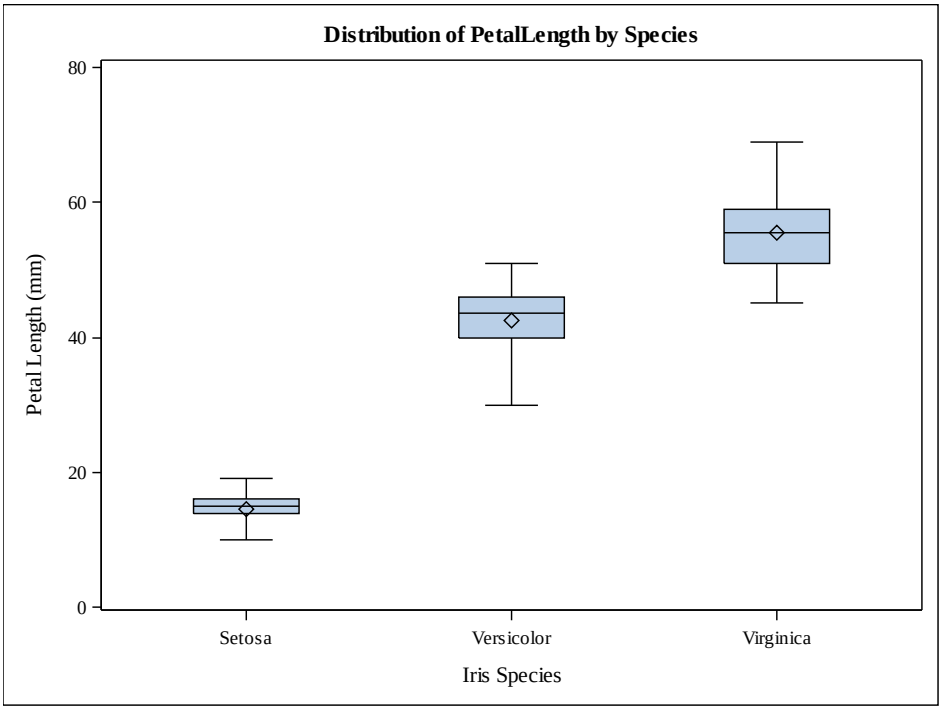
Exercise 1



a) The box plots above indicate that Setosa tends to have smaller petal lengths than the other two species and Virginica tends to larger petal lengths, though there is some overlap between Versicolor and Virginica petal lengths. The petal lengths for Setosa also seem to have small spread. From the plots alone we will not be able to determine if the differences in location or spread of petal lengths is significant across species.

b) Basic descriptive statistics are shown as follows. The mean and median values are close to 40 (37.58 and 43.5, respectively), but there is a pretty large standard deviation (17.653) and the difference between the largest and smallest lengths is 59 mm, so there is a very large spread in values across the various species. The larger value of median than mean may indicate a left skewness for the data, and in fact we a skewness of -.275 which is small but noticeable.

Moments			
N	150	**Sum Weights**	150
Mean	37.58	**Sum Observations**	5637
Std Deviation	17.6529823	**Variance**	311.627785
Skewness	-0.2748842	**Kurtosis**	-1.4021034
Uncorrected SS	258271	**Corrected SS**	46432.54
Coeff Variation	46.9744075	**Std Error Mean**	1.44135997

Basic Statistical Measures			
Location		**Variability**	
Mean	37.58000	**Std Deviation**	17.65298
Median	43.50000	**Variance**	311.62779
Mode	14.00000	**Range**	59.00000
		Interquartile Range	35.00000

