Getting Comfortable with your Data in SAS - Visualization and Descriptive Statistics

A Michelle Edwards, Ph.D., MLIS

September 30, 2019

- Quick Review
- Research Question
- Data Visualization - what is it?
  - Five General Principles behind Data Visualization:
  - Graphic Design Principles
- Data Types
  - Quantitative
  - Qualitative
- Digging into Visualizing Examples
  - Visualizing ONE number (univariate)
    - Bar Chart
    - Histograms
    - Boxplots
  - Visualizing MORE than ONE number (bivariate or multivariate)
    - Scatterplot
    - Bar chart with means
    - More SAS graph examples
- Descriptive Statistics
  - PROC UNIVARIATE
  - PROC MEANS
  - PROC FREQ
- Workshop Review
  - Data Visualization
  - Descriptive Statistics

Note: This is part of my Computer setup - you will NOT see this!! Please ignore

```
## SAS found at C:/Program Files/SASHome/SASFoundation/9.4/sas.exe
```

```
## sas, saslog, sashtml, sashtml5, and sashtmllog & sashtml5log engines
```

```
##    are now ready to use.
```

End of Note

Quick Review

Last workshop we talked about collecting our data, cleaning it to a point that would allow us to enter it into Excel and eventually into SAS or R.

Items we discussed included:

- Best practices for naming our variables or column headings
- Knowing your data - what were the expected ranges of values? typos?
- How to get the data into SAS

Research Question

Remember how all of our research is driven by a research question. We do not go out and collect data just for the fun of it. We collect data with a purpose and that purpose will be linked back to your research question.

Now that we have our data collected and loaded into our statistical package, let's start to get more comfortable with it, by visualizing it and then running some descriptive statistics.

Data Visualization - what is it?

Data visualization can mean many different things to different people. To me, it all comes down to the purpose. WHY? and for WHOM? Think of data visualization as a tool to help you tell your story. I've always seen data analysis or statistical analysis as a way to help me to tell the story of my research. There are so many "tools" that you can use, so let's think of data visualization and statistical analyses as 2 tools that you will use to tell your research story.

Remember that visualizing your data is "critical to the understanding of the contents of your data." We use data visualization to help us examine, scrutinize and validate our data.

There are a great many resources out there for us to use as a guide to the world of Data Visualization. I will use 2 in particular - but please use the ones that speak to you. The resources I like to use are:

- "Effective Data Visualization: The Right Chart for the Right Data" (2017) by Stephanie D.H. Evergreen
- "Making Sense of Data II: A Practical Guide to Data Visualization, Advanced Data Mining Methods, and Applications" (2009) by Glenn J. Myatt and Wayne P. Johnson

Five General Principles behind Data Visualization:

1. Show the data
2. Simplify - you want to keep the message simple and remove all the flowery bits of your visualizations
3. Reduce the clutter - do you REALLY need all those grids or ticks on your graph??
4. Revise your visualizations - creating a graph or a table should be viewed as part of your writing. You don't write something and leave it. You write, you revise, you may write again and revise again. Treat you visualizations in the same manner.
5. Be HONEST! This may sound funny - but let's face it, there are times where the visualizations we create may have an element of exaggeration added in. Those y-axes - where do they start? at 0 or somewhere else? Are we exaggerated the differences between those lines?

Graphic Design Principles

Yes there are some graphic design principles that you may want to consider as you begin to think about creating your data visualizations. This is a link to a powerpoint presentation that I use when talking about these principles. Please take a few minutes when you have the time to review them. Each has a visual to explain the principle.

Data Types

The data that we colect from our research projects may not always be the same type. What do I mean by a Data Type? Let's review the basic types:

Quantitative

- Measures a quantity
  - Continuous
    - a measure of something
  - Categorical
    - Nominal: you are in one group or another, there is NO order to the groups
    - Ordinal: you are in one group or another, there IS an order to the groups
- Examples of:
  - Continuous: height, weight, age,
  - Nominal: Yes/No, Program of Study, Gender
  - Ordinal: shirt size, LIkert scale, Year of study

Qualitative

- Measures a quality
- Descriptive, words,

Digging into Visualizing Examples

Let's begin by loading data into our program - SAS. The data we will be using is the Fisher's Iris dataset that is part of the SAS program. To view this dataset please use the following SAS statements.

```
Data iris;
  set sashelp.iris;
Run;

Proc print data=iris (obs=15);
Run;
```

Obs	Species	Sepal Length	Sepal Width	Petal Length	Petal Width
1	Setosa	50	33	14	2
2	Setosa	46	34	14	3
3	Setosa	46	36	10	2
4	Setosa	51	33	17	5
5	Setosa	55	35	13	2
6	Setosa	48	31	16	2
7	Setosa	52	34	14	2
8	Setosa	49	36	14	1
9	Setosa	44	32	13	2
10	Setosa	50	35	16	6
11	Setosa	44	30	13	2
12	Setosa	47	32	16	2
13	Setosa	48	30	14	3

```
14    Setosa       51        38       16        2
15    Setosa       48        34       19        2
```

Note that there are 5 variables:

- Species of the Iris measured
- Sepal length measured
- Sepal width measured
- Petal length measured
- Petal width measured

Exercise 1

1. What type of data are each of the 5 variables available in the Iris dataset?

Visualizing ONE number (univariate)

Let's start with the Species variable. We have 1 variable - so univariate - and we want to see what's happening with it.
What kind of visualizations could you do?

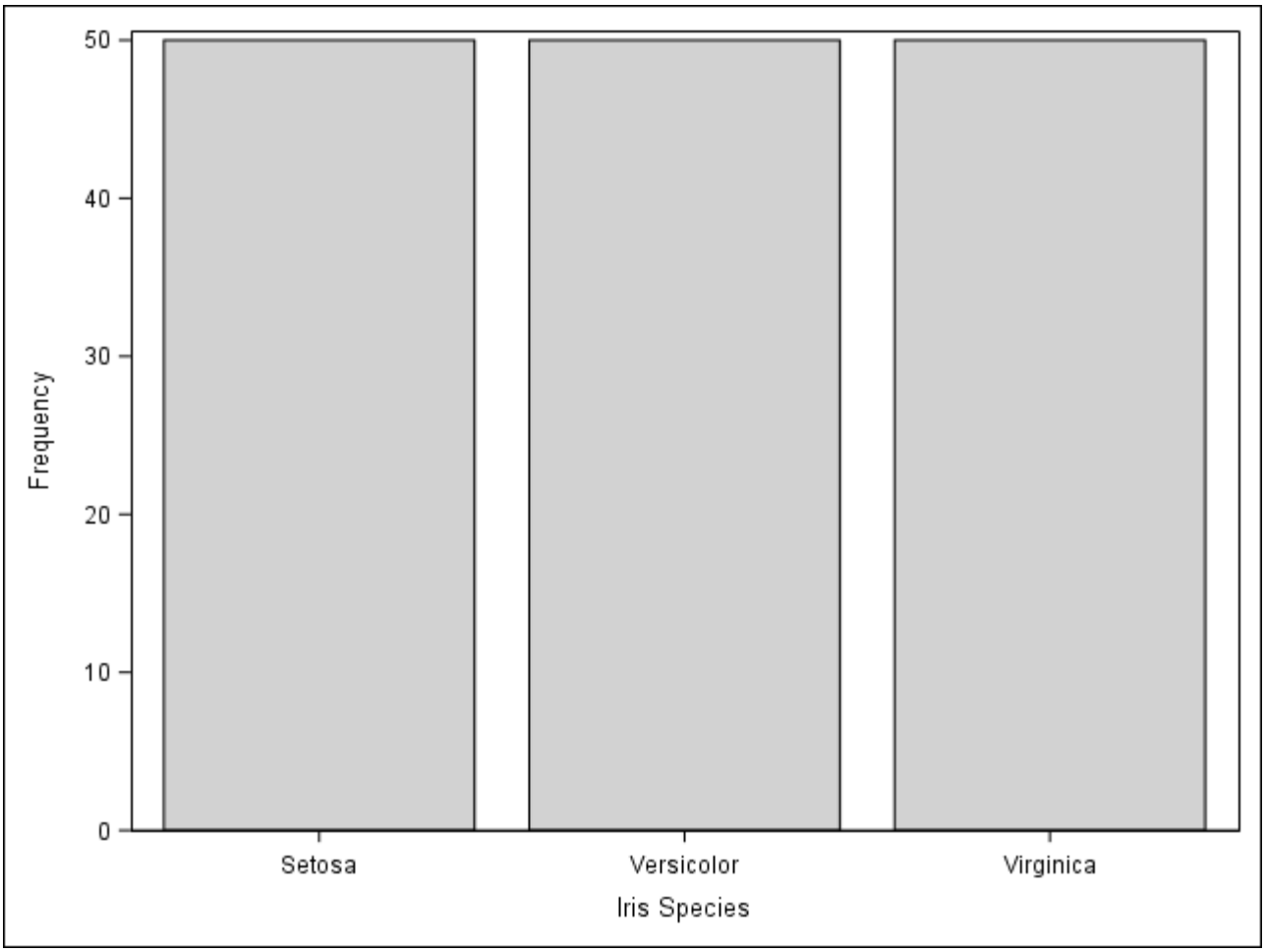
- __
- __
- __

Bar Chart

Bar chart can show us the frequency or the count of observations in Species, and we could also break it up for each species. In SAS, to create a bar chart we will use the PROC SGPLOT

```
Data iris;
  set sashelp.iris;
Run;

Proc sgplot data=iris;
  vbar species;
Run;
```



