HW 1 Report

Yongchao Qiao

Part A

1. Report the summary statistics of variable *X*

As Table 1 shows below, variable X has mean as 30.601, with variance as 18.628, standard deviation as 4.316, standard error as 0.136, coefficient variation as 14.104. The skewness and kurtosis of variable X are 0.151 and -0.212 respectively, both of which are near 0, so it infers that variable X might be normally distributed. Also, the median of variable X is 31 which is similar with the mean. Besides, the minimum is 18, with Q1 as 28, Q3 as 34 and maximum as 43.

Table 1: Summary Statistics

Mean	30.601	Variance	18.628	SD	4.316
Standard error	0.136	CV	14.104	Skewness	0.151
Kurtosis	-0.212	Minimum	18	Q1	28
Median	31	Q3	34	Maximum	43

2. Test for location: $\mu_0 = 33$

$$H_0 : \mu_0 = 33, \quad H_1 : \mu_0 \neq 33$$

In this step, though it has not been verified that the variable X follows the normal distribution, we just assume it does to continue the student's t test (Although it might not be valid). Also, for the sign and signed rank tests, we just assume variable X has the symmetric distribution (It is might not be right, though the skewness is near 0). The results of location tests are shown as table 2. Since P values of student's t test, sign test and signed rank test are all less than 0.05, so we can reject the null hypothesis that $\mu_0 = 33$, for student's test with respect to mean while for sign and signed rank tests with respect to median.

Table 2: Location tests

Test	Statistics		P-value	
Student's t	t	-17.577	Pr>|t|	<0.0001
Sign	M	-210	Pr>=|M|	<0.0001
Signed Rank	S	-125871	Pr>=|S|	<0.0001

3. Test for normal distribution

H_0 : The variable X follows the normal distribution

H_1 : The variable X does not follow the normal distribution

Table 3: Normal distribution tests

Test	Statistics		P-value	
Kolmogorov-Smirnov	D	0.056	Pr>D	<0.010
Cramer-von Mises	W-sq	0.468	Pr>W-sq	<0.005
Anderson-Darling	A-sq	2.696	Pr>A-sq	<0.005

The results of normal distribution tests are shown as table 3. Since P values of Cramer-von Mises, Kolmogorov-Smirnov and Anderson-Darling are all less than 0.05, so we can reject the null hypothesis that the variable X follows the normal distribution. Specifically, Kolmogorov-Smirnov test focuses on the largest vertical difference, while Cramer-von Mises focuses on the overall squared difference with Anderson-Darling focusing on the weighted overall squared difference between the observed distribution and the theoretical normal distribution. However, no matter from which aspect, the variable X does not follow the normal distribution.

Figure 1

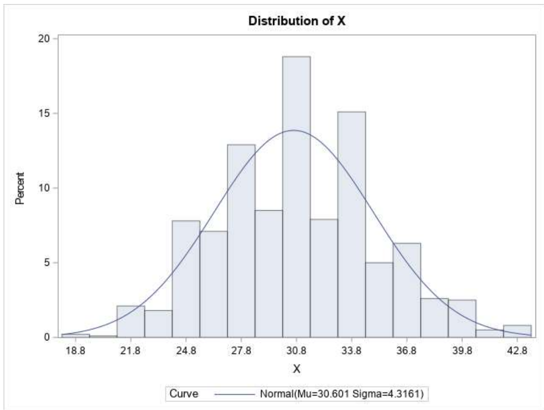
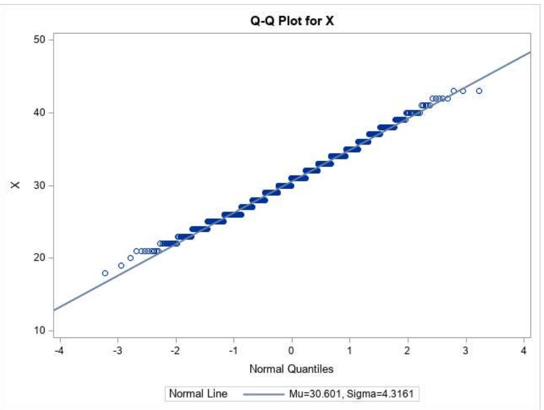


Figure 2



The histogram with fitted distribution and the Q-Q plot are shown as above. Technically, the observed distribution has much difference between the theoretical normal distribution. And for the Q-Q plot left end of pattern is above the line with right end of pattern below the line, which indicates short tails at both ends of the data distribution. Also, the Q-Q plot shows staircase pattern, which indicates that the variable X has been rounded or discrete. All these suggested that the variable X does not follow the normal distribution which is consistent with previous conclusion of tests.

Part B

1. Distributions of X in different groups of Y

The histograms and boxplot of X based on different values of Y are shown as below. Generally, the distribution of X when Y=0 is quite similar with that when Y=1 while when Y=2, the distribution is slightly different from that in previous groups. Specifically, the mean values are, around 30.6, almost the same in different Y groups with almost the same standard deviation as 4.3. Comparing the distribution of X when Y=0 and Y=1, they have the same median as 31, the same maximum observation below upper fence as 43, the same interquartile range as 7, the same Q1 as 27 and the same Q3 as 34 with the only significant difference as the minimum observation above lower fence, being 18, of Y=1 being smaller than that of Y=0 as 21. So X has almost the same distribution when Y=0 and Y=1. Comparing the distribution of X when Y=2 and Y=0&1, they do not have the same median because of 30 for Y=2 and 31 for Y=0&1; When Y=2, the maximum observation below upper fence, as 40, is lower than former two groups with the interquartile range, as 5, being lower than former two groups. Also, when Y=2, the Q3 is 33 which is lower than former two groups, while the Q1 is 28 which is greater than that of Y=0&1. Besides, when Y=2, the minimum observation above lower fence, being 21, is same with that of Y=0. In addition, when Y=2, there are extreme values, as 42, above the upper fence.

Overall, variable X has almost the same distribution when Y=0 and Y=1 while the distribution of variable X when Y=2 is slightly different from former two groups.

Figure 3 Y=0

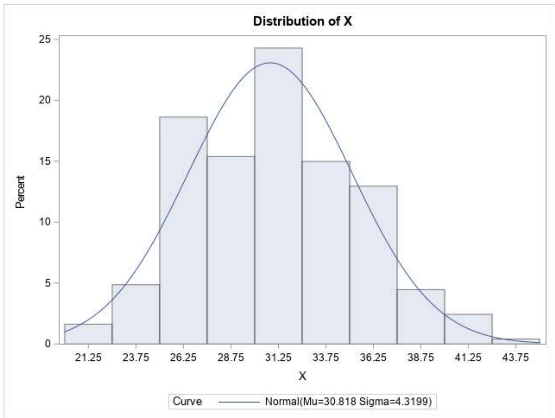


Figure 4 Y=1

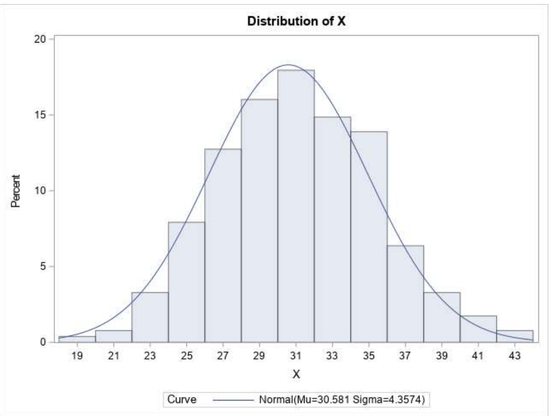


Figure 5 Y=2

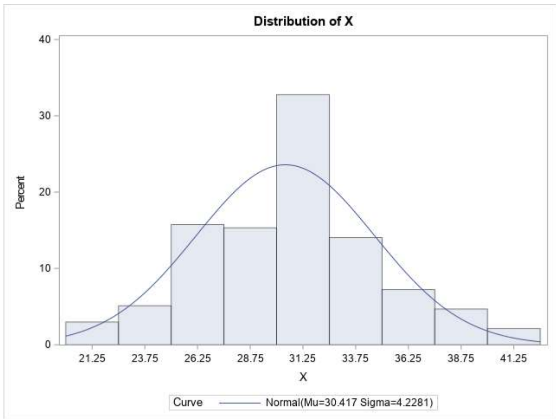
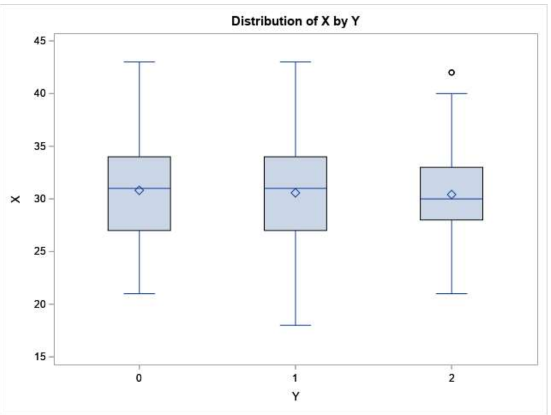


Figure 6 Boxplot



2. Distributions of X in different groups of Z

The histograms and boxplot of X based on different values of Z are shown as below.

Figure 7 Z=0

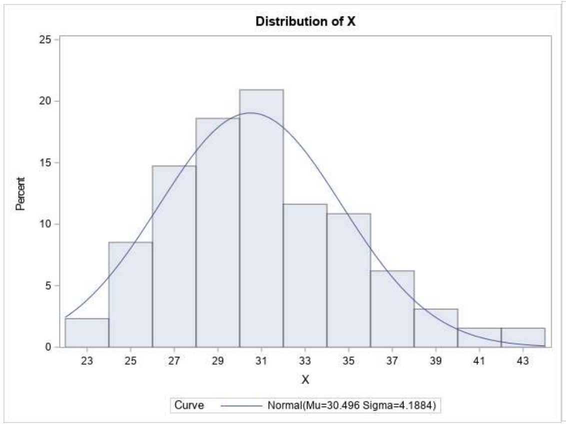


Figure 8 Z=1

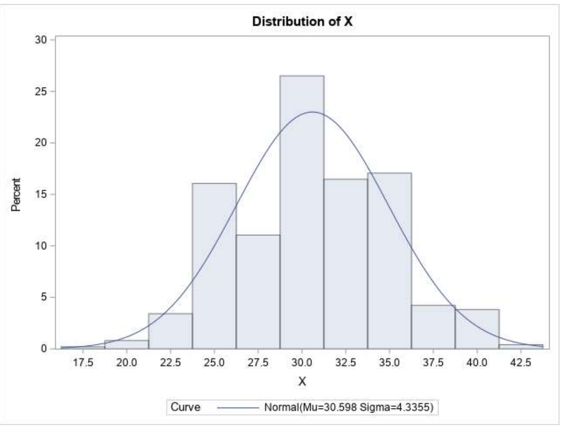


Figure 9 Z=2

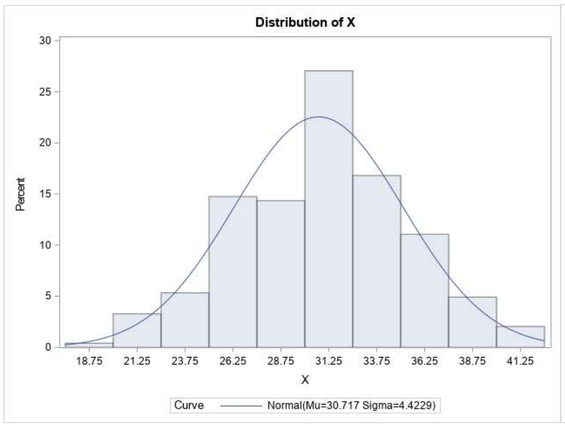
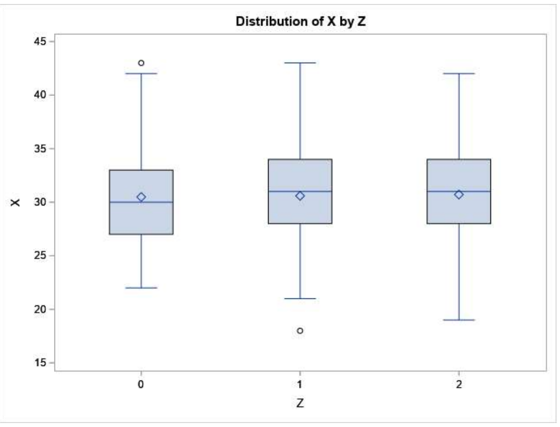


Figure 10 Boxplot



Generally, the distribution of X when Z=1 is quite similar with that when Z=2 while when Z=0, the distribution is slightly different from that in previous groups. Specifically, the mean values are, around 30.6, almost the same in different Z groups with almost the same standard deviation as 4.3. In three different groups based on values of Z, the interquartile range are same as 6. Comparing the distribution of X when Z=1 and Z=2, they have the same median as 31, the same Q1 as 28 and the same Q3 as 34. However, the maximum observation below upper fence of Z=1 is 43, which is greater than that of Z=2 as 42 with the minimum observation above lower fence, as21, being greater than that of Z=2 as 19. In addition, when Z=1, there is extreme value, as 18, below the lower fence. So X has similar distribution when Z=1 and Z=2. Comparing the distribution of X when Z=0 and Y=1&2, they do not have the same median because of 30 for Z=0 and 31 for Z=1&2; When Z=0, the maximum observation below upper fence, as 42, is lower than group Z=1 but same with group Z=2. Also, when Z=0, the Q3 is 33 which is lower than former two groups, while the Q1 is 27 which is still lower than that of Z=1&2. Besides, when Z=0, the minimum observation above lower fence, being 22, is greater than that of Z=1&2. In addition, when Z=0, there is an extreme value, as 43, above the upper fence.

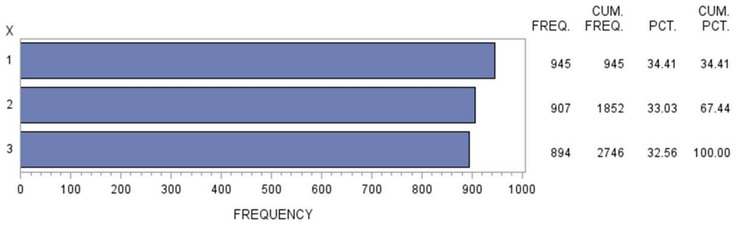
Overall, variable X has similar distributions in different groups based on variable Z although there are slight differences among these groups.

Part C

1. The bar chart for the observed frequencies (Z) with X as grouping variable

The bar chart is shown as below. It indicates that when X=1, Z is 945, being 34.41% of the total frequencies; when X=2, Z is 907, being 33.03% of the total frequencies; when X=3, Z is 894, being 32.56% of the total frequencies.

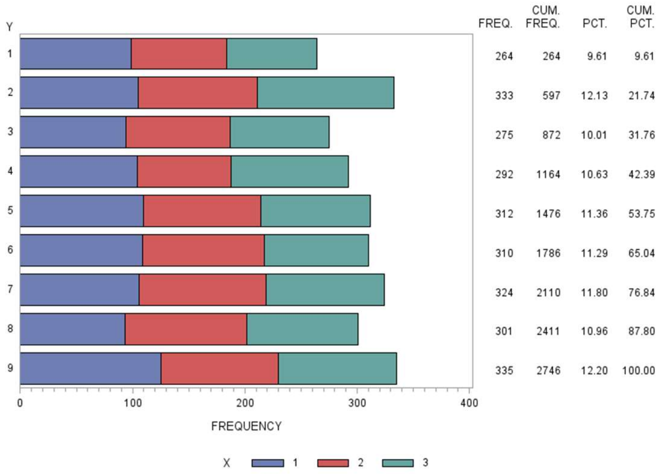
Figure 11 Bar chart of Z with X as grouping variable



2. The bar chart for Z with Y as major grouping variable and X as sub-grouping variable

The bar chart is shown as below. It indicates that when Y=1, Z is 264, being 9.61% of the total frequencies; when Y=2, Z is 333, being 12.13% of the total frequencies; when Y=3, Z is 275, being 10.01% of the total frequencies; when Y=4, Z is 292, being 10.63% of the total frequencies; when Y=5, Z is 312, being 11.36% of the total frequencies; when Y=6, Z is 310, being 11.29% of the total frequencies; when Y=7, Z is 324, being 11.80% of the total frequencies; when Y=8, Z is 301, being 10.96% of the total frequencies; when Y=9, Z is 335, being 12.20% of the total frequencies.

Figure 12 Bar chart of Z with Y as major grouping variable and X as sub-grouping variable



Appendix:

```
/*Part A*/
proc import datafile = 'E:/GW/Textbook/Data Analysis/HW1/HW1a.csv' /*read the file into sas*/
dbms = csv /*specify the format of the file*/
out=work.HW1a;     /*specify the saved dataset in sas*/
getnames=yes;     /*get the name of the variables from the original file*/
run;    /*run this procedure*/

proc print data=HW1a;    /* print the dataset saved in sas*/
run; /*run the print procedure*/

proc univariate data=HW1a mu0=33;     /*use univariate procedure to analyze variable X in HW1a and test for location u0 = 33*/
var x;       /*specify the variable X */
histogram x /normal;       /* Draw the histogram of variable X and fit the data with the normal distribution*/
qqplot x /normal(mu=30.601 sigma=4.316066);    /* Construct the Q-Q plot for variable X */
run; /* Run this procedure */

/* Part B */
proc import datafile = 'E:/GW/Textbook/Data Analysis/HW1/HW1b.csv' /*read the file into sas*/
dbms = csv /*specify the format of the file*/
out=work.HW1b;     /*specify the saved dataset in sas*/
getnames=yes;     /*get the name of the variables from the original file*/
run;    /*run this procedure*/

proc print data=HW1b;    /* print the dataset saved in sas*/
run; /*run the print procedure*/
/*/* By variable Y */
proc sort data = HW1b; /* Sort the data by variable Y with the ascending order*/
by y; /* By variable Y */
run;    /* Run this procedure */

proc univariate data=HW1b;       /*Use univariate procedure to analyze variable X in HW1b */
var x;       /*specify the variable X */
histogram x/normal;     /* Draw the histogram of variable X and fit the data with the normal distribution*/
by y;    /* By the variable Y */
run;    /* Run this procedure */
```

```
proc boxplot data=HW1b;        /* Use boxplot procedure to analyze variable X in HW1b */
plot x*y/boxstyle=schematic;    /* Plot the boxplot of variable X against variable Y */
run; /* Run this procedure */

/*/* By variable Z */
proc sort data = HW1b; /* Sort the data by variable Z with the ascending order*/
by z; /* By variable Z */
run;    /* Run this procedure */

proc univariate data=HW1b;        /*Use univariate procedure to analyze variable X in HW1b
*/
var x;        /*specify the variable X */
histogram x/normal;      /* Draw the histogram of variable X and fit the data with the normal
distribution*/
by z;    /* By the variable Z */
run;    /* Run this procedure */

proc boxplot data=HW1b;        /* Use boxplot procedure to analyze variable X in HW1b */
plot x*z/boxstyle=schematic;    /* Plot the boxplot of variable X against variable Z and
specifies the style of the box-and-whiskers plots displayed as schematic*/
run; /* Run this procedure */

/* Part C */
proc import datafile = 'E:/GW/Textbook/Data Analysis/HW1/HW1c.csv' /*read the file into
sas*/
dbms = csv /*specify the format of the file*/
out=work.HW1c;      /*specify the saved dataset in sas*/
getnames=yes;      /*get the name of the variables from the original file*/
run;    /*run this procedure*/

proc print data=HW1c;    /* print the dataset saved in sas*/
run; /*run the print procedure*/
proc gchart data=HW1c;
hbar x/freq=z discrete; /* Plot the bar chart of Z with X as grouping variable*/
run; /* */

proc gchart data=HW1c;
hbar y/subgroup=x freq=z discrete; /* Plot the bar chart of Z with Y as major grouping
variable and X as sub-grouping variable*/
run;
```