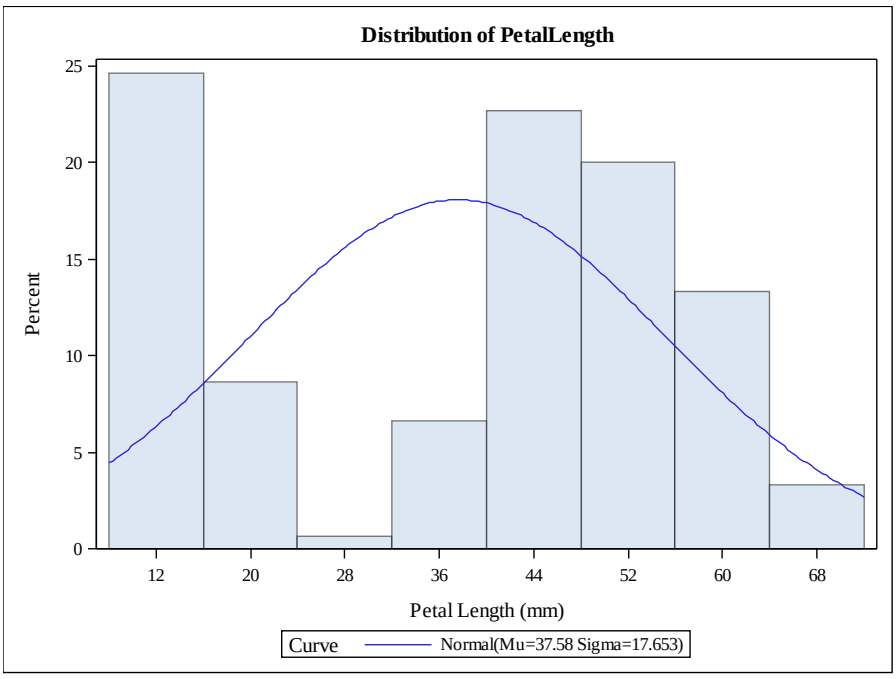
Note: The mode displayed is the smallest of 2 modes with a count of 13.

Tests of normality will give us a quantitative check, and histograms and probability plots can give us visual checks of normality.

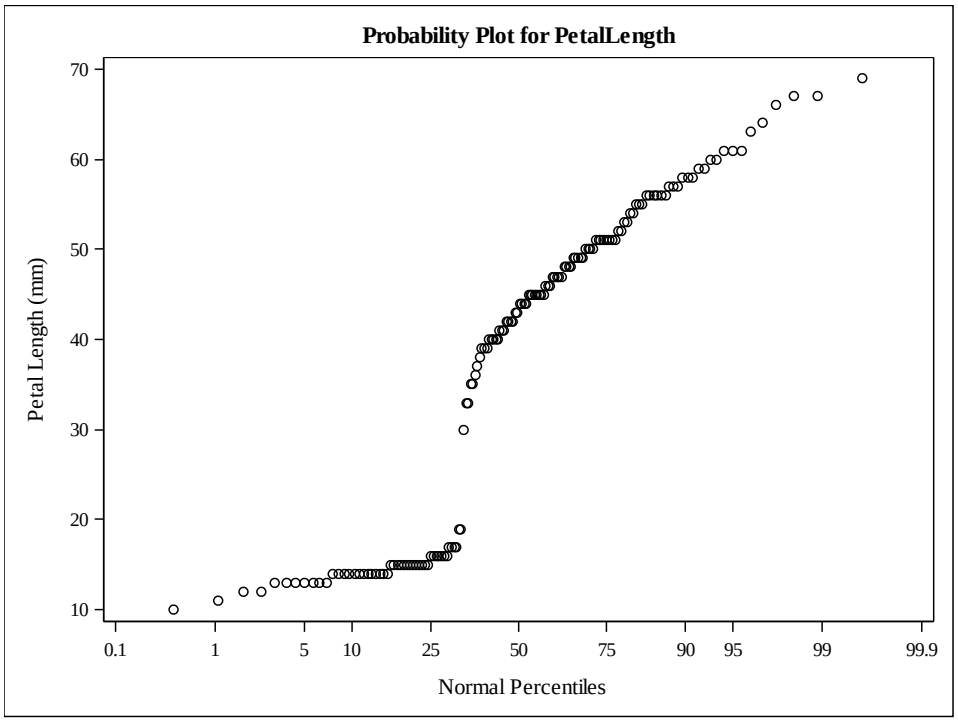
Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.876268	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.198154	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	1.222285	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	7.678546	**Pr > A-Sq**	<0.0050

All of the tests for normality are quite significant at a .05 level, indicating that an assumption that all of the petal lengths came from a single normal population should be rejected. It is pretty unlikely that a sample from a single normal population would generate this data.