Exercise 2

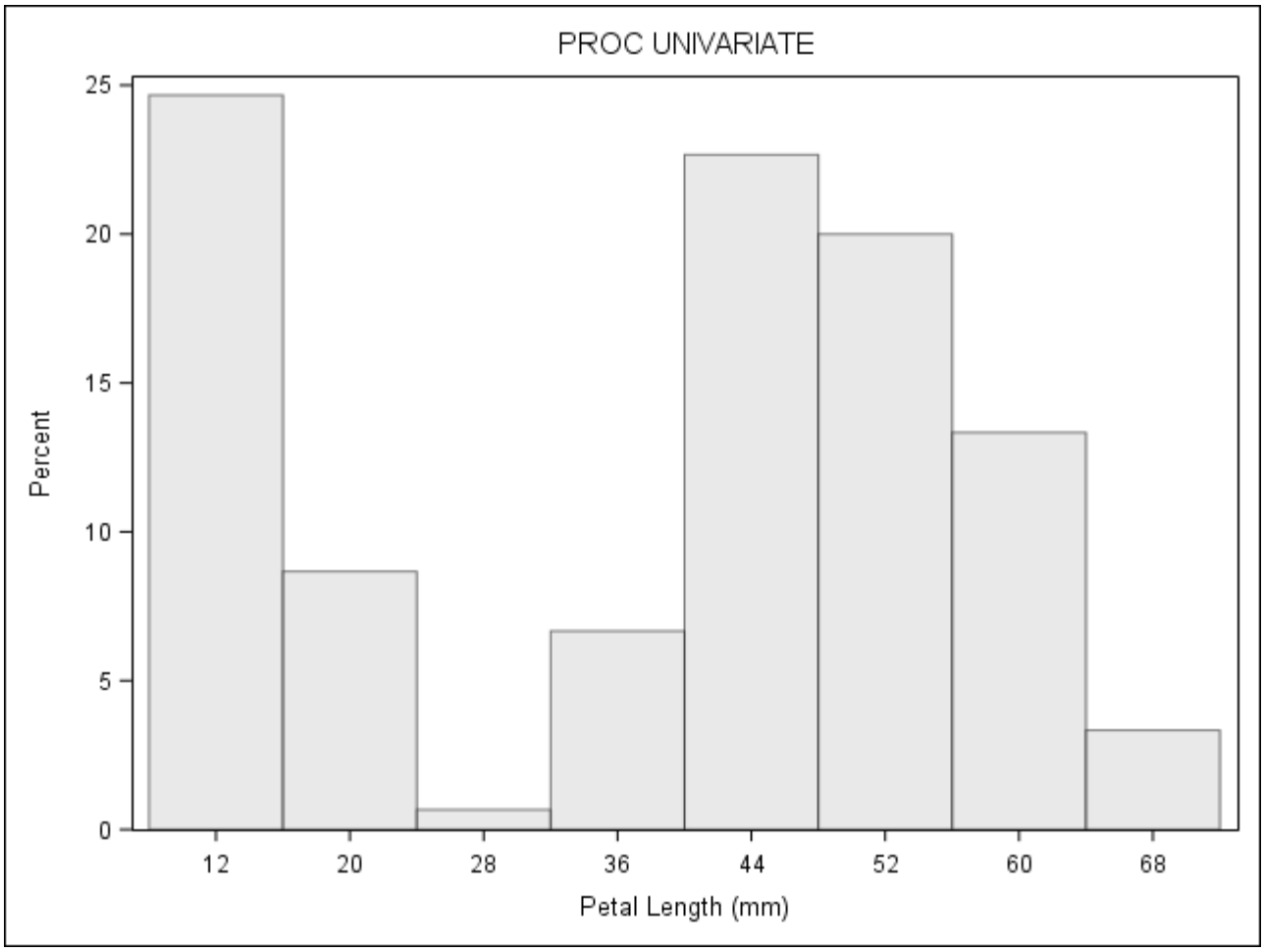
1. What story does this graph tell?
2. When would you use a barchart with data that are continuous? Such as petal width?

Histograms

SAS source code modfied from SAS blogs

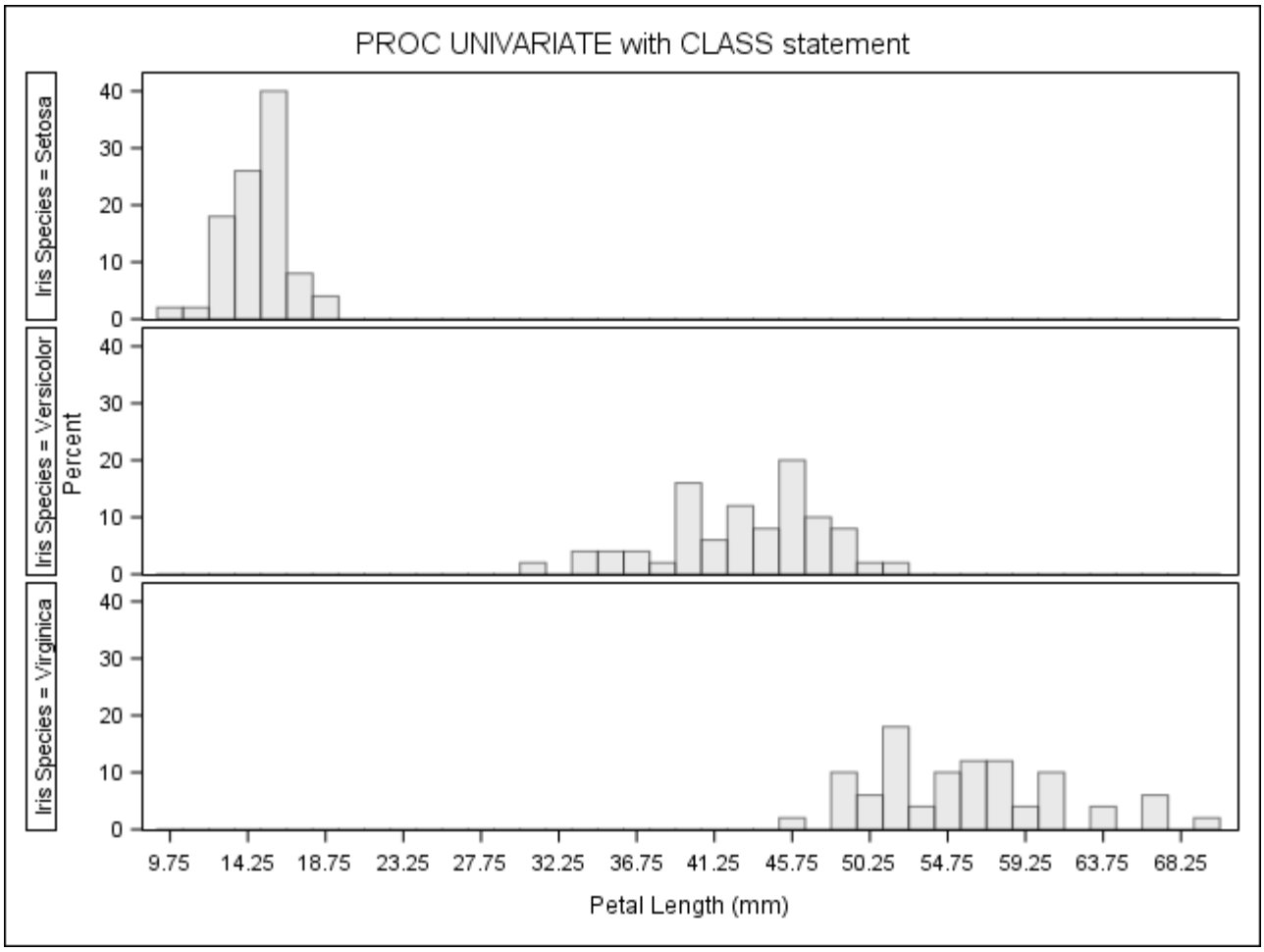
What is a histogram? Show frequency or counts of your data in a group or an interval. Generally used for continous data. Let's try it on petal length.

```
proc univariate data=sashelp.iris;
  var PetalLength;
  histogram PetalLength / odstitle="PROC UNIVARIATE";
  ods select histogram;
run;
```



We know that we have 3 species of Irises in this dataset, so let's try to pull them apart, by creating a histogram for each species, by adding a **CLASS** statement to the code we just used.

```
proc univariate data=sashelp.iris;
  class species;
  var PetalLength;
  histogram PetalLength / nrows=3 odstitle="PROC UNIVARIATE with CLASS statement";
  ods select histogram;
run;
```



This exercise has 2 purposes:

- to visualize the data
- to learn SAS coding to visualize out data

Exercise 3

1. Working with a partner, review the last SAS code and take turns explaining what each line means.
2. Create another set of histograms for Sepal Length.

