STAT 448 HW #1

Tianqi Wu

2018/ 9/4

Problem 1.(a)

Here is the code that produces the scatter plot matrix:

```
proc sgscatter data=fishy;
  title "Scatterplot Matrix for fishy Data";
  matrix weight length1 length3 height;
run;
```

From the plot, we can see that there are strong positive relationships among the numeric variables. There is nonlinear relationship between **weight** and others while there are linear relationships among others (**length1, length2 and height**).

Problem 1.(b)

Here is the code that produces the correlation analyses:

```
proc corr data = fishy;
  title "Linear Correlation Analyses for fishy Data";
run;

proc corr data = fishy spearman;
  title "Nonlinear Correlation Analyses for fishy Data";
run;
```

For both linear and nonlinear correlation analyses, the correlation values among **weight**, **length1** and **length3** are quite large (around 0.95) with p-value < 0.0001. It indicates that there are strong positive correlation among the three variables. The correlation values between height and others are relatively smaller (around 0.7) with p-value < 0.0001 which indicates weaker positive correlations.

Linear Correlation Analyses for fishy Data

The CORR Procedure

4 Variables:	weight length1 length3 height

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
weight	60	530.76500	306.48388	31846	5.90000	1100
length1	60	28.76333	6.95494	1726	7.50000	41.10000
length3	60	34.78833	8.43956	2087	8.80000	46.60000
height	60	11.79366	4.35158	707.61970	2.11200	18.95700

Pearson Correlation Coefficients, N = 60
Prob > |r| under H0: Rho=0

	weight	length1	length3	height
weight	1.00000	0.94733 <.0001	0.95547 <.0001	0.80092 <.0001
length1	0.94733 <.0001	1.00000	0.97523 <.0001	0.72718 <.0001
length3	0.95547 <.0001	0.97523 <.0001	1.00000	0.85738 <.0001
height	0.80092 <.0001	0.72718 <.0001	0.85738 <.0001	1.00000

Nonlinear Correlation Analyses for fishy Data

The CORR Procedure

4 Variables:	weight length1 length3 height

Simple Statistics

Variable	N	Mean	Std Dev	Median	Minimum	Maximum
weight	60	530.76500	306.48388	535.00000	5.90000	1100
length1	60	28.76333	6.95494	29.40000	7.50000	41.10000
length3	60	34.78833	8.43956	36.45000	8.80000	46.60000
height	60	11.79366	4.35158	12.46200	2.11200	18.95700

Spearman Correlation Coefficients, N = 60
Prob > |r| under H0: Rho=0

	weight	length1	length3	height
weight	1.00000	0.96886 <.0001	0.98376 <.0001	0.77370 <.0001
length1	0.96886 <.0001	1.00000	0.97197 <.0001	0.67223 <.0001
length3	0.98376 <.0001	0.97197 <.0001	1.00000	0.79799 <.0001
height	0.77370 <.0001	0.67223 <.0001	0.79799 <.0001	1.00000

Problem 1.(c)

Here is the code that produces the correlation analyses when grouped:

```
proc corr data = fishy nosimple;
  by species;
  title "Linear Correlation Analyses by group for fishy
Data";
run;

proc corr data = fishy spearman nosimple;
  by species;
  title "Nonlinear Correlation Analyses by group for fishy
Data";
run;
```

For both linear and nonlinear correlation analyses for different species, the correlation values among all the variables are quite large (around 0.97) with p-value < 0.0001. It indicates that there are stronger positive correlation among all the variables after grouping by species. Also, the nonlinear one shows relatively higher correlations values overall.

Linear Correlation Analyses by group for fishy Data

The CORR Procedure

species=Bream

4 Variables:	weight length1 length3 height

Pearson Correlation Coefficients, N = 30
Prob > |r| under H0: Rho=0

	weight	length1	length3	height
weight	1.00000	0.96107 <.0001	0.97049 <.0001	0.97533 <.0001
length1	0.96107 <.0001	1.00000	0.99631 <.0001	0.94616 <.0001
length3	0.97049 <.0001	0.99631 <.0001	1.00000	0.95891 <.0001
height	0.97533 <.0001	0.94616 <.0001	0.95891 <.0001	1.00000