The histogram and probability plot would also give us a qualitative indication that the data set deviates quite a bit from normal, and hence the underlying population probably isn't normal. We see two different maxima with a dip between them in the histogram, and the histogram does not match the estimated normal distribution very well.



The probability plot that follows also has two distinct fairly straight regions with a big jump in between. This coincides with the peaks and the gap in the histogram. If the data were reasonably normal, the points on the probability plot would fall pretty closely along a straight line.



c) Results for individual species follow. We see that the spread (standard deviation and range) within species is much smaller (e.g. standard deviations of 1.74, 4.7, and 5.52 for Setosa, Versicolor and Virginica, respectively). We also see that the mean of 14.62 and median of 15 for Setosa is much further away from the overall mean and median than are the means and medians of Versicolor and Virginica.

The skewness for Setosa is pretty small at 0.11. The skewness is more noticeable for the other two species, with Versicolor having a moderate left skew of -0.61 and Virginica having a moderate right skew of 0.55.

Discussion of normality checks follows the results.

Iris Species=Setosa

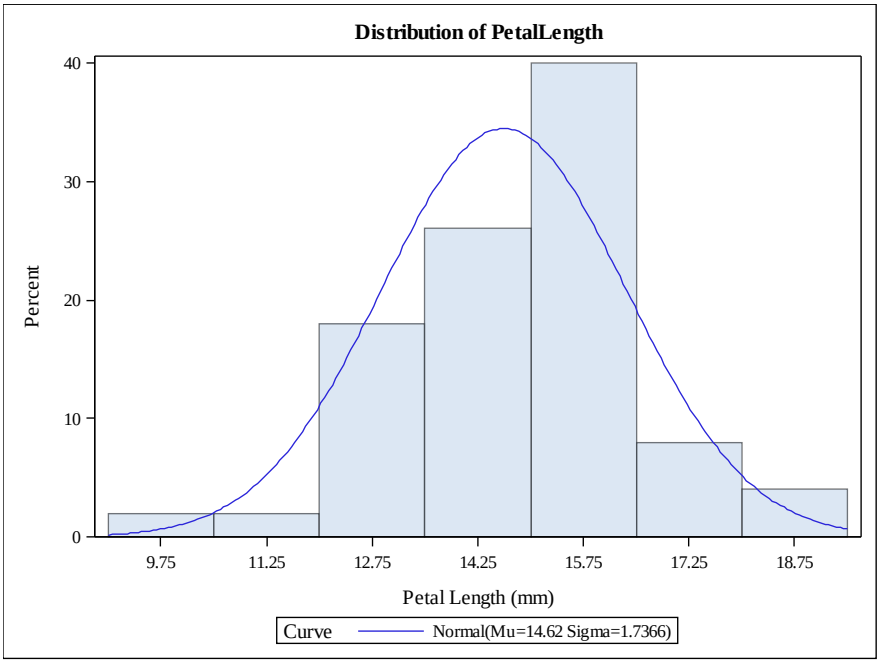
Moments			
N	50	**Sum Weights**	50
Mean	14.62	**Sum Observations**	731
Std Deviation	1.73663996	**Variance**	3.01591837
Skewness	0.1063939	**Kurtosis**	1.02157611
Uncorrected SS	10835	**Corrected SS**	147.78
Coeff Variation	11.8785223	**Std Error Mean**	0.24559798