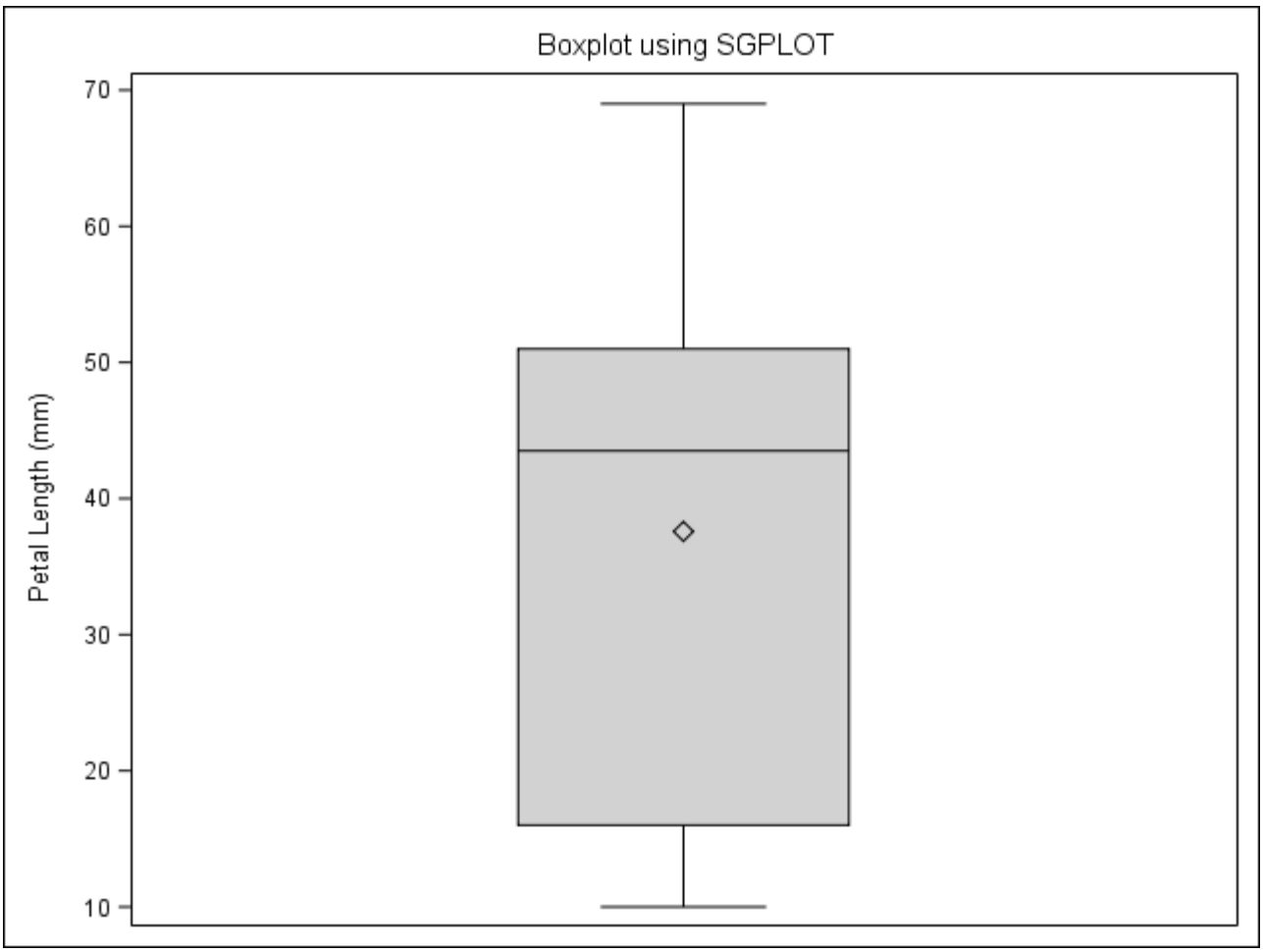
There are more options available to you as you work with histograms, please see the SAS blog posted above for more information and ways to visualize these data with histograms.

Boxplots

What is a boxplot? A standardized, visual representation of the distribution of your data. Oh now that's a mouthful! It's a way of looking at your data to see whether the values you've collected are evenly distributed across the range you collected. Oish! How bout this - bell curve - but in a box format? A classic visual of a box plot will show you the 4 quartiles of your data - so where is 25% of your data, 50% (median), 75%. It will also show you where the mean of your data resides. It can also show you if you have any outliers - any data points that are beyond the expected normal distribution. I know you've all seen these, and you will use them when or rather if you conduct any GLMM analysis. These are a very handy plot to help examine your residuals.

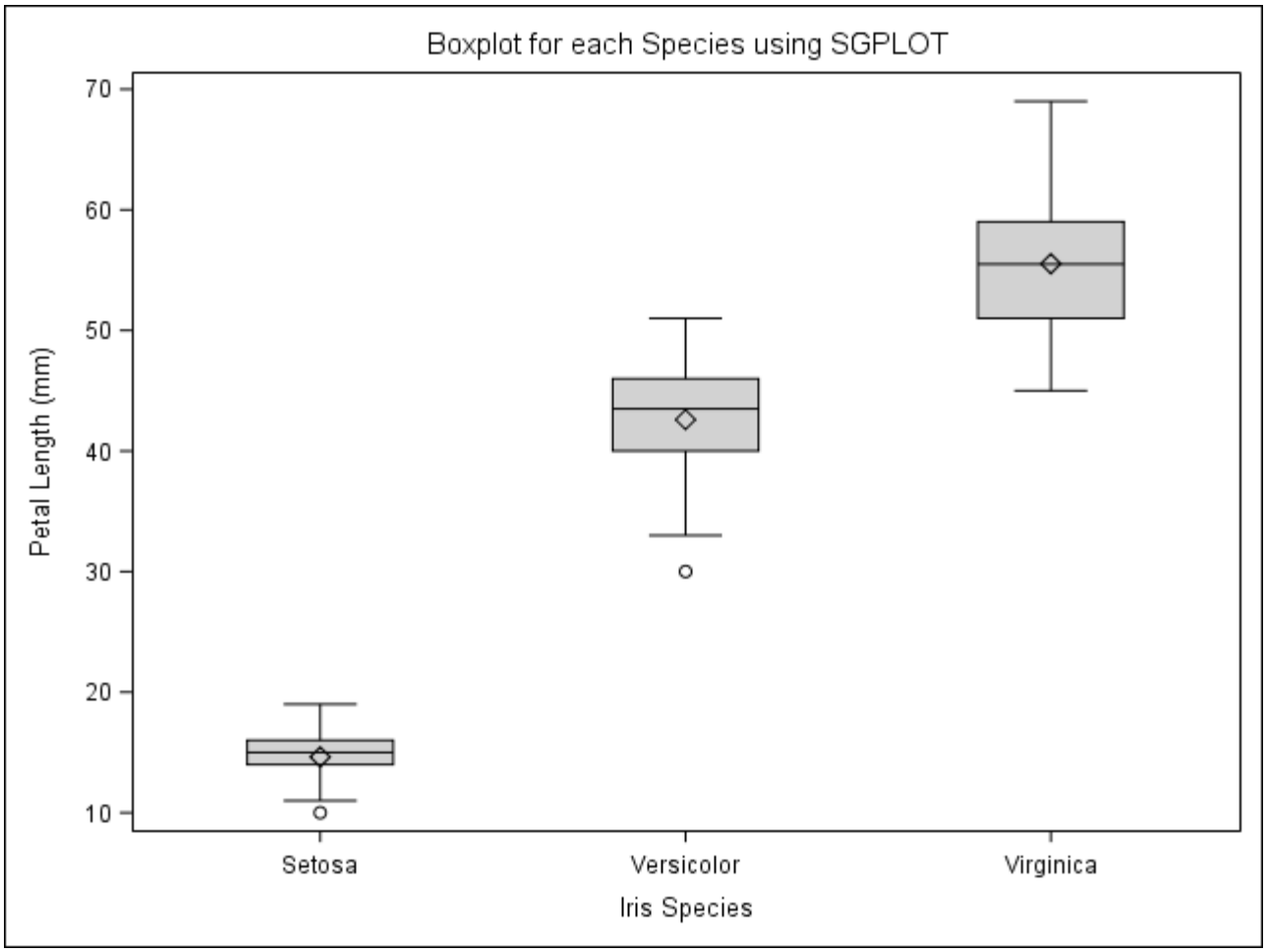
Alrighty, less chatting and more doing. Let's continue to work with the Fisher's Iris dataset and try a basic boxplot using the PROC SGPLOT.

```
proc sgplot data=sashelp.iris;
   vbox PetalLength ;
   title "Boxplot using SGPLOT";
run;
```



Let's try that boxplot again, but asking for each species to be separate.

```
proc sgplot data=sashelp.iris;
   vbox PetalLength /category=Species;
   title "Boxplot for each Species using SGPLOT ";
run;
```



Exercise 4

1. With a partner, review these last plots and describe what you are seeing? What story is the data trying to tell?
2. Starting to think about descriptive statistics, what numbers/values/statistics might you want to add here to help fill in the story?