Nonlinear Correlation Analyses by group for fishy Data

The CORR Procedure

species=Bream

4 Variables:	weight length1 length3 height

Spearman Correlation Coefficients, N = 30
Prob > |r| under H0: Rho=0

	weight	length1	length3	height
weight	1.00000	0.97248 <.0001	0.97124 <.0001	0.96815 <.0001
length1	0.97248 <.0001	1.00000	0.99633 <.0001	0.96951 <.0001
length3	0.97124 <.0001	0.99633 <.0001	1.00000	0.96661 <.0001
height	0.96815 <.0001	0.96951 <.0001	0.96661 <.0001	1.00000

Linear Correlation Analyses by group for fishy Data

The CORR Procedure

species=Perch

4 Variables:	weight length1 length3 height

Pearson Correlation Coefficients, N = 30
Prob > |r| under H0: Rho=0

	weight	length1	length3	height
weight	1.00000	0.95726 <.0001	0.95793 <.0001	0.96261 <.0001
length1	0.95726 <.0001	1.00000	0.99939 <.0001	0.99100 <.0001
length3	0.95793 <.0001	0.99939 <.0001	1.00000	0.99153 <.0001
height	0.96261 <.0001	0.99100 <.0001	0.99153 <.0001	1.00000

Nonlinear Correlation Analyses by group for fishy Data

The CORR Procedure

species=Perch

4 Variables:	weight length1 length3 height

Spearman Correlation Coefficients, N = 30
Prob > |r| under H0: Rho=0

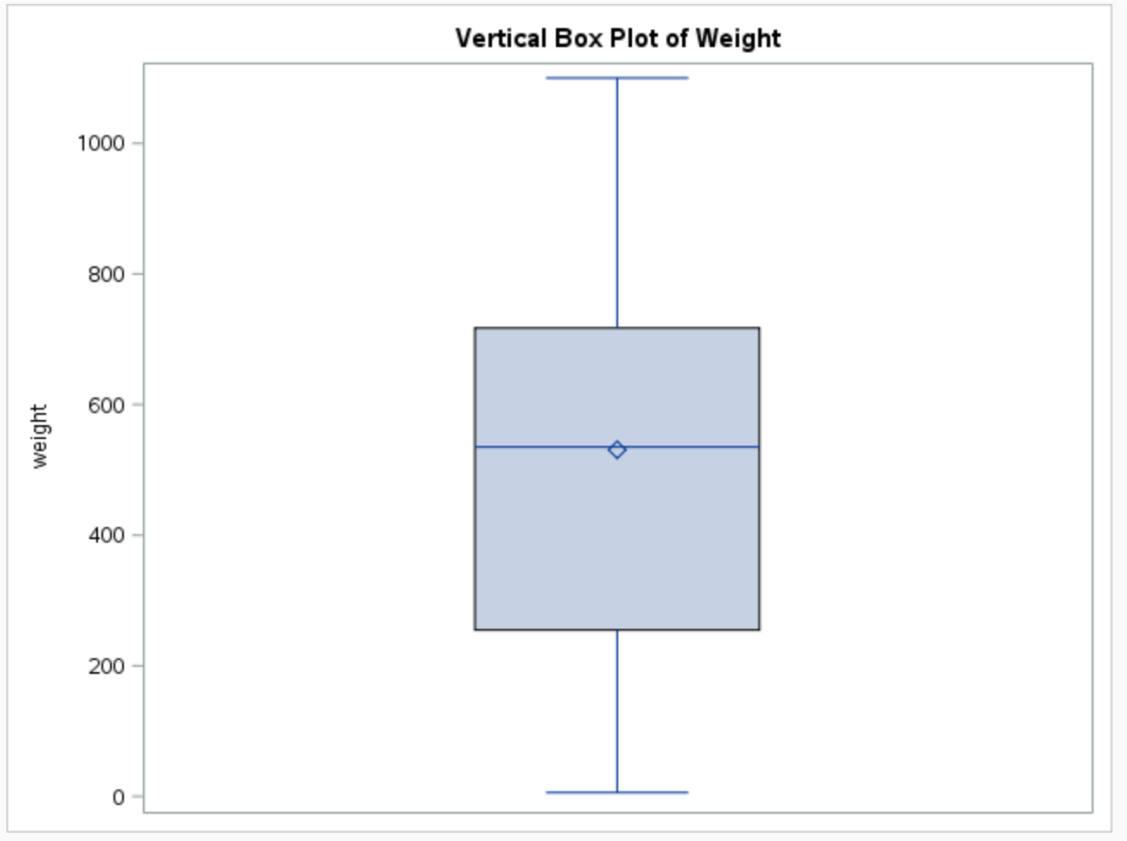
	weight	length1	length3	height
weight	1.00000	0.98886 <.0001	0.98841 <.0001	0.98263 <.0001
length1	0.98886 <.0001	1.00000	0.99989 <.0001	0.99120 <.0001
length3	0.98841 <.0001	0.99989 <.0001	1.00000	0.99154 <.0001
height	0.98263 <.0001	0.99120 <.0001	0.99154 <.0001	1.00000

