $y = 1$, the mean is 23.252 and the Std Dev is 4.043. When $y = 2$, the mean is 23.195 and the Std Dec is 4.040. In my opinion there are some different between each mean and std. From the figure 4, the boxplot show the quantile distribution of x.It show that they have similar median but the Q1 and Q3 quantile are a little bit different.. In conclusion, it can say that they have different distribution.

Figura 5. Histogram of variable x with different y values

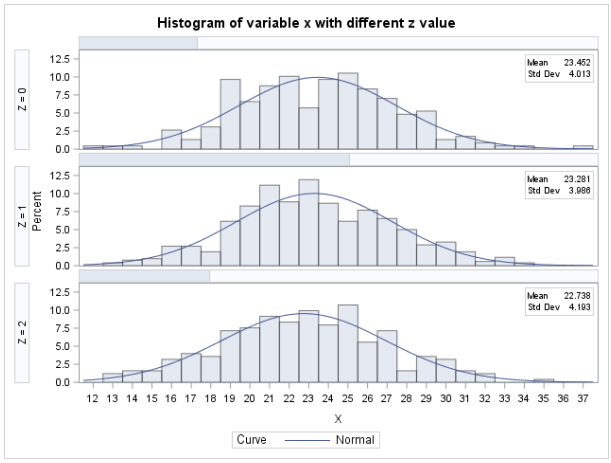
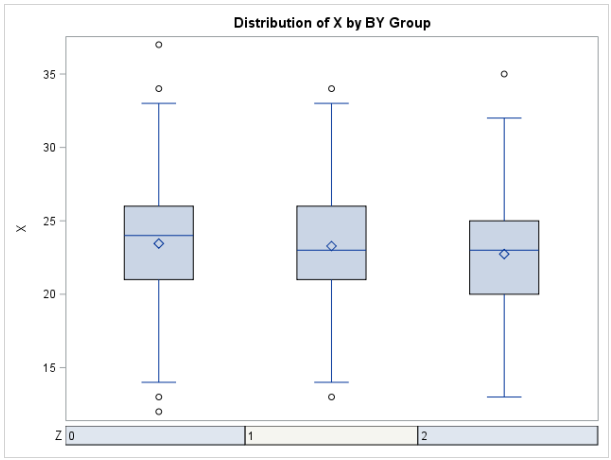


Figura 6. Boxplot of variable x with different y values



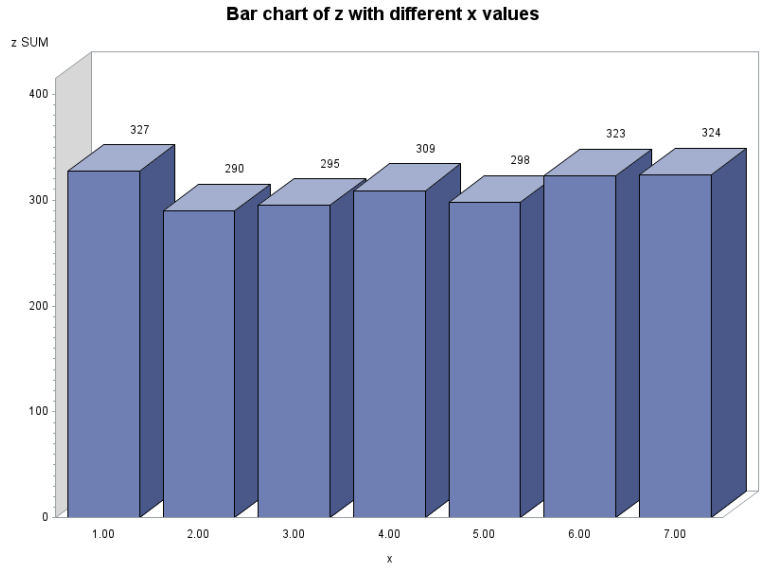
From the figure 5 and figure 6, it show the distribution and boxplot of variable x with class z. Focus on figure 3, on the right side of plot, it show the mean and standard deviation of x. When $z = 0$, the mean is 23.452 and the std Dev if 4.013. when $z = 1$, the mean is 23.281 and the Std Dev is 3.986. When $z = 2$, the mean is 22.738 and the Std Dec is 4.193. In my opinion, compare with the class y, class z have more influence on variable x. the histogram are not smooth. It is conspicuous that when $z = 0$, there is a big gap in the middle of histogram.From the figure 4, the boxplot show the quantile distribution of x.It show that they have different median and the Q1 and Q3 quantile. what's more the boxplot of $z = 2$ is lower then the other. In conclusion, it can say that they have different distribution.

2. Part B

In the part B, the data set have three variable in which x and y is categorical variable and z is frequency variable.

2.1. Bar chat for z with x as groping variable

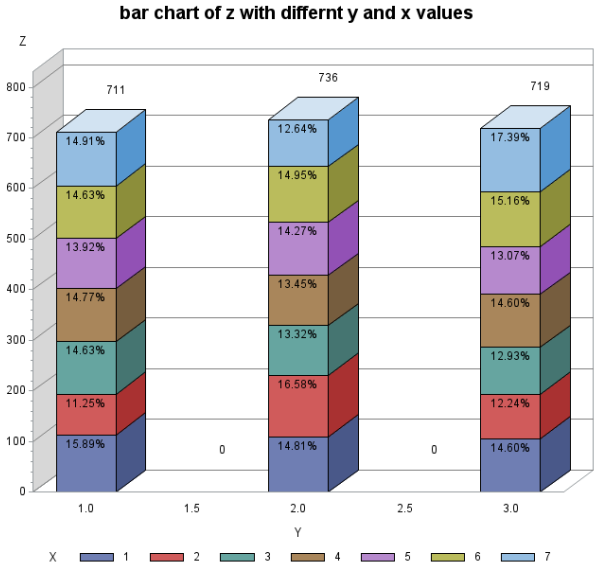
Figura 7. Bar chart of z with different x values



From above figure 9, it show the bar chart of z with different x value. The number above each bar is the sum of z. It can conclusion that when x=1, z=327.When x=2,z=290. When x=3,z=295.When x=4,z=309. When x=5,z=298. When x=6,z=323. When x=7,z=324.

2.2. Bar chat for z with x and y as groping variable

Figura 8. Bar chart of z with different y and x values



From figure 8, it show that the bar chart of z with different y and x value. The main class is variable y and the sub-class is variable x. When y=1, z=711 which there are 15.99 percentage of x=1,11.25 percentage of x=2,14.63 percentage of x=3,14.77 percentage of x=4,13.92 percentage of x=5,14.63 percentage of x=6,14.91 percentage of x=7.When y=2, z=735 which there are 14.87 percentage of x=1,16.58 percentage of x=2,13.32 percentage of x=3,13.45 percentage of x=4,14.27 percentage of x=5,14.95 percentage of x=6,12.64 percentage of x=7.When y=3, z=719 which there are 14.6 percentage of x=1,12.24 percentage of x=2,12.53 percentage of x=3,14.6 percentage of x=4,13.07 percentage of x=5,15.15 percentage of x=6,17.39 percentage of x=7.