Visualizing MORE than ONE number (bivariate or multivariate)

When we have more than one variable that we are interested in plotting at the same time, then we have more options. Maybe a scatterplot to see if there is a relationship between the variables, or a side-by-side barchart to see how things may be different? Let's take a quick peak at a scatterplot and a barchart.

Scatterplot

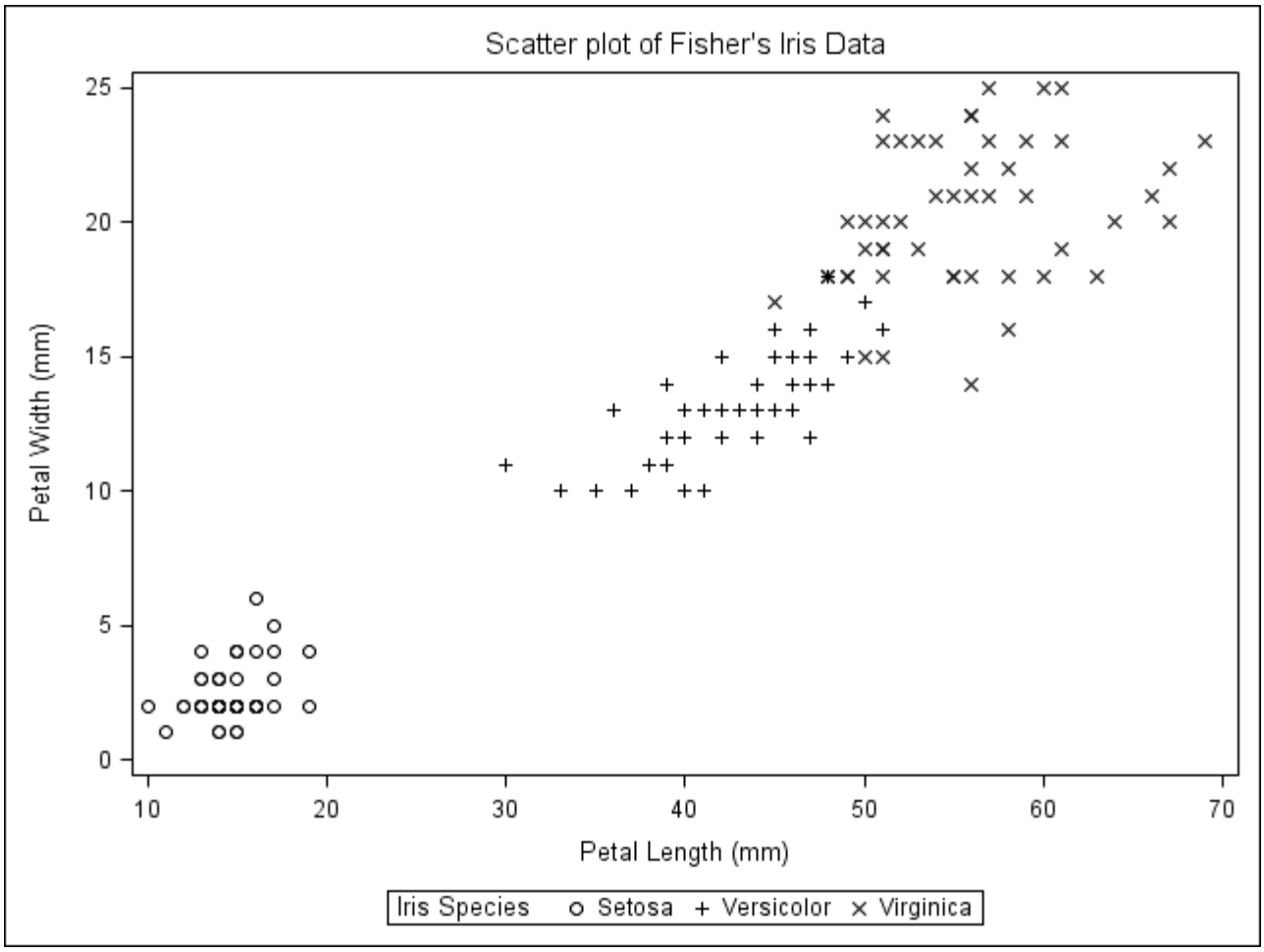
As the name suggests we are creating a plot with one variable as the Y-axis and the second variable as the X-axis. Let's try a scatterplot with petal width as our Y and petal length as our X.

```
proc sgplot data=sashelp.iris;
   scatter x =petallength y=petalwidth;
   title "Scatter plot of Fisher's Iris Data ";
run;
```

The PROC SGPLOT creates a plot that allows us to modify the type of dot placed on the graph, its colour, its size, etc… If you need to use PROC PLOT, note that the dot placed on the graph is a letter, A - Z, representing how many values are on the same spot.

Our first plot has all data, but let's try to separate the 3 species once again. With the boxplot we used a category statement. With the scatterplot we are going to use a group statement.

```
proc sgplot data=sashelp.iris;
   scatter x =petallength y=petalwidth / group=species;
   title "Scatter plot of Fisher's Iris Data ";
run;
```



You will notice that the plot produced in these notes, the default change was the shape of the dot for each species. Due to the method that I am using to run SAS, this will be the default. However, on your laptops, the default may be a change in colour. Blue circles for Setosa, red circles for Versicolor, and green circles for Virginica. If you are using the Animal Biosciences SAS server - your output will match the output in these notes.