Problem 1.(d)

Here is the code that produces the boxplot of the **weight** variable:

```
proc sgplot data=fishy;
 title "Vertical Box Plot of Weight";
 vbox weight;
run;
```

From the boxplot, the minimum is 0 and the maximum is around 1100. The mean and median is close (around 500). The upper quantile is around 700 and lower quantile is around 250. Hence, the IQR is around 450. The plot may be symmetric since there is no obvious skewness looking above and below the median. The box does not spread much and there is no outlier considering the range (Q1-1.5*IQR, Q3+1.5*IQR) = (-425 ~ 1375).

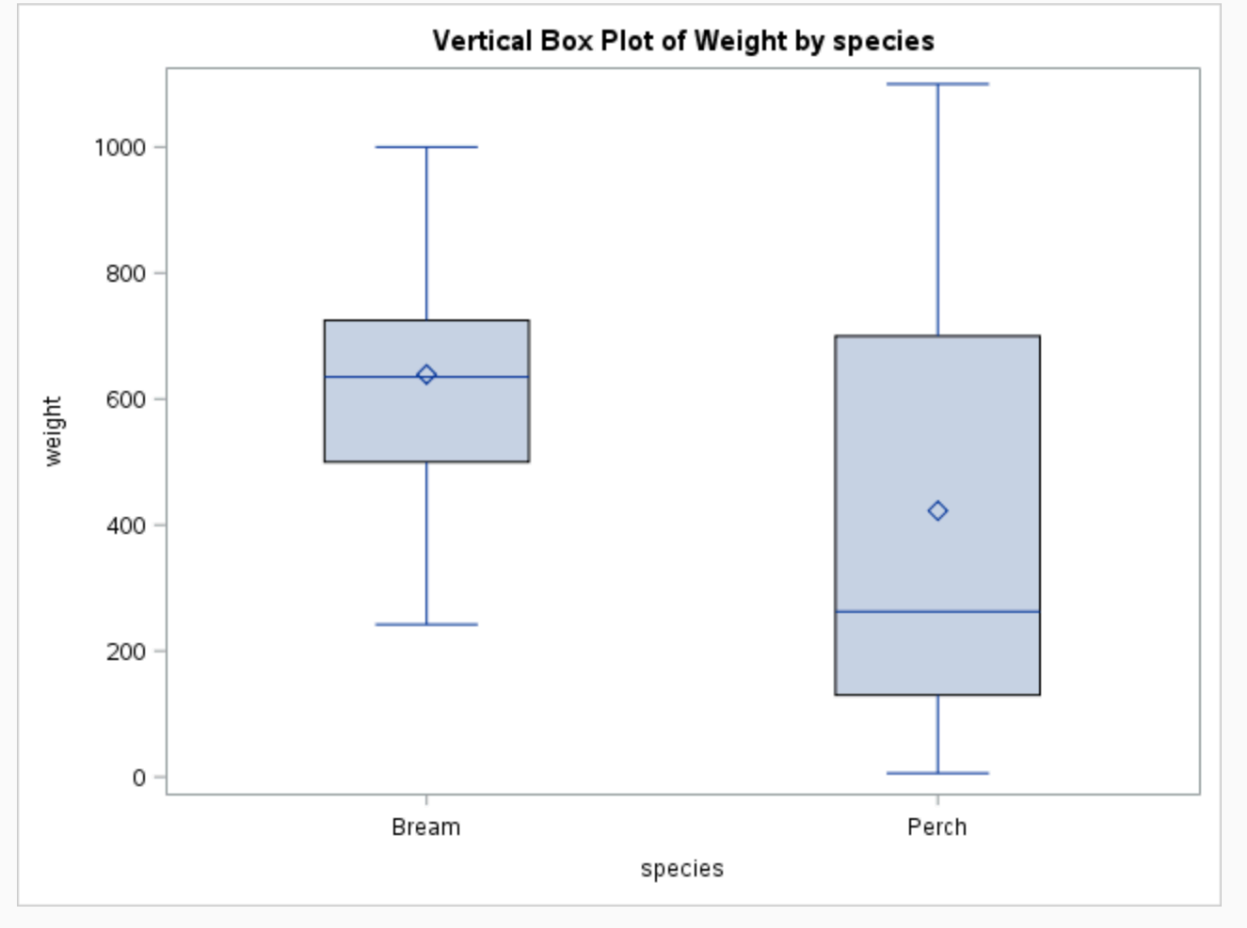


Problem 1.(e)

Here is the code that produces the boxplots of the **weight** variable by species:

```