Basic Statistical Measures			
Location		**Variability**	
Mean	14.62000	**Std Deviation**	1.73664
Median	15.00000	**Variance**	3.01592
Mode	14.00000	**Range**	9.00000
		Interquartile Range	2.00000

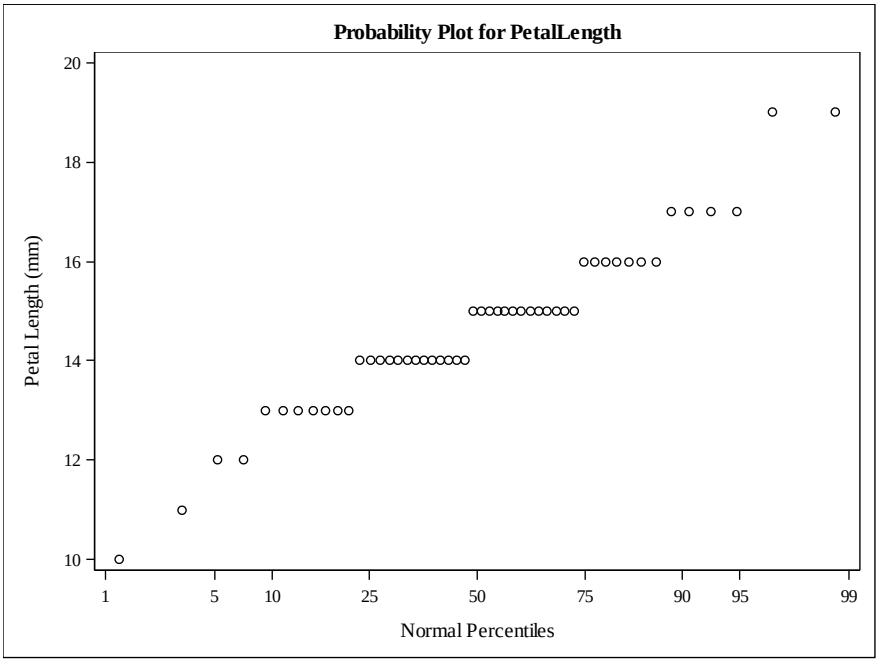
Note: The mode displayed is the smallest of 2 modes with a count of 13.

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.954977	**Pr < W**	0.0548
Kolmogorov-Smirnov	**D**	0.153398	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	0.189745	**Pr > W-Sq**	0.0070
Anderson-Darling	**A-Sq**	1.007324	**Pr > A-Sq**	0.0111

Iris Species=Setosa



Iris Species=Setosa



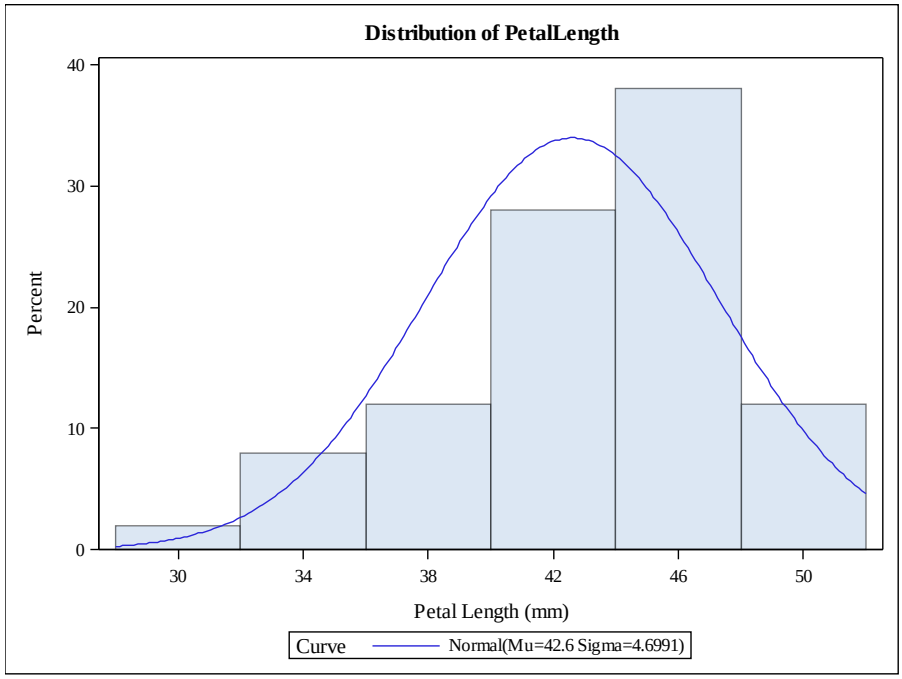
Iris Species=Versicolor

Moments			
N	50	**Sum Weights**	50
Mean	42.6	**Sum Observations**	2130
Std Deviation	4.69910977	**Variance**	22.0816327
Skewness	-0.6065077	**Kurtosis**	0.0479033
Uncorrected SS	91820	**Corrected SS**	1082
Coeff Variation	11.0307741	**Std Error Mean**	0.66455448

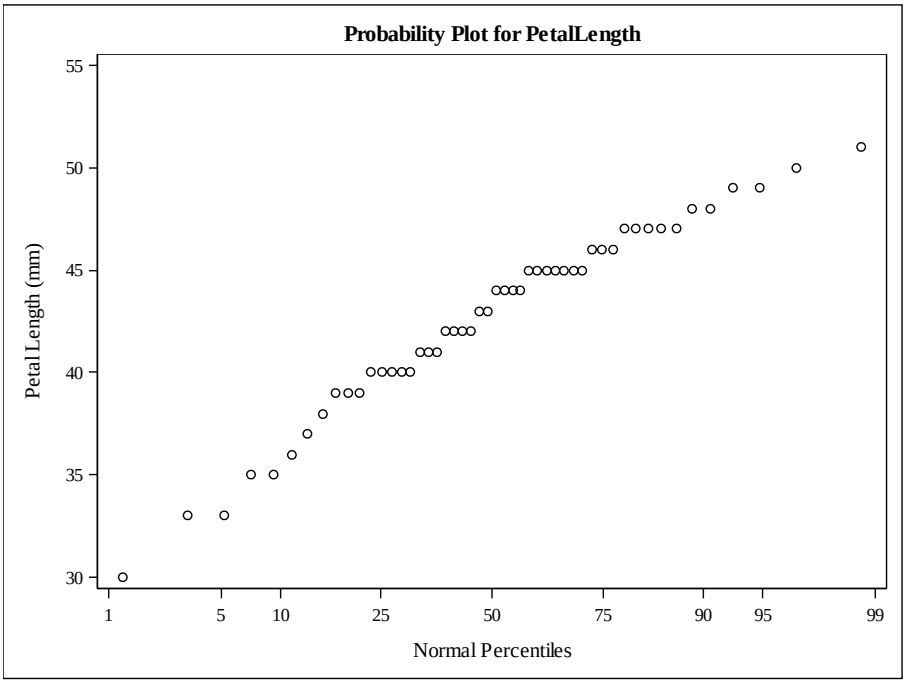
Basic Statistical Measures			
Location		**Variability**	
Mean	42.60000	**Std Deviation**	4.69911
Median	43.50000	**Variance**	22.08163
Mode	45.00000	**Range**	21.00000
		Interquartile Range	6.00000

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.966004	**Pr < W**	0.1585
Kolmogorov-Smirnov	**D**	0.117121	**Pr > D**	0.0855
Cramer-von Mises	**W-Sq**	0.090004	**Pr > W-Sq**	0.1506
Anderson-Darling	**A-Sq**	0.555056	**Pr > A-Sq**	0.1479

Iris Species=Versicolor



Iris Species=Versicolor



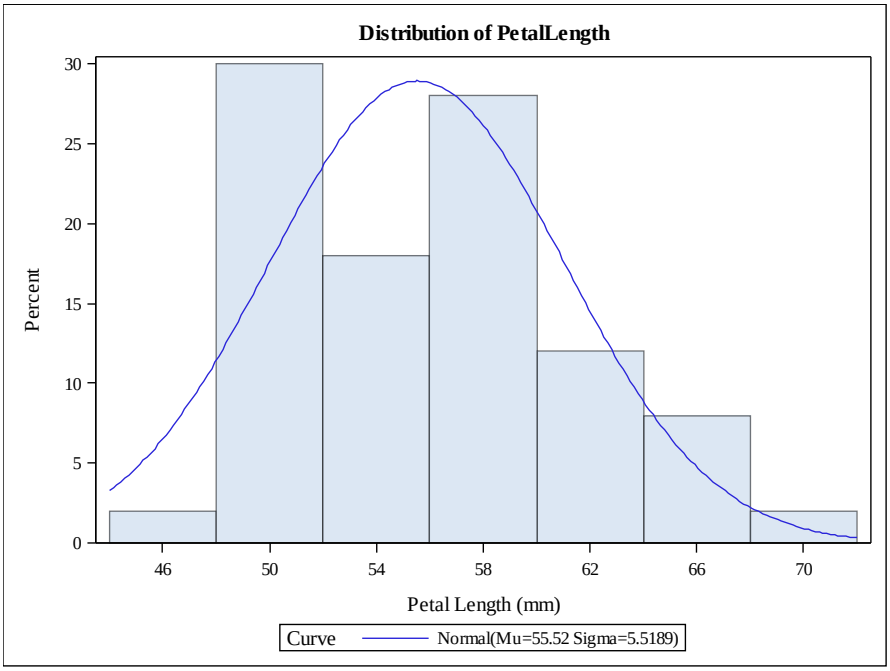
Iris Species=Virginica

Moments			
N	50	**Sum Weights**	50
Mean	55.52	**Sum Observations**	2776
Std Deviation	5.51894696	**Variance**	30.4587755
Skewness	0.54944459	**Kurtosis**	-0.1537786
Uncorrected SS	155616	**Corrected SS**	1492.48
Coeff Variation	9.94046642	**Std Error Mean**	0.78049696

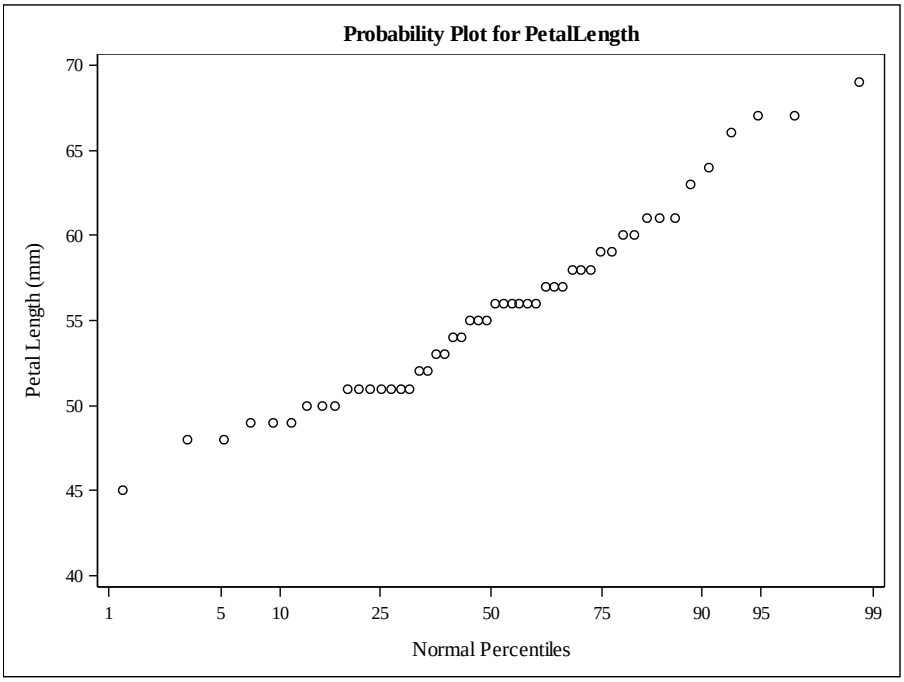
Basic Statistical Measures			
Location		**Variability**	
Mean	55.52000	**Std Deviation**	5.51895
Median	55.50000	**Variance**	30.45878
Mode	51.00000	**Range**	24.00000
		Interquartile Range	8.00000

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.962186	**Pr < W**	0.1098
Kolmogorov-Smirnov	**D**	0.113606	**Pr > D**	0.1036
Cramer-von Mises	**W-Sq**	0.086306	**Pr > W-Sq**	0.1725
Anderson-Darling	**A-Sq**	0.608956	**Pr > A-Sq**	0.1088

Iris Species=Virginica



Iris Species=Virginica



Visually, normality doesn't seem unreasonable for any of the species—the histograms do not look too different from bell shapes and the probability plots are pretty close to a straight line. Looking at the tests of normality, we see that all tests for Versicolor and Virginica are insignificant at a .05 level, so we would not reject the assumption of normality for these two species. For Setosa, though, we see some possible cause for concern. Only Shapiro-Wilk is insignificant at a .05 level, but the p-value of .0548 is pretty close to the cutoff. Shapiro-Wilk is specific to normal distributions, so it should be a better test of normality than the others. We should note that there is some indication of deviation from normality for Setosa given how significant the other test statistics are. T tests, for instance, are fairly robust with respect to small deviations from normality, so we should still be safe using a t-test on the Setosa lengths

provided the test stronger rejects or fails to reject the null. If the p-value for such a test were not far from .05, we would want to switch to a rank-based test to be safe.

Exercise 2

a) We found normality to be a very bad assumption for the full sample, so we should not trust the t test here. Signed rank assumes symmetry. We do not have a quantitative test of symmetry, but the histogram doesn't look very symmetric and the skewness was -0.27. If we consider the fact that the Versicolor and Virginica values seem to overlap a fair amount and the Setosa seem pretty well separated from the rest, symmetry is not a safe assumption, so we should rely on the sign test. The test statistic of -8 has a p-value 0.208. So we would not reject the null hypothesis and conclude that there is no significant difference between 45 and the true population median for petal lengths.

Tests for Location: Mu0=45				
Test	**Statistic**		**p Value**	
Student's t	**t**	-5.14792	**Pr > |t|**	<.0001
Sign	**M**	-8	**Pr >= |M|**	0.2080
Signed Rank	**S**	-1817	**Pr >= |S|**	0.0002

b) Virginica was reasonably close to normal, so we can rely on a t test to compare the mean of the Virginica population to the null value of 43.5 (the median for the full sample). We want a one-sided test because we want the alternative hypothesis to be that the Virginica mean is significantly greater than the null value.

DF	t Value	Pr > t
49	15.40	<.0001

With a test statistic of 15.4 and a p-value less than .0001, we would reject the null hypothesis in favor of the alternative that the mean Viriginica petal length is significantly greater than 43.5.

c) Based on the results of Exercise 1, the species Versicolor and Virginica are reasonably normal. So we can use a t test to compare the mean/median difference between Versicolor and Virginica.

Species	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Versicolor		42.6000	41.2645	43.9355	4.6991	3.9253	5.8557
Virginica		55.5200	53.9515	57.0885	5.5189	4.6102	6.8773
Diff (1-2)	**Pooled**	-12.9200	-14.9543	-10.8857	5.1254	4.4974	5.9590
Diff (1-2)	**Satterthwaite**	-12.9200	-14.9549	-10.8851			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	98	-12.60	<.0001
Satterthwaite	Unequal	95.57	-12.60	<.0001

Equality of Variances				
Method	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Folded F	49	49	1.38	0.2637

The equality of variances test has an insignificant p-value 0.2637, which means we can rely on the assumption that the two species have equal population variance. It would lead us to the pooled t-test. With a test statistic -12.6 and a p-value less than .0001, we would reject the null hypothesis in favor of the alternative that one of the populations tends to have larger values than the other. If we looked at the samples, we would see that the distribution of Virginica petal lengths tends to have larger values than the distribution of Versicolor. The confidence limits for the difference (not required for this exercise) show that Virginica petal lengths are estimated to be about 13mm larger on average, with a 95% confidence region of roughly 11mm to 15mm.

Exercise 3

a) The correlation matrix for the population of Versicolor and Virginica together shows moderate to strong positive correlations in all pairs of the four length and width measurements and all correlations are statistically significant. This tells us that if one of the measurement increases, the other three measurements will tend to increase as well. The length measures and the petal measures have pretty strong correlations with values over 0.8. The other correlations are more moderate with values between 0.5 and 0.6.

Pearson Correlation Coefficients, N = 100 Prob > |r| under H0: Rho=0				
	SepalLength	**SepalWidth**	**PetalLength**	**PetalWidth**
SepalLength Sepal Length (mm)	1.00000	0.55385 <.0001	0.82848 <.0001	0.59371 <.0001
SepalWidth Sepal Width (mm)	0.55385 <.0001	1.00000	0.51980 <.0001	0.56620 <.0001
PetalLength Petal Length (mm)	0.82848 <.0001	0.51980 <.0001	1.00000	0.82335 <.0001
PetalWidth Petal Width (mm)	0.59371 <.0001	0.56620 <.0001	0.82335 <.0001	1.00000

b) For Versicolor, the correlation matrix shows positive correlations in all pairs of the four length and width measurements similar to those in the combined data set. The length and petal measurement correlations are slightly smaller, falling between 0.75 and 0.8. The correlation of 0.66 between widths is slightly stronger than in the combined data set.

Iris Species=Versicolor

Pearson Correlation Coefficients, N = 50 Prob > |r| under H0: Rho=0				
	SepalLength	**SepalWidth**	**PetalLength**	**PetalWidth**
SepalLength Sepal Length (mm)	1.00000	0.52591 <.0001	0.75405 <.0001	0.54646 <.0001
SepalWidth Sepal Width (mm)	0.52591 <.0001	1.00000	0.56052 <.0001	0.66400 <.0001
PetalLength Petal Length (mm)	0.75405 <.0001	0.56052 <.0001	1.00000	0.78667 <.0001
PetalWidth Petal Width (mm)	0.54646 <.0001	0.66400 <.0001	0.78667 <.0001	1.00000

For Virginica, although all of the relationships are still significant at .05 level, the magnitudes of the correlation (and also the level of significance) between petal width and sepal length, petal length are much smaller than that in the combined data set. The correlation between petal and sepal lengths is, however, a bit stronger than in the combined sample.

Our main conclusions about the differences in correlations with respect to these two species is that there is a stronger proportion relationship between length measures in a Virginica and a weaker correspondence in proportions of for other pairs of length and width measurements.

Iris Species=Virginica

Pearson Correlation Coefficients, N = 50 Prob > |r| under H0: Rho=0				
	SepalLength	**SepalWidth**	**PetalLength**	**PetalWidth**
SepalLength Sepal Length (mm)	1.00000	0.45723 0.0008	0.86422 <.0001	0.28111 0.0480
SepalWidth Sepal Width (mm)	0.45723 0.0008	1.00000	0.40104 0.0039	0.53773 <.0001
PetalLength Petal Length (mm)	0.86422 <.0001	0.40104 0.0039	1.00000	0.32211 0.0225
PetalWidth Petal Width (mm)	0.28111 0.0480	0.53773 <.0001	0.32211 0.0225	1.00000