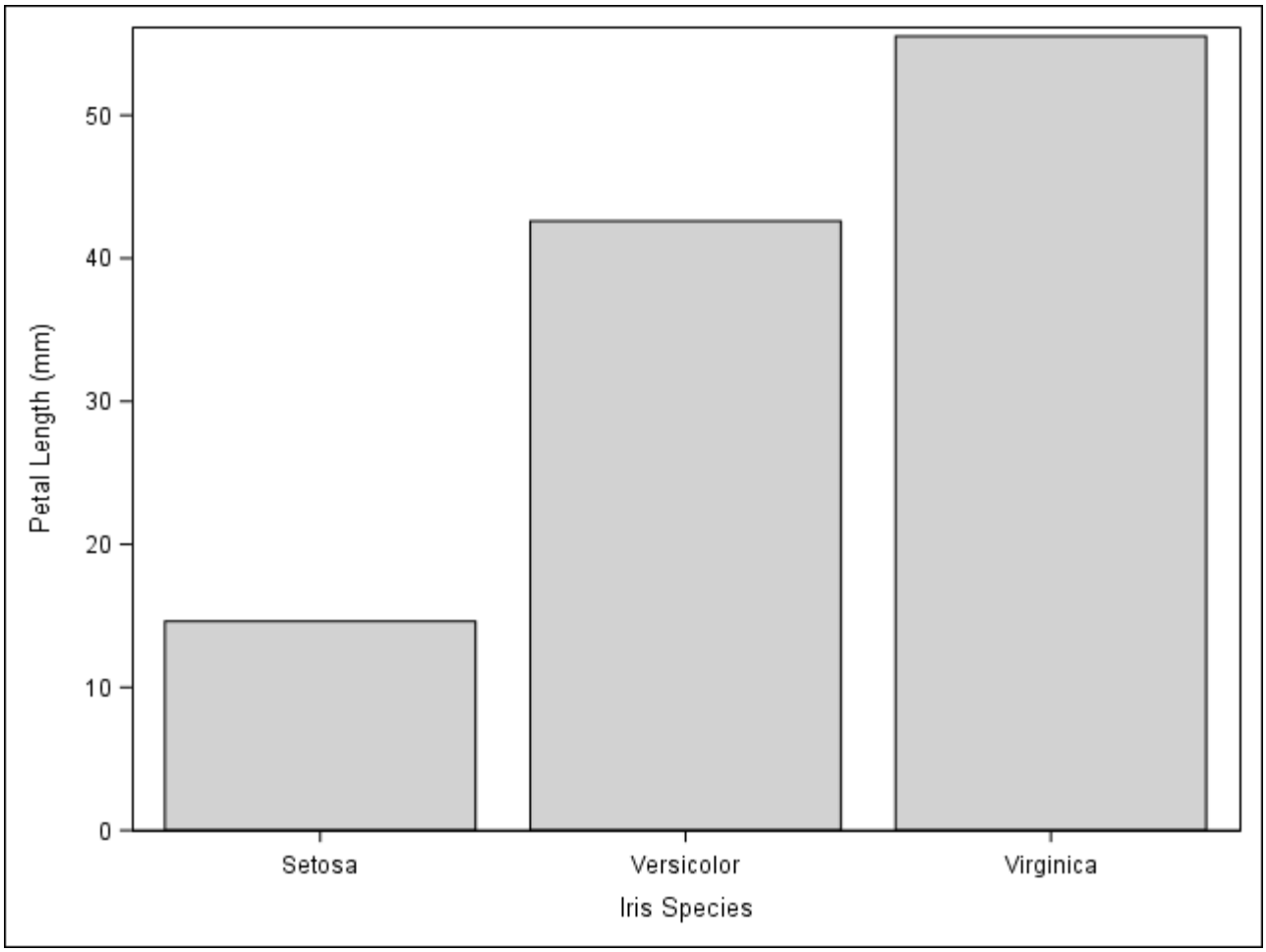
Exercise 5

1. With a partner work out the story that this data visualization is telling?
2. Thinking ahead to the statistics - what might you want to run to help fill in this story?

Bar chart with means

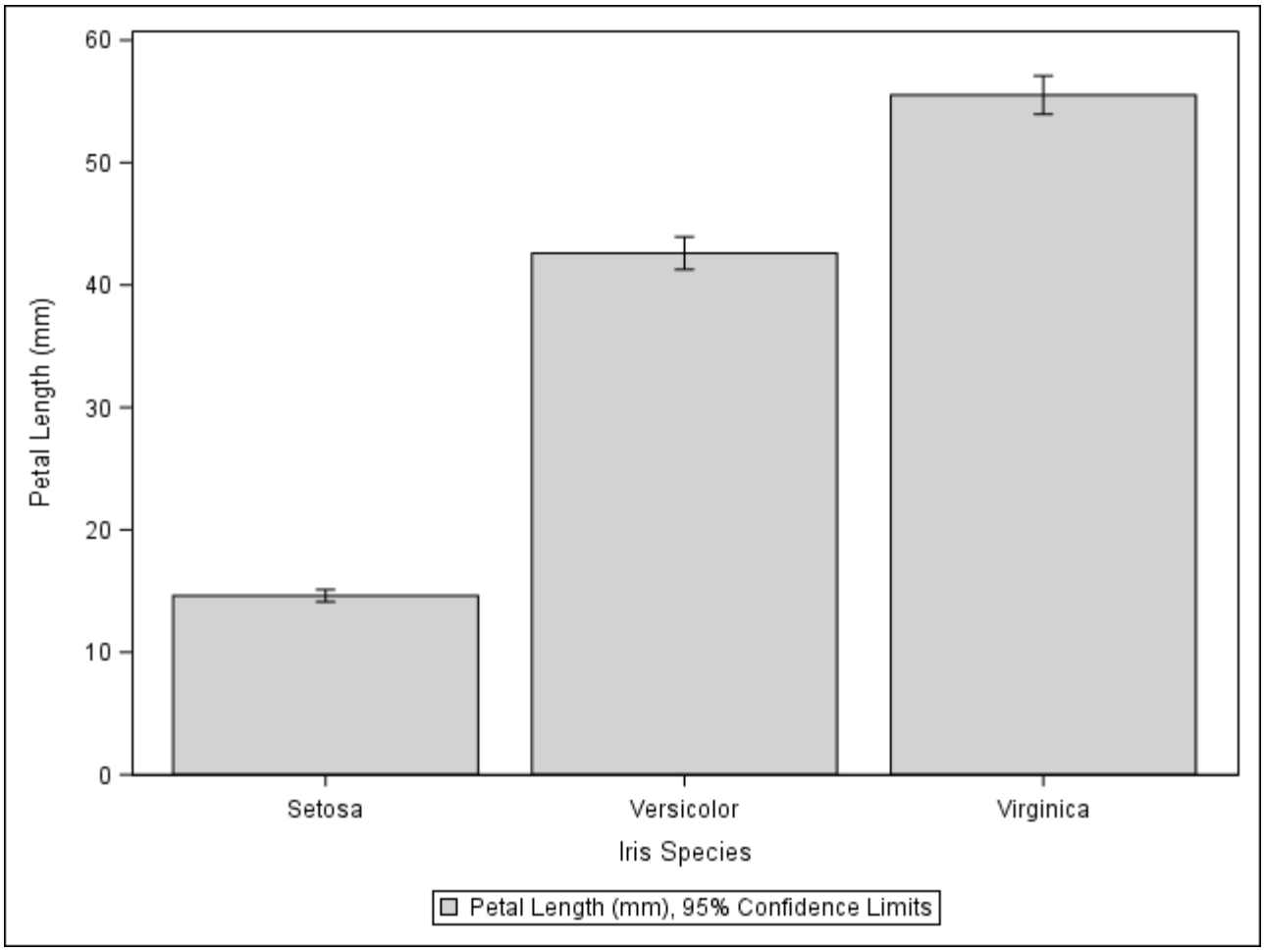
We often like to see our data represented by our groups in a bar chart. So let's use the PROC SGPLOT coding we used earlier but add a few options. Rather than representing the number of observations for each species, this time, let's have the bars represent the mean of Petal Length.

```
Proc sgplot data=sashelp.iris;
  vbar species /response=petallength stat=mean;
Run;
```



Now let's add one more piece of information. We are going to add the 95% confidence limits of each mean. To add the standard error bars, we would need to use PROC MEANS to calculate the errors and save them in a dataset which we would then use to create the plots. Let's keep it simple for this example and use the confidence limits.

```
Proc sgplot data=sashelp.iris;
  vbar species /response=petallength stat=mean limitstat=clm ;
Run;
```



Please note that we can change the appearance of these graphs and add more information such as the mean, etc… Always keep in mind, the purpose of the visualization.

More SAS graph examples

There are so many different ways to visualize your data in the SAS program. For more examples and some of the differences between SAS Studio and SAS 9.4 please visit Creating graphs in SAS workshop notes on the OACstats Blog

Descriptive Statistics

The next step on your data visualization and analysis path, are descriptive statistics. As much as we would all love to jump ahead and get to the “meat” of our research and run our models, we should all specnd time getting comfortable with your data. We’ve just spent time visualizing the data - getting a sense as to what the data looks like, but now let’s get a better sense of the numbers behind some of the visualizations.

Descriptive statistics is exactly that - statistics that will help you describe your data. Class statistics may include:

- mean, median, mode
- frequencies
- distributions or ranges
- measures of variation

We reviewed the different types of data above. If you need to review these again, please take a couple of moments to reread the different data types. Why you may ask? The type of data you are working with till depict, in most cases, what

type of statistic you will calculate.

PROC UNIVARIATE

We already used thie PROC in SAS to create our histograms. In the code we used above, we selected to ONLY see the histograms. Let's try rerunning the same code we used above, but remove the **ods select histogram** statment.

```
proc univariate data=sashelp.iris;
  var PetalLength;
  histogram PetalLength / odstitle="PROC UNIVARIATE";
run;
```

Variable: PetalLength (Petal Length (mm))

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 for Location: Mu0=0

Test	Statistic		p Value	
Student's t	**t**	26.0726	**Pr > |t|**	<.0001
Sign	**M**	75	**Pr >= |M|**	<.0001
Signed Rank	**S**	5662.5	**Pr >= |S|**	<.0001

Quantiles (Definition 5)

Level	Quantile
100% Max	69.0
99%	67.0
95%	61.0
90%	58.0
75% Q3	51.0
50% Median	43.5
25% Q1	16.0
10%	14.0
5%	13.0
1%	11.0
0% Min	10.0