proc sgplot data=fishy;
  title "Vertical Box Plot of Weight by species";
  vbox weight / category=species;
run;
```

From the boxplots, **Bream** has similar mean and median (around 650) but **Perch** has distinct mean(400) and median(220). **Bream** does not spread as much as **Perch** and **Perch** has smaller minimum and larger maximum. Since **Bream** has a much smaller IQR, outliers may exist from the observation. **Bream** can be seen as symmetric but **Perch** skewed to the small values of weight.



Problem 1.(f)

Here is the code that produces the basic statistics, histogram and QQ plots of weight and length1:

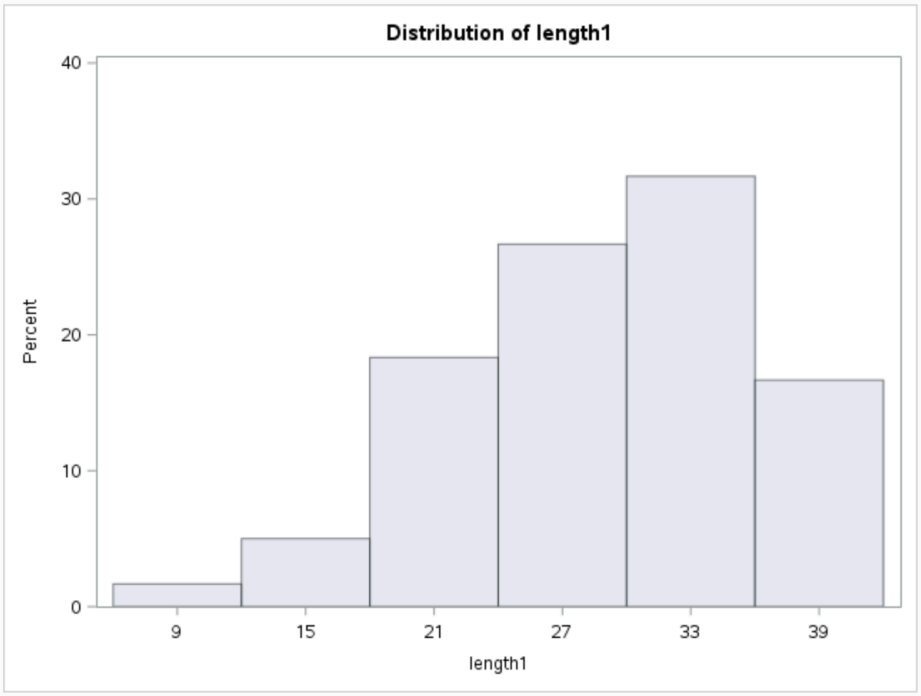
```
proc univariate data=fishy;
 var length1 weight;
 histogram;
 qqplot;
run;
```

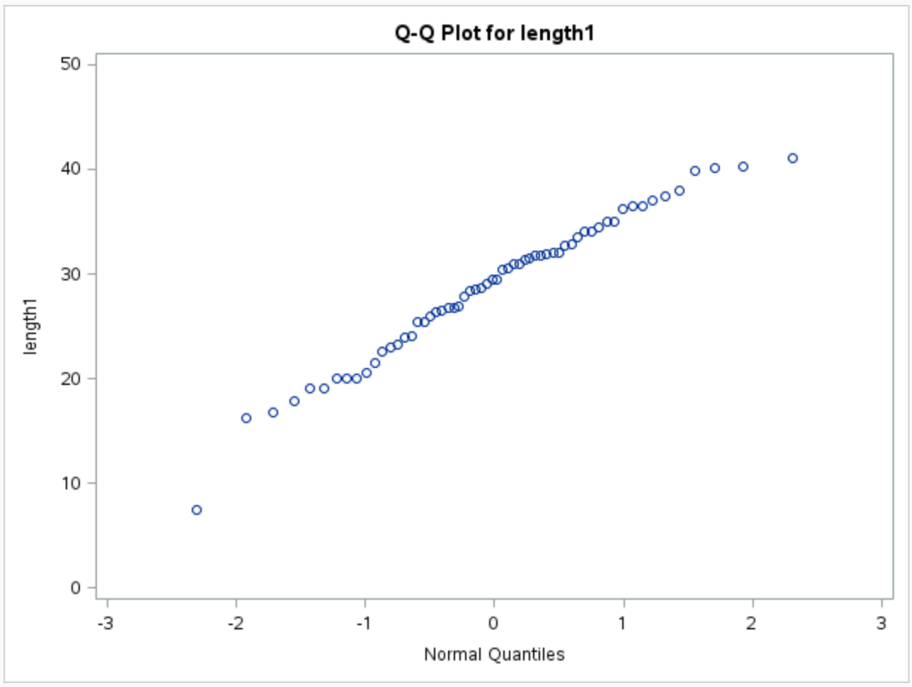
From the results, **length1** shows nearly straight line from the QQ plot which indicates normality while **weight** shows nonlinear property which does not indicate normality. **Length1** has skewness of -0.50 but **weight** has skewness of 0.05. **Also**, weight has relatively large variance. From the histograms, the distribution of **lenght1** is almost bell shaped but **weight** is more equally distributed.

The UNIVARIATE Procedure
Variable: length1

Moments			
N	60	Sum Weights	60
Mean	28.7633333	Sum Observations	1725.8
Std Deviation	6.9549389	Variance	48.3711751
Skewness	-0.495219	Kurtosis	0.20263693
Uncorrected SS	52493.66	Corrected SS	2853.89933
Coeff Variation	24.179878	Std Error Mean	0.89787875

Basic Statistical Measures			
Location		Variability	
Mean	28.76333	Std Deviation	6.95494
Median	29.40000	Variance	48.37118
Mode	20.00000	Range	33.60000
		Interquartile Range	9.80000

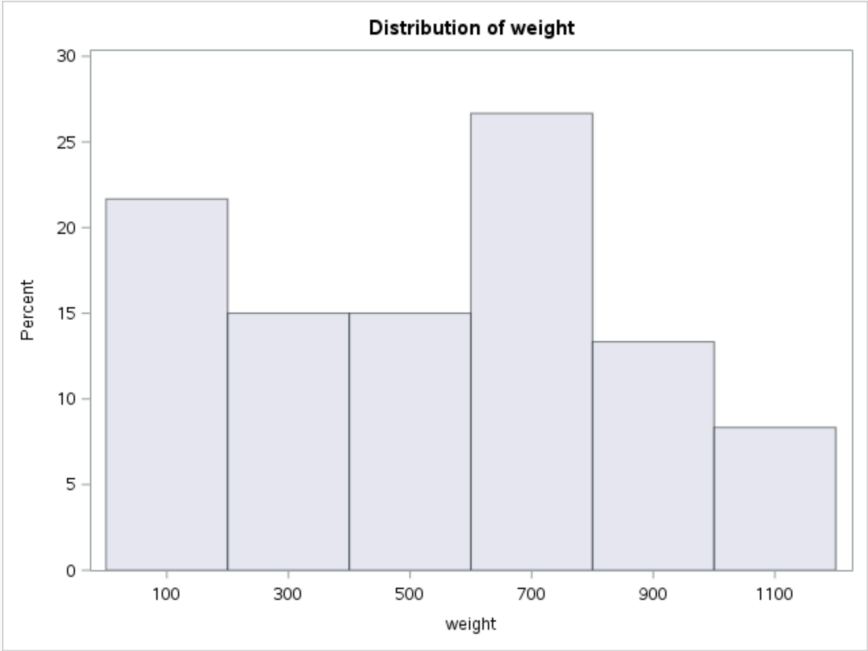

Distribution of length1
Percent
40
30
20
10
0
9
15
21
27
33
39
length1

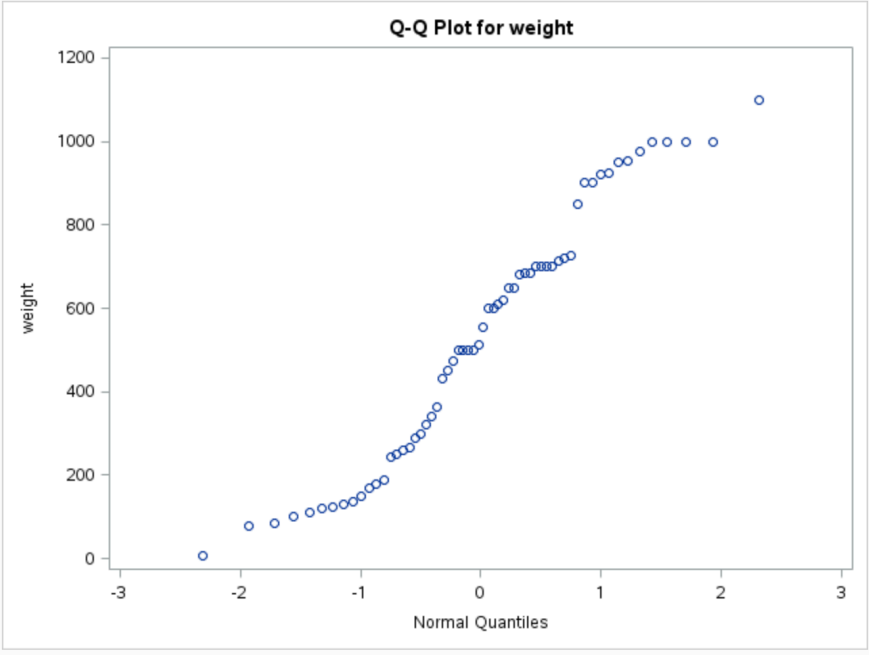

Q-Q Plot for length1
length1
50
40
30
20
10
0
-3
-2
-1
0
1
2
3
Normal Quantiles

The UNIVARIATE Procedure
Variable: weight

Moments			
N	60	Sum Weights	60
Mean	530.765	Sum Observations	31845.9
Std Deviation	306.483878	Variance	93932.3677
Skewness	0.04996808	Kurtosis	-1.1702522
Uncorrected SS	22444698.8	Corrected SS	5542009.7
Coeff Variation	57.7437997	Std Error Mean	39.5668986

Basic Statistical Measures			
Location		Variability	
Mean	530.7650	Std Deviation	306.48388
Median	535.0000	Variance	93932
Mode	500.0000	Range	1094
		Interquartile Range	462.00000





Problem 1.(g)

Here is the code that produces the basic statistics, histogram and QQ plots of height and length3 by species:

```
proc univariate data=fishy;
 var length3 height;
 by species;
 histogram;
 qqplot;
run;
```

From the results, length3, height of Bream and length3 of Perch show nearly straight lines from the QQ plots which indicates normality while height of Perch shows nonlinear property which does not indicate normality.Variables of Bream has greater mean and smaller variance. All of the variables have small skewness with no outliers.

The UNIVARIATE Procedure
Variable: length3

species=Bream

Moments			
N	30	Sum Weights	30
Mean	38.5733333	Sum Observations	1157.2
Std Deviation	4.4025019	Variance	19.382023
Skewness	-0.0759458	Kurtosis	-0.584556
Uncorrected SS	45199.14	Corrected SS	562.078667
Coeff Variation	11.4133302	Std Error Mean	0.8037832

Basic Statistical Measures			
Location		Variability	
Mean	38.57333	Std Deviation	4.40250
Median	38.65000	Variance	19.38202
Mode	36.20000	Range	16.50000
		Interquartile Range	5.30000

The UNIVARIATE Procedure
Variable: height

species=Bream

Moments			
N	30	Sum Weights	30
Mean	15.36119	Sum Observations	460.8357
Std Deviation	2.0415227	Variance	4.16781495
Skewness	0.05931866	Kurtosis	-0.7338656
Uncorrected SS	7199.85138	Corrected SS	120.866634
Coeff Variation	13.2901338	Std Error Mean	0.37272935

Basic Statistical Measures			
Location		Variability	
Mