Homework 1

Jiangyan Feng (NetID: jf8)

Exercise 1

a)

Table 1 Basic Statistical Measures for Area Based on All Data

Basic Statistical Measures			
Location		**Variability**	
Mean	14.84752	**Std Deviation**	2.90970
Median	14.35500	**Variance**	8.46635
Mode	11.23000	**Range**	10.59000
		Interquartile Range	5.06000

Table 2 Moments for Area Based on All Data

Moments			
N	210	**Sum Weights**	210
Mean	14.8475238	**Sum Observations**	3117.98
Std Deviation	2.90969943	**Variance**	8.46635078
Skewness	0.39988919	**Kurtosis**	-1.0842659
Uncorrected SS	48063.7496	**Corrected SS**	1769.46731
Coeff Variation	19.5972033	**Std Error Mean**	0.20078834

The basic descriptive statistics for area for all of the data are shown as in Table 1.

Based on Table 1, the mean and median is very similar. The typical value is 14.85 in terms of mean and it is 14.36 in terms of median.

From Table 2, we observe the skewness is ~0.4, which is quite different from 0. This suggests that the data may be right skewed with long tail on the right side. This is consistent with the observation that mean is larger than median from Table 1.

From Table 1, we note that range is 10.59. This suggests the difference between maximum and minimum area is around 10.59. The difference is relatively large considering the mean and median value of the area.

b)

variety=Canadian

Moments			
N	70	**Sum Weights**	70
Mean	11.8738571	**Sum Observations**	831.17
Std Deviation	0.72300358	**Variance**	0.52273418
Skewness	0.23447131	**Kurtosis**	-0.8711666
Uncorrected SS	9905.2625	**Corrected SS**	36.0686586
Coeff Variation	6.08903724	**Std Error Mean**	0.08641546

Basic Statistical Measures			
Location		**Variability**	
Mean	11.87386	**Std Deviation**	0.72300
Median	11.83500	**Variance**	0.52273
Mode	11.18000	**Range**	2.78000
		Interquartile Range	1.18000

variety=Kama

Moments			
N	70	**Sum Weights**	70
Mean	14.3344286	**Sum Observations**	1003.41
Std Deviation	1.21570357	**Variance**	1.47793518
Skewness	-0.2356035	**Kurtosis**	0.06817022
Uncorrected SS	14485.2865	**Corrected SS**	101.977527
Coeff Variation	8.48100478	**Std Error Mean**	0.14530437

Basic Statistical Measures			
Location		**Variability**	
Mean	14.33443	**Std Deviation**	1.21570
Median	14.35500	**Variance**	1.47794
Mode	14.11000	**Range**	5.85000
		Interquartile Range	1.31000

variety=Rosa

Moments			
N	70	**Sum Weights**	70
Mean	18.3342857	**Sum Observations**	1283.4
Std Deviation	1.43949626	**Variance**	2.07214948
Skewness	-0.3618779	**Kurtosis**	-0.4397939
Uncorrected SS	23673.2006	**Corrected SS**	142.978314
Coeff Variation	7.85138992	**Std Error Mean**	0.17205271

Basic Statistical Measures			
Location		**Variability**	
Mean	18.33429	**Std Deviation**	1.43950
Median	18.72000	**Variance**	2.07215
Mode	15.38000	**Range**	5.80000
		Interquartile Range	1.82000

The basic descriptive statistics for area for 3 different varieties are shown as above.

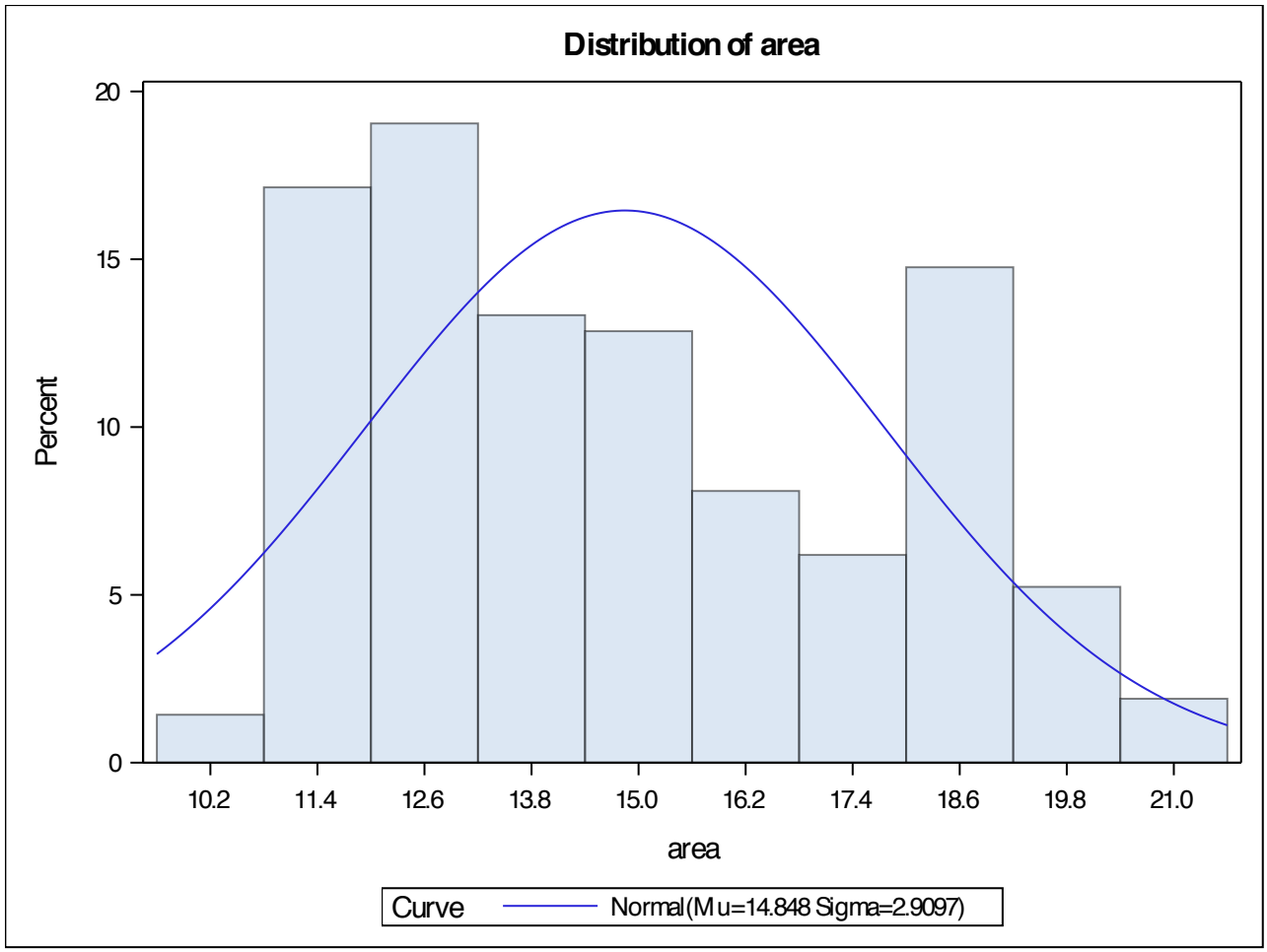
Similarities:
1. Compared with the range and variance we observed when using all data, we observe the range decreases when considering different varieties. The similar range value suggests that the variability within each variety is similar.
2. We observe that mean is similar to median in all 3 varieties. However, the skewness data for all three varieties is slightly away from 0, suggesting the existence of skewness in all three varieties.

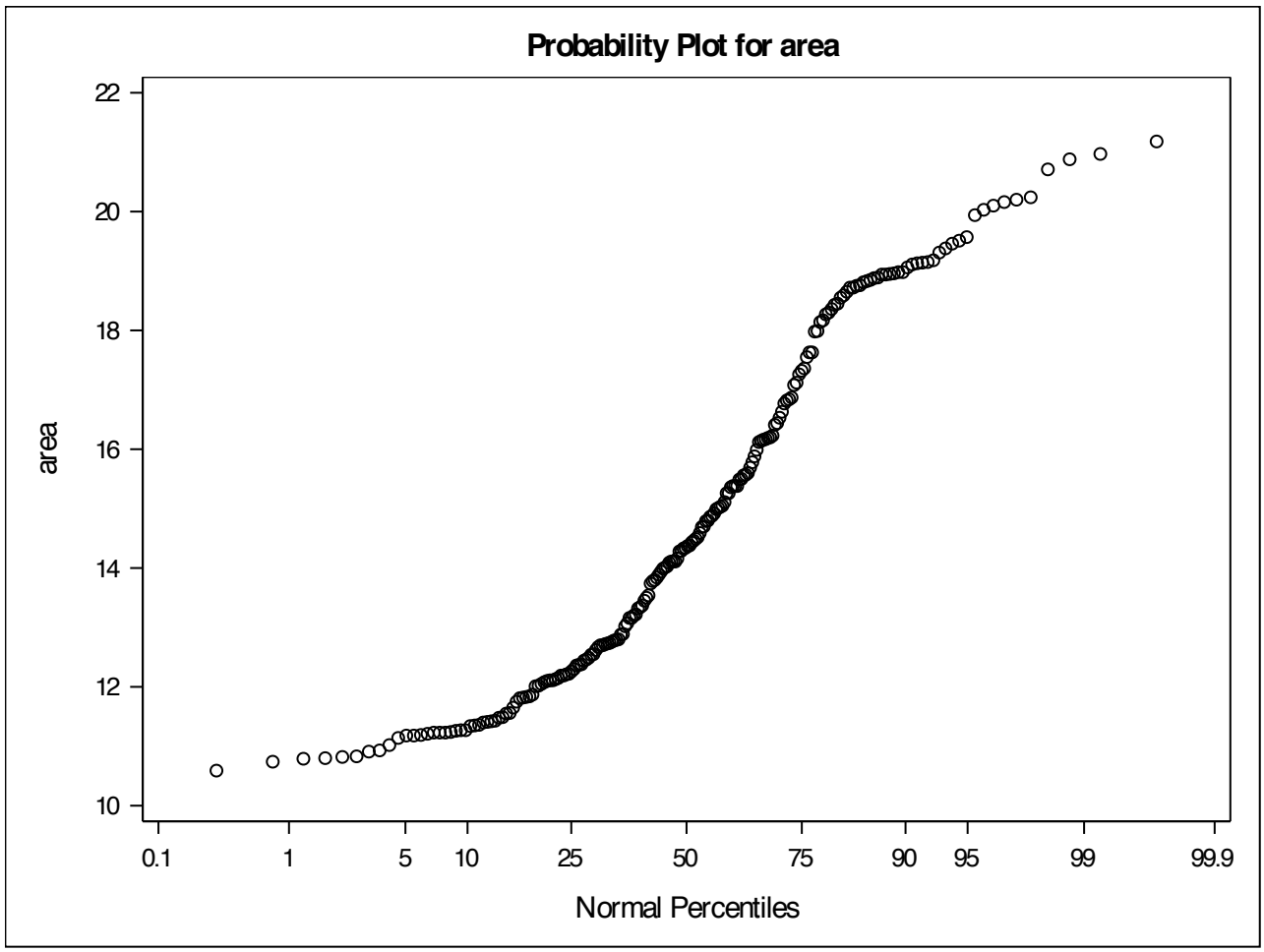
Differences:
1. Based on the mean and median value, we note that the typical values are very different. Specifically, Canadian < Kama < Rosa.
2. Based on the skewness data, the skewness for variety = Canadian is positive, suggesting a right skewed distribution. However, the skewness for variety = Kama and variety = Rosa is negative, suggesting left skewed distribution.

Exercise 2
a)

Variable: area

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.932594	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.106806	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	0.658401	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	4.448231	**Pr > A-Sq**	<0.0050





Visually:
From the histogram, we note the distribution of area data is not normal. Instead, it seems that the distribution is right skewed.
From the probability plot, we note that the line is not straight enough. It slowly grows in the very beginning. Then, the slope increases dramatically and then decreases dramatically. This may suggest a right skewed distribution. Therefore, the assumption of normality may not be reasonable.

Quantitatively:
According to Shapiro-Wilk test, p Value is very small (< 0.0001) compared with 0.05. So we conclude that it is unlikely the area distribution is normal. Therefore, the normality assumption is not reasonable.

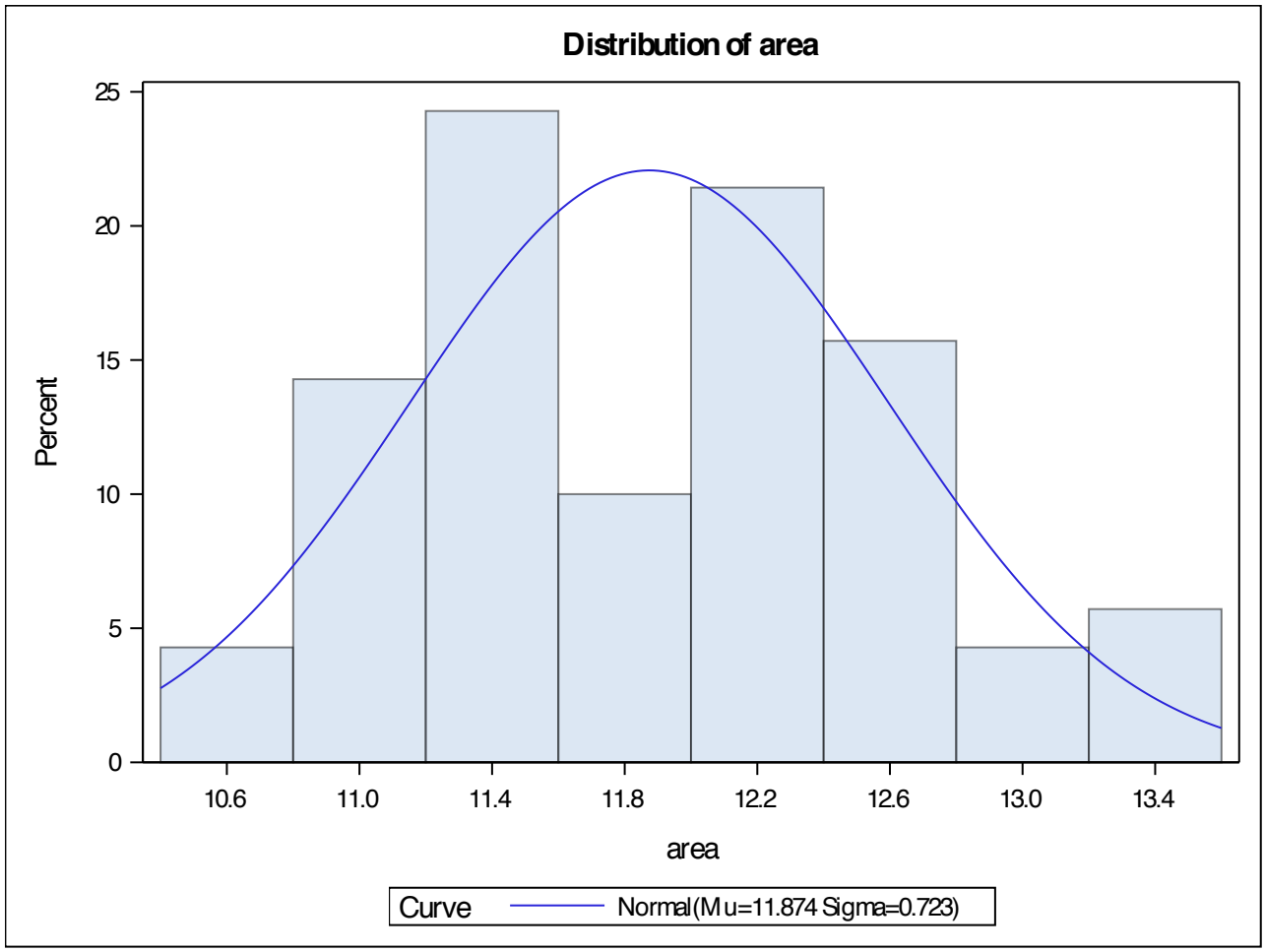
b)