Extreme Observations

Lowest		Highest	
Value	**Obs**	**Value**	**Obs**
10	3	64	118
11	18	66	112

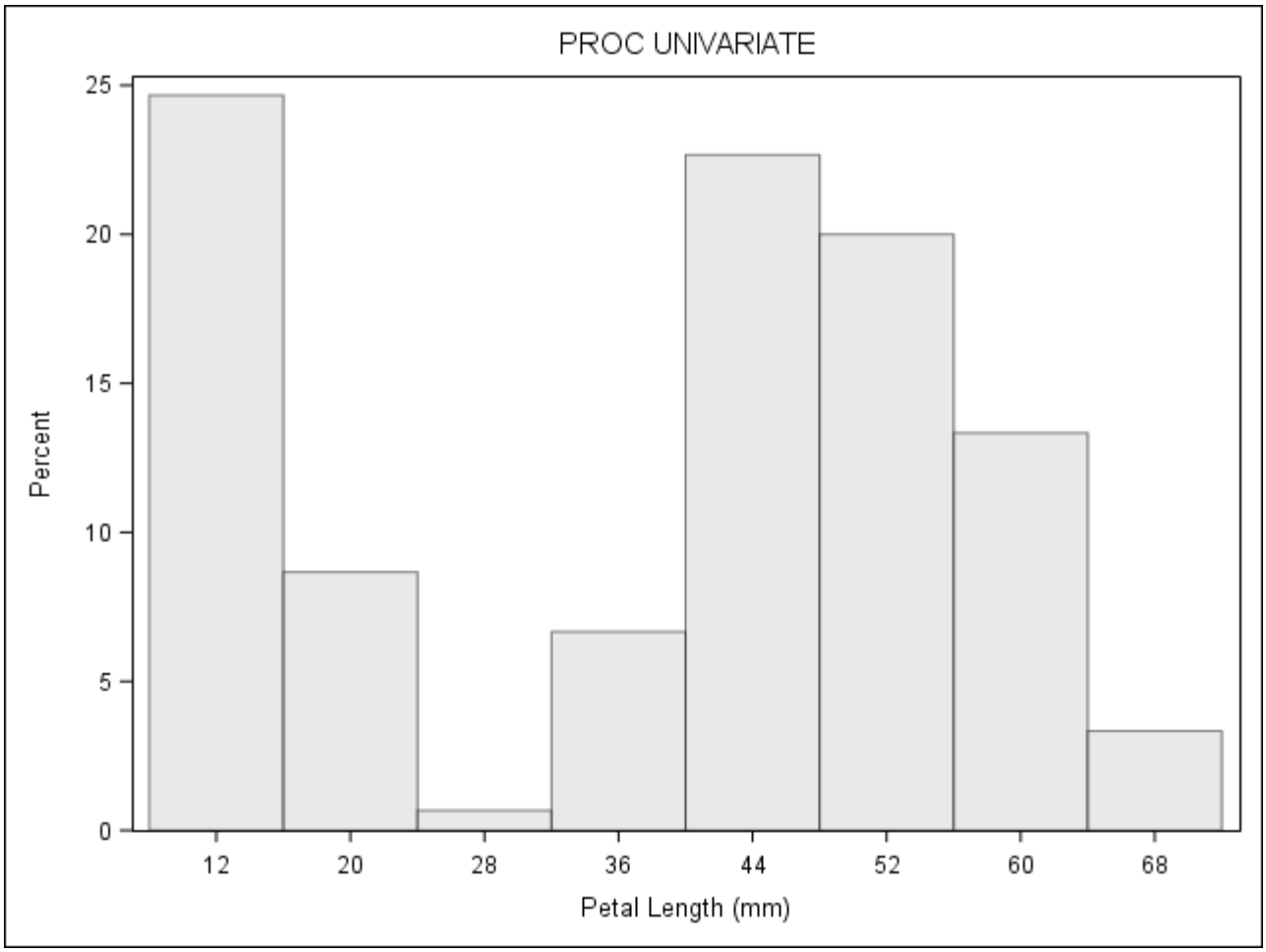
12	19	67	110
12	17	67	120
13	44	69	135

WOW!!! Lots of information. Let's work through the output together.

PROC UNIVARIATE is a great place to start getting comfortable with your data. It provides a great overview of the data you collected. But note that it is for the entire dataset and not broken out by the 3 species as we've done before.

So let's add that CLASS statement and see what happens next.

```
proc univariate data=sashelp.iris;
  class species;
  var PetalLength;
  histogram PetalLength / nrows =3 odstitle="PROC UNIVARIATE";
run;
```



WOW!!! Lots of information. Let's work through the output together.

PROC UNIVARIATE is a great place to start getting comfortable with your data. It provides a great overview of the data you collected. But note that it is for the entire dataset and not broken out by the 3 species as we've done before.

So let's add that CLASS statement and see what happens next.

```
proc univariate data=sashelp.iris;
  class species;
  var PetalLength;
  histogram PetalLength / nrows =3 odstitle="PROC UNIVARIATE";
run;
```

Variable: PetalLength (Petal Length (mm))
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 Location: Mu0=0

Test	**Statistic**		**p Value**	
Student's t	**t**	59.52818	**Pr > |t|**	<.0001
Sign	**M**	25	**Pr >= |M|**	<.0001
Signed Rank	**S**	637.5	**Pr >= |S|**	<.0001

Quantiles (Definition 5)

Level	**Quantile**
100% Max	19
99%	19

95%	17
90%	17
75% Q3	16
50% Median	15
25% Q1	14
10%	13
5%	12
1%	10
0% Min	10

Extreme Observations

Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
10	3	17	31
11	18	17	45
12	19	17	47
12	17	19	15
13	44	19	20

Variable: PetalLength (Petal Length (mm))
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 Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	64.1031	**Pr > |t|**	<.0001
Sign	**M**	25	**Pr >= |M|**	<.0001
Signed Rank	**S**	637.5	**Pr >= |S|**	<.0001

Quantiles (Definition 5)

Level	Quantile
100% Max	51.0
99%	51.0
95%	49.0
90%	48.0
75% Q3	46.0
50% Median	43.5
25% Q1	40.0
10%	35.5
5%	33.0
1%	30.0
0% Min	30.0

Extreme Observations

Lowest		Highest	
Value	**Obs**	**Value**	**Obs**
30	91	48	99
33	73	49	88
33	66	49	95
35	79	50	100
35	71	51	55

Variable: PetalLength (Petal Length (mm))
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

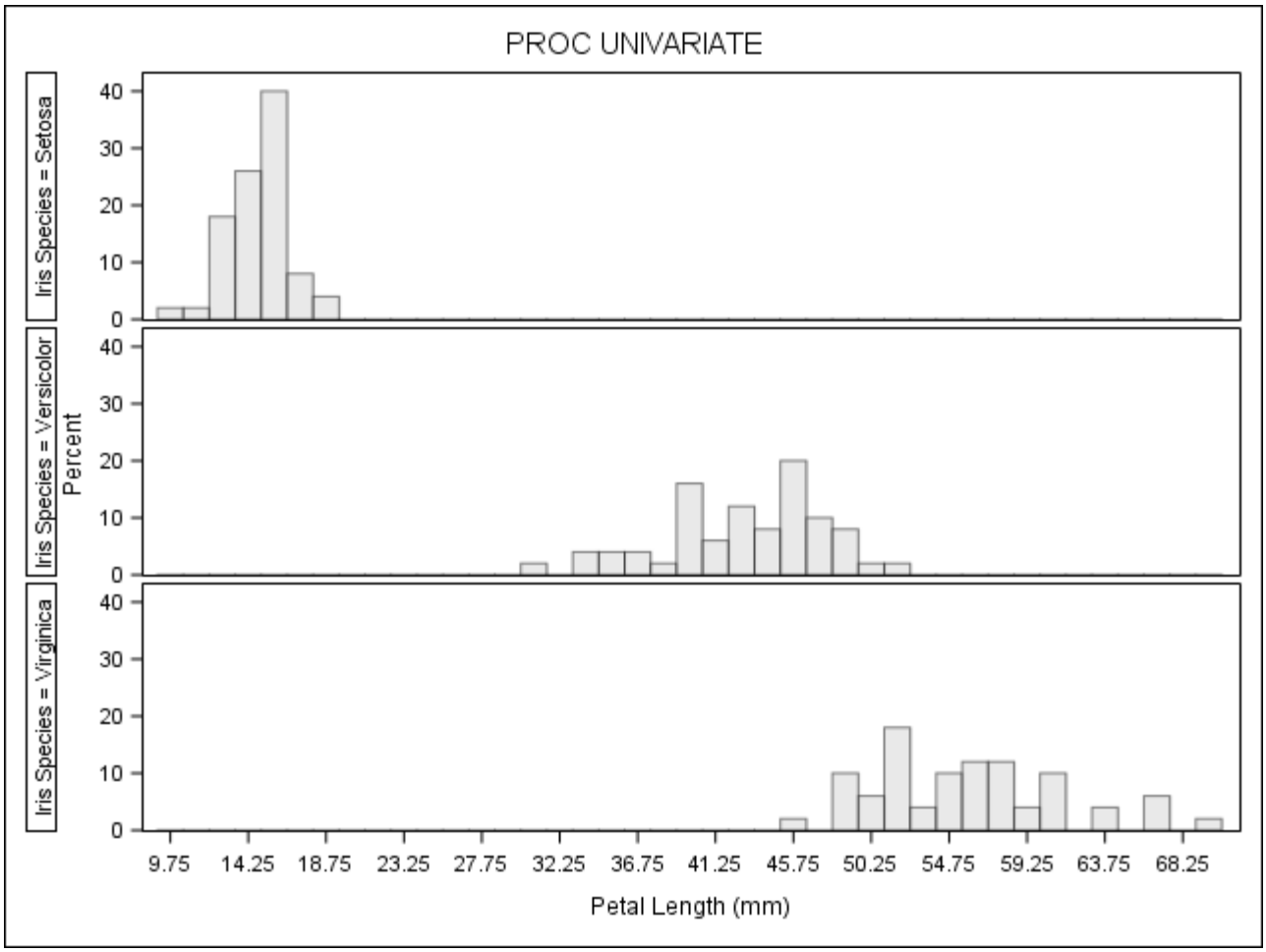
Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	71.13417	**Pr > |t|**	<.0001
Sign	**M**	25	**Pr >= |M|**	<.0001
Signed Rank	**S**	637.5	**Pr >= |S|**	<.0001

Quantiles (Definition 5)

Level	Quantile
100% Max	69.0
99%	69.0
95%	67.0
90%	63.5
75% Q3	59.0
50% Median	55.5
25% Q1	51.0
10%	49.0
5%	48.0
1%	45.0
0% Min	45.0

Extreme Observations

Lowest		Highest	
Value	**Obs**	**Value**	**Obs**
45	113	64	118
48	133	66	112
48	126	67	110
49	139	67	120
49	123	69	135



PROC UNIVARIATE provides a lot of information, but there's one more piece of information that we might want to see, how about Normality? A way to test whether are data comes from a normal distribution. If you add the option of **NORMAL PLOT** at the end of the PROC UNIVARIATE statement, we will obtain a bit more information. Let's try it out!

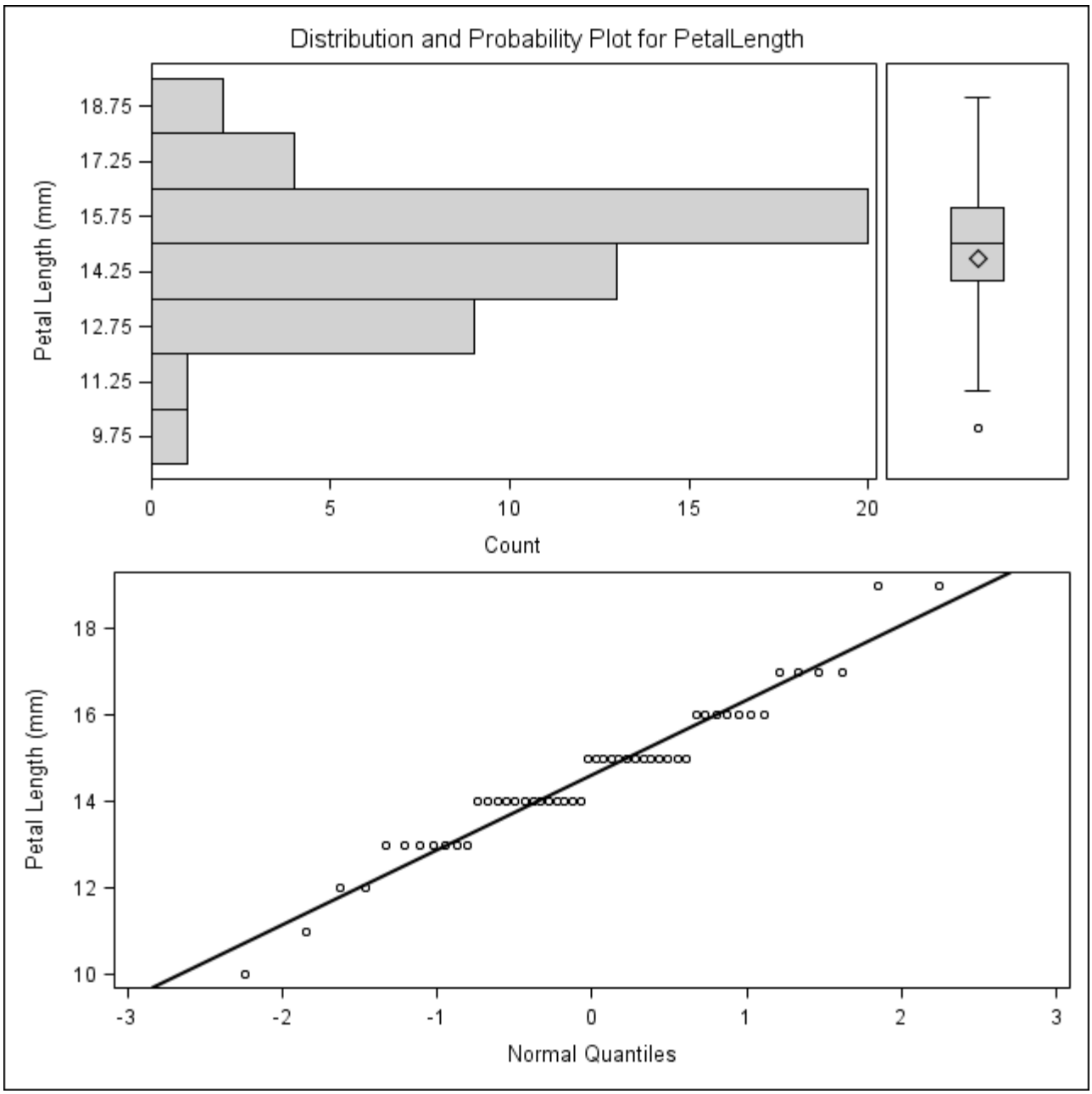
```
proc univariate data=sashelp.iris normal plot;
  class species;
  var PetalLength;
  histogram PetalLength / nrows =3 odstitle="PROC UNIVARIATE";
  ods select TestsForNormality Plots;
run;
```

Variable: PetalLength (Petal Length (mm))
Species = Setosa

Tests for Normality

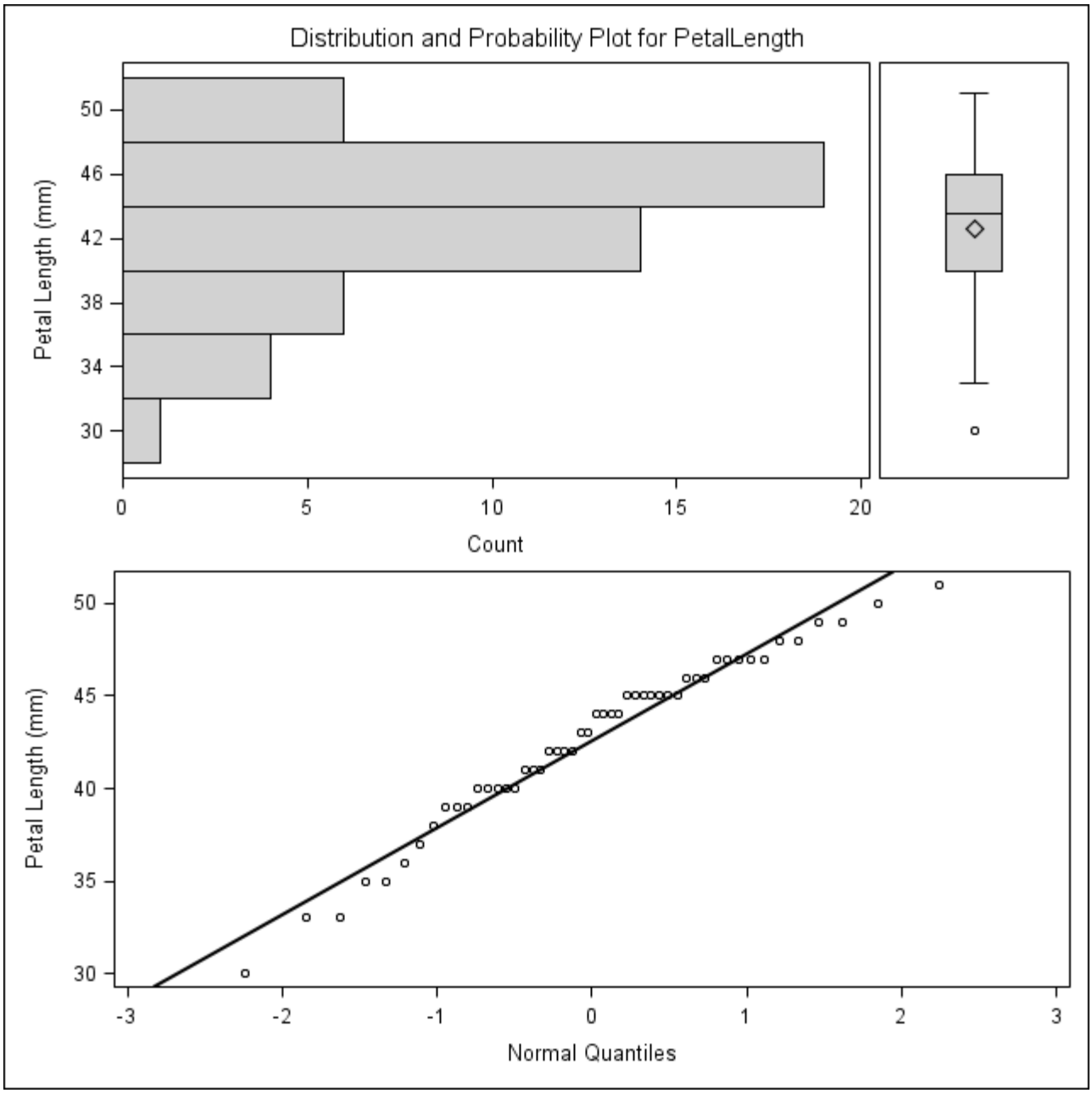
Test	Statistic		p Value	
Shapiro-Wilk	**W**	0.954977	**Pr < W**	0.0548

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



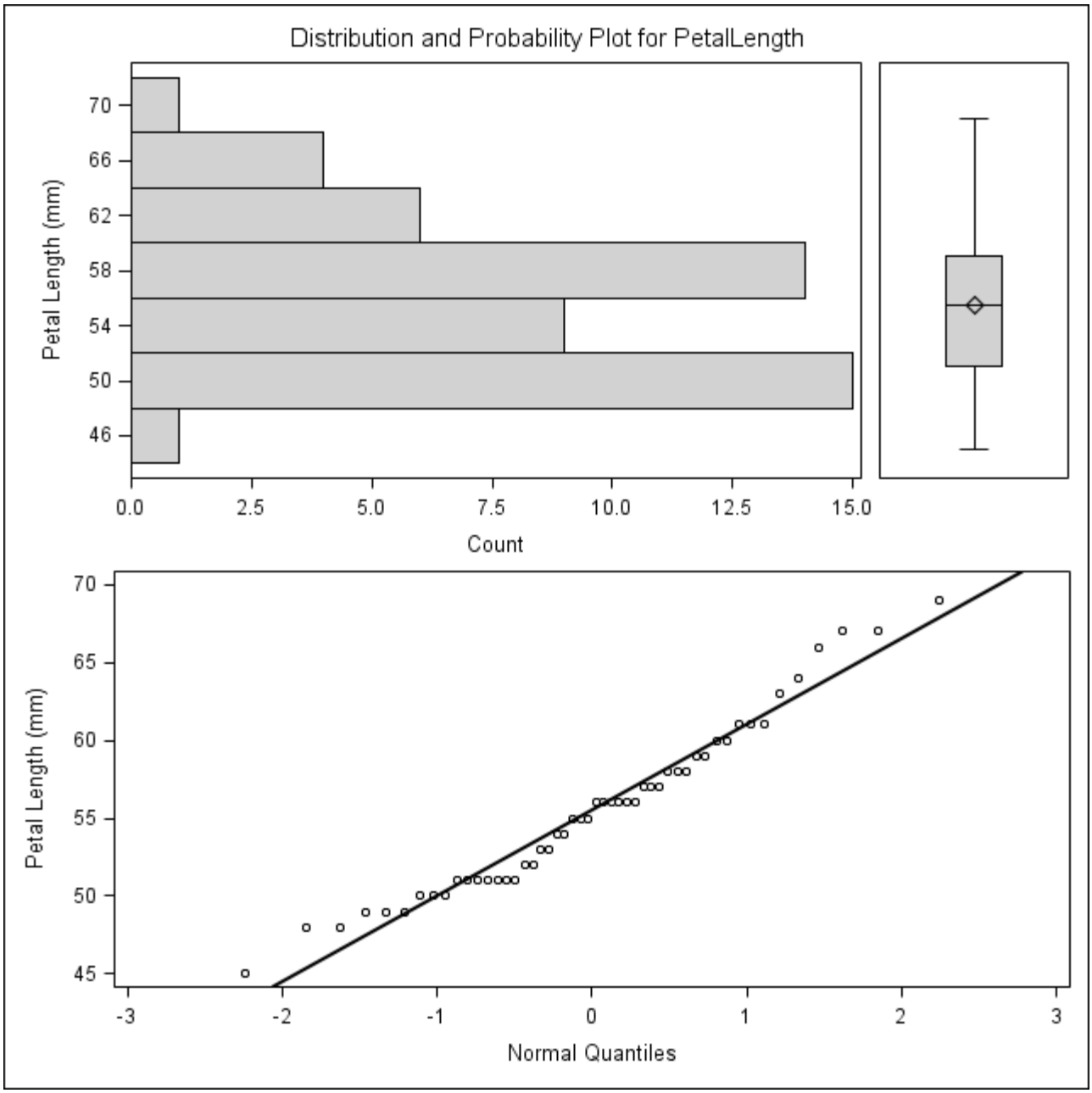
Variable: PetalLength (Petal Length (mm))
Species = Versicolor

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


Distribution and Probability Plot for PetalLength
Petal Length (mm)
50
46
42
38
34
30
0
5
10
15
20
Count
Petal Length (mm)
50
45
40
35
30
-3
-2
-1
0
1
2
3
Normal Quantiles

Variable: PetalLength (Petal Length (mm))
Species = Virginica

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



Exercise 6

1. What do these results mean? What story are they telling about our data?

PROC MEANS

As you've seen PROC UNIVARIATE can provide us with a LOT of information about our data. But what if we are only interested in the means of our variable? Do we have to see ALL that information? Nope - we can calculate the means by using a different PROC - MEANS.

Let's start with the basic PROC MEANS - we will continue to work with petallength variable.

```
proc means data=sashelp.iris;
  var PetalLength;
run;
```

Analysis Variable : PetalLength Petal Length (mm)				
N	**Mean**	**Std Dev**	**Minimum**	**Maximum**
150	37.5800000	17.6529823	10.0000000	69.0000000

The default output for PROC MEANS is:

- N
- Mean
- StdDev - standard deviation
- Minimum
- Maximum

What if we want to see the Standard Error and maybe the Sum as well. We can do that by telling SAS what statistics we would like to see, by adding them after the PROC MEANS statement. These are keywords available:

**Descriptive statistics:

- CLM NMISS
- CSS RANGE
- CV SKEWNESS|SKEW
- KURTOSIS|KURT STDDEV|STD
- LCLM STDERR
- MAX SUM
- MEAN SUMWGT
- MIN UCLM
- MODE USS
- N VAR

**Quantile statistics:

- MEDIAN|P50 Q3|P75
- P1 P90
- P5 P95
- P10 P99
- Q1|P25 QRANGE

**Hypothesis testing:

- PROBT|PRT T

```
proc means data=sashelp.iris mean max min stderr sum;
  var PetalLength;
run;
```

Analysis Variable : PetalLength Petal Length (mm)				
Mean	**Maximum**	**Minimum**	**Std Error**	**Sum**

37.5800000	69.0000000	10.0000000	1.4413600	5637.00

Notice how I was also able to rearrage the table by the order of the statistics I requested?

Exercise 7

1. Create a PROC MEANS that will show the mean, stddev, max, min, sum for Petal Length for each species.
2. Create a PROC MEANS that will show the mean, stddev, max, min, sum for Sepal Length for each species.

So we have tackled PROC UNIVARIATE and PROC MEANS, the last one I'd like to introduce you to is PROC FREQ

PROC FREQ

As the name alludes - this is used to calculate frequencies. Our data will not tell us too many stories from the frequency perspective, but we'll walk through the example nonetheless.

```
Proc freq data=sashelp.iris;
  tables species;
Run;
```

Iris Species

Species	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Setosa	50	33.33	50	33.33
Versicolor	50	33.33	100	66.67
Virginica	50	33.33	150	100.00

Exercise 8

1. With a partner discuss the results of the PROC FREQ.
2. Should we run a PROC FREQ on any other of our variables in the Fisher's Iris dataset? Why or why not?

Workshop Review

Data Visualization

- 5 General Principles Behind Data Visualization
- Graphic Design Principles
- Different Data Types
- Different examples
    - Bar Chart

    - Histograms
    - Boxplots
    - Scatterplot
    - many more….

Descriptive Statistics

- PROC UNIVARIATE
- PROC MEANS
- PROC FREQ