Variable: area

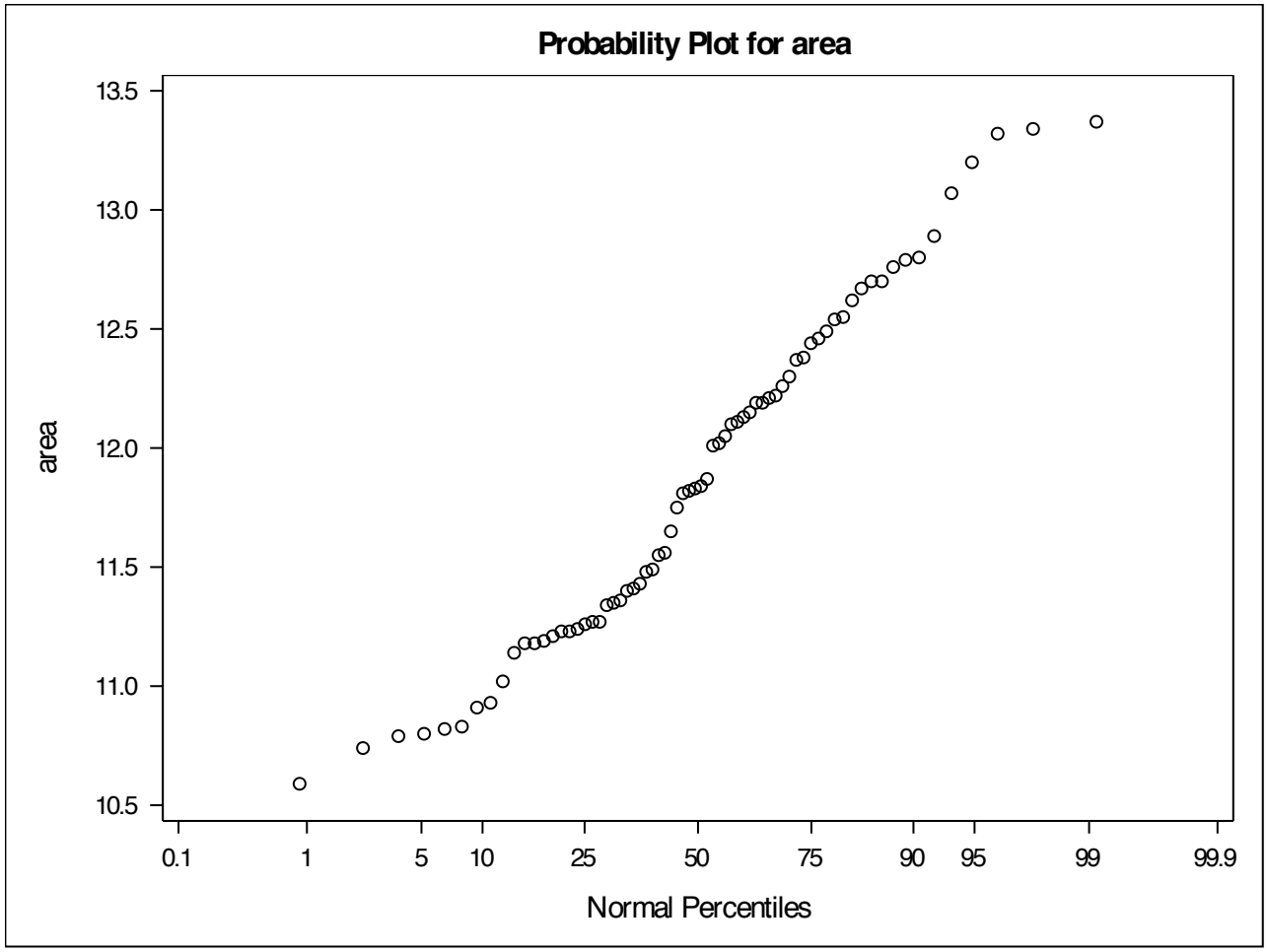
variety=Canadian

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.966281	**Pr < W**	0.0562
Kolmogorov-Smirnov	**D**	0.102263	**Pr > D**	0.0698
Cramer-von Mises	**W-Sq**	0.11633	**Pr > W-Sq**	0.0701
Anderson-Darling	**A-Sq**	0.695368	**Pr > A-Sq**	0.0702

variety=Canadian



variety=Canadian

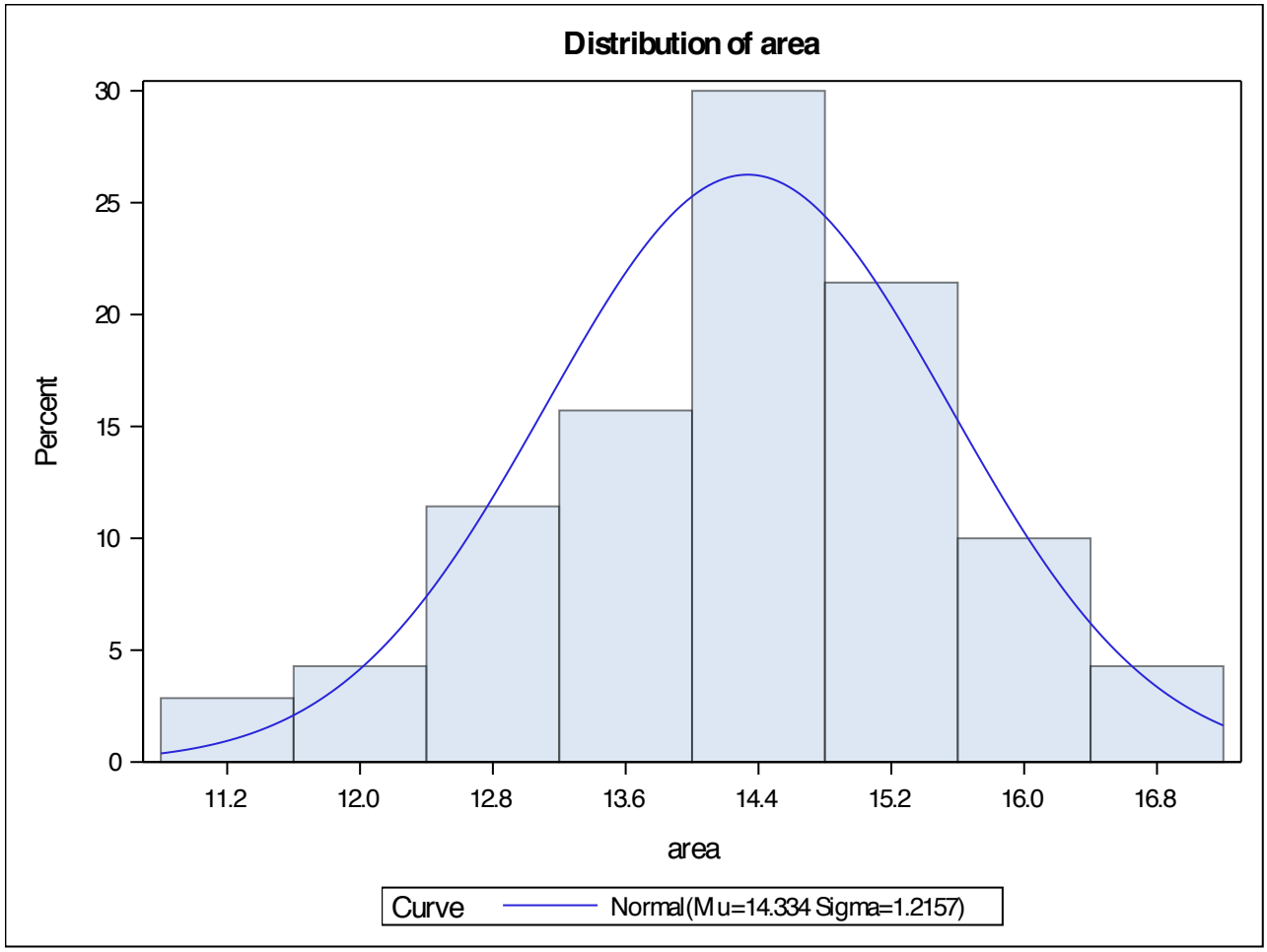


Variable: area

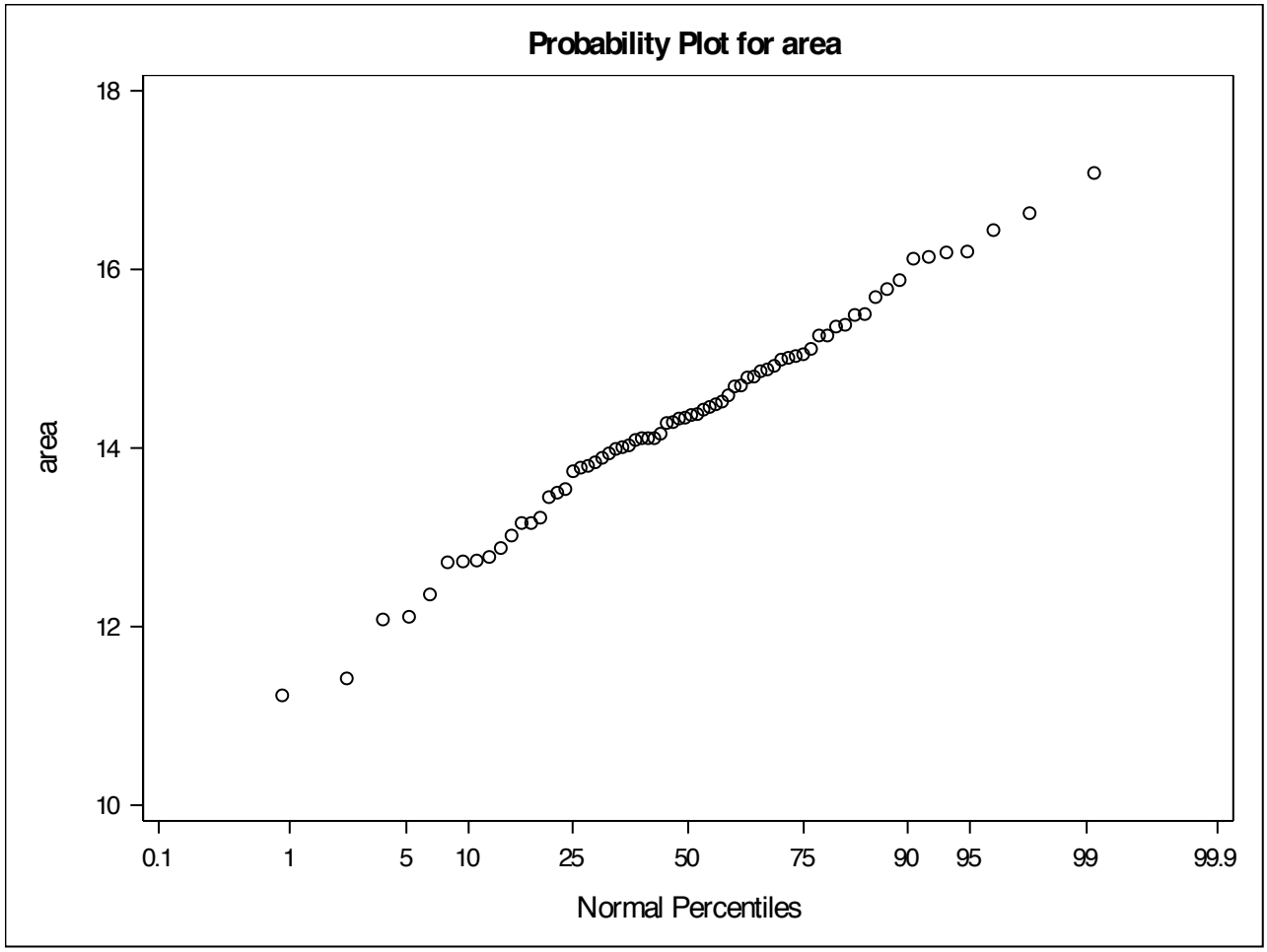
variety=Kama

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.990077	**Pr < W**	0.8581
Kolmogorov-Smirnov	**D**	0.069578	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.039198	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.234289	**Pr > A-Sq**	>0.2500

variety=Kama



variety=Kama

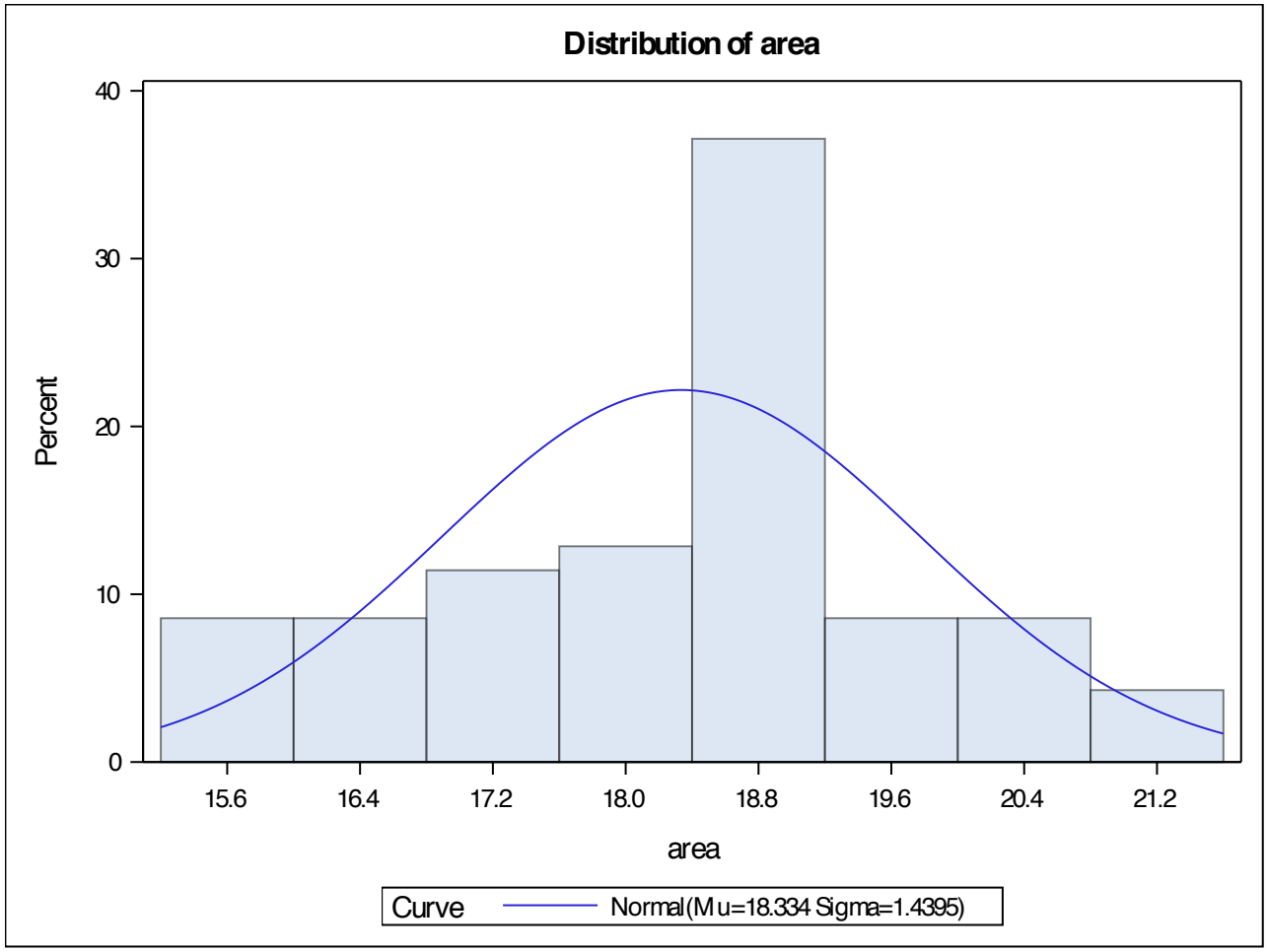


Variable: area

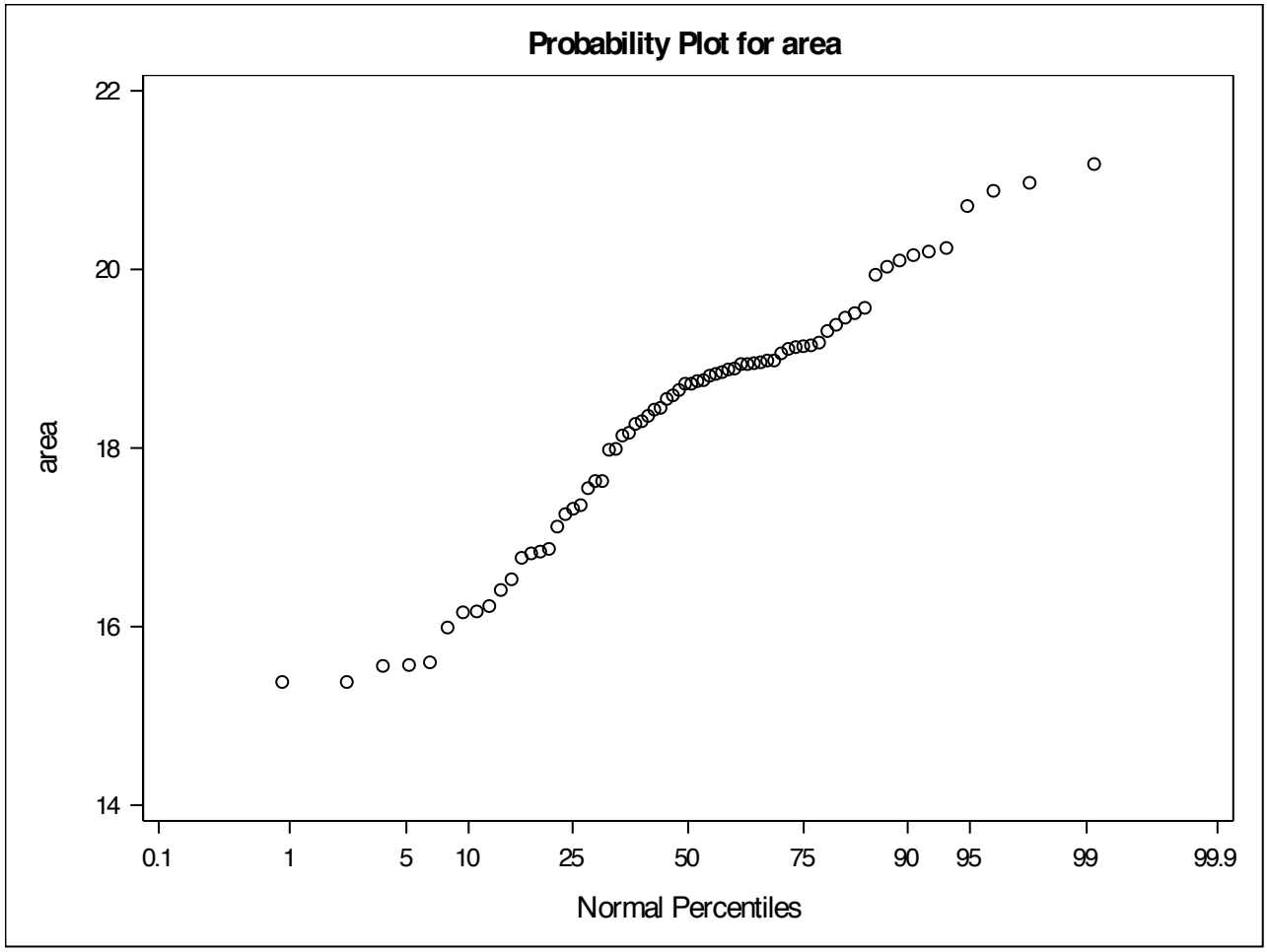
variety=Rosa

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.959688	**Pr < W**	0.0241
Kolmogorov-Smirnov	**D**	0.119917	**Pr > D**	0.0139
Cramer-von Mises	**W-Sq**	0.214445	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	1.103965	**Pr > A-Sq**	0.0067

variety=Rosa



variety=Rosa



The normality test results, histogram, and probability plot for all three varieties are shown as above.

According to Shapiro-Wilk test, W value for variety = Canadian, Kama, and Rosa is: 0.966281, 0.990077, 0.959688, respectively. Since the closer the W value towards 1, the more normal is the distribution, we note that seeds of Rosa may deviate from normality the most. Next, p value for variety = Canadian, Kama, and Rosa is: 0.0562, 0.8581, 0.0241, respectively. Only p value for Rosa is less than 0.05, whereas the p value for the other two varieties is larger than 0.05. This suggests that there is significant evidence against the null hypothesis that is the area distribution is normal for Rosa. In other words, the wheat seeds of Rosa show significant differences from normality in their area.

Exercise 3

a)

Variable: area

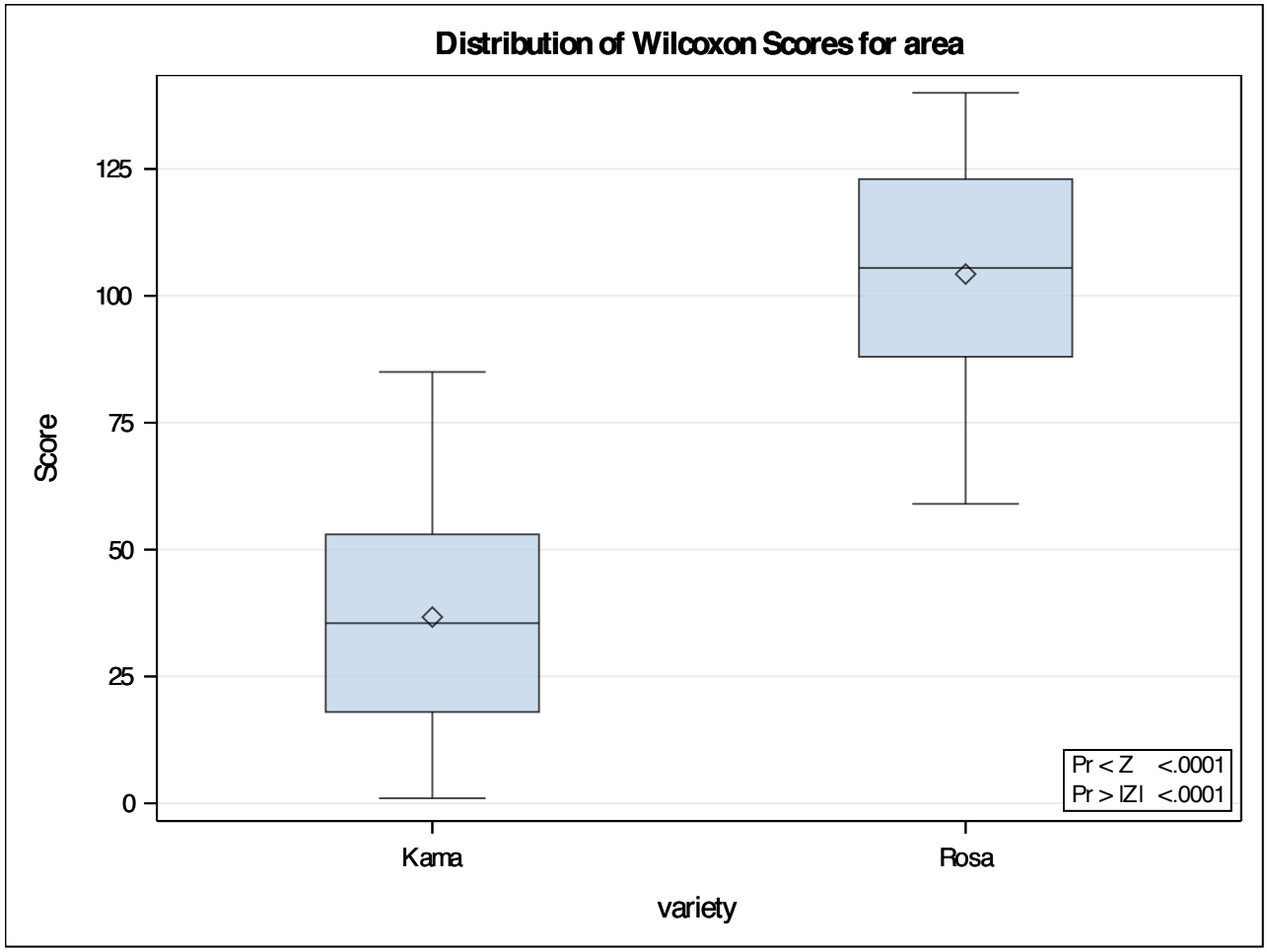
Tests for Location: Mu0=14				
Test	**Statistic**		**p Value**	
Student's t	**t**	4.220981	**Pr > |t|**	<.0001
Sign	**M**	11	**Pr >= |M|**	0.1471
Signed Rank	**S**	2790.5	**Pr >= |S|**	0.0014

From Exercise 2, we find that the area distribution of all data is not normal and not symmetric. Therefore, we obtain conclusion based on the Sign test. According to Sign test, p Value is 0.1471, which is greater than 0.05. This suggests that the median of area data should be around 14.

b)

Wilcoxon Scores (Rank Sums) for Variable area Classified by Variable variety					
variety	**N**	**Sum of Scores**	**Expected Under H0**	**Std Dev Under H0**	**Mean Score**
Kama	70	2569.0	4935.0	239.944238	36.70
Rosa	70	7301.0	4935.0	239.944238	104.30
Average scores were used for ties.					

Wilcoxon Two-Sample Test					
				t Approximation	
Statistic	**Z**	**Pr < Z**	**Pr > |Z|**	**Pr < Z**	**Pr > |Z|**
2569.000	-9.8585	<.0001	<.0001	<.0001	<.0001
Z includes a continuity correction of 0.5.					



From Exercise 2, we find that the area distribution of Rosa is not normal. Therefore, we use the Wilcoxon rank sum test. Visually, from the above distribution plot, the area for Rosa is clearly larger than Kama. Quantitatively based on the two-sided result that is Pr > |Z| is < 0.0001, this value is smaller than 0.05, suggesting that the median of these two is very different. From the one-sided result that is Pr < Z is < 0.0001, we conclude that it is unlikely for the area value of Rosa to be smaller than Kama. Taken together, we conclude that the claim that Rosa seeds have smaller area than Kama seeds is wrong.

Exercise 4
a)

Pearson Correlation Coefficients, N = 210 Prob > |r| under H0: Rho=0			
	area	**compactness**	**width**
area	1.00000	0.60829 <.0001	0.97077 <.0001
compactness	0.60829 <.0001	1.00000	0.76163 <.0001
width	0.97077 <.0001	0.76163 <.0001	1.00000

Spearman Correlation Coefficients, N = 210 Prob > |r| under H0: Rho=0			
	area	**compactness**	**width**
area	1.00000	0.63850 <.0001	0.97489 <.0001
compactness	0.63850 <.0001	1.00000	0.76273 <.0001
width	0.97489 <.0001	0.76273 <.0001	1.00000

To check the correlation among area, compactness, and width, we performed both Pearson and Spearman correlation analysis. The results from these two tests are similar.

Based on the above results, we observe that the correlation coefficients are all positive and the p values are all smaller than 0.05. Therefore, it indicates that all three variables are positively correlated. We also observe that the correlation coefficients are different for each pair of variables. Based on the magnitude, we conclude that the pairwise correlation strength can be ordered as: area vs. width > compactness vs. width > area vs. compactness.

For the population it was sampled from, we may expect similar correlations. We might infer that the area and width of the population would be strongly positively correlated. We might also expect that there is positive correlation between width and compactness. There might be positive correlation between area and compactness too, but this correlation is much weaker compared with the previous two.

b)

variety=Canadian

Pearson Correlation Coefficients, N = 70 Prob > |r| under H0: Rho=0			
	area	**compactness**	**width**
area	1.00000	0.54676 <.0001	0.86382 <.0001
compactness	0.54676 <.0001	1.00000	0.86131 <.0001
width	0.86382 <.0001	0.86131 <.0001	1.00000

Spearman Correlation Coefficients, N = 70 Prob > |r| under H0: Rho=0			
	area	**compactness**	**width**
area	1.00000	0.53005 <.0001	0.85455 <.0001
compactness	0.53005 <.0001	1.00000	0.85218 <.0001
width	0.85455 <.0001	0.85218 <.0001	1.00000

variety=Kama

Pearson Correlation Coefficients, N = 70 Prob > |r| under H0: Rho=0			
	area	**compactness**	**width**
area	1.00000	0.37104 0.0016	0.90007 <.0001
compactness	0.37104 0.0016	1.00000	0.66657 <.0001
width	0.90007 <.0001	0.66657 <.0001	1.00000

Spearman Correlation Coefficients, N = 70 Prob > |r| under H0: Rho=0			
	area	**compactness**	**width**
area	1.00000	0.34061 0.0039	0.86749 <.0001
compactness	0.34061 0.0039	1.00000	0.68059 <.0001
width	0.86749 <.0001	0.68059 <.0001	1.00000

variety=Rosa

Pearson Correlation Coefficients, N = 70 Prob > |r| under H0: Rho=0			
	area	**compactness**	**width**
area	1.00000	0.27263 0.0224	0.88049 <.0001
compactness	0.27263 0.0224	1.00000	0.64373 <.0001
width	0.88049 <.0001	0.64373 <.0001	1.00000

Spearman Correlation Coefficients, N = 70 Prob > |r| under H0: Rho=0			
	area	**compactness**	**width**
area	1.00000	0.19515 0.1055	0.83561 <.0001
compactness	0.19515 0.1055	1.00000	0.62171 <.0001
width	0.83561 <.0001	0.62171 <.0001	1.00000

To check the correlation among area, compactness, and width, we performed both Pearson and Spearman correlation analysis for all three varieties separately.

For variety = Canadian, the results between Pearson and Spearman is similar. Based on the correlation coefficients and p values, we conclude that there is positive correlation between each pair of the three measures. Specifically, the correlation strength can be ordered as: area vs. width > compactness vs. width > area vs. compactness. Difference between part a, the correlation strength between area and width is similar with the correlation between compactness and width.

For variety = Kama, the results between Pearson and Spearman is similar. Based on the correlation coefficients and p values, we conclude that there is positive correlation between each pair of the three measures. Specifically, the correlation strength can be ordered as: area vs. width > compactness vs. width > area vs. compactness. However, the correlation between area and compactness is very weak since the correlation coefficient is ~0.37. In contrast, the correlation between area and width is very strong with the coefficient ~0.90.

For variety = Rosa, the results between Pearson and Spearman is different. Since Rosa area data is not normal as shown previously, we will focus on the Spearman results. Based on the correlation coefficients and p values, we conclude that there is no

correlation between area and compactness; whereas there is positive correlation between area vs. width and compactness vs. width. Similar with two other varieties, the correlation between area and width is stronger than the correlation between compactness and width.

Compared with the results in part a, there are the following differences:
(1) The correlation coefficient between area and width for each variety is smaller compared the one with all data.
(2) The correlation between area and compactness does not exist for variety = Rosa whereas there is strong positive correlation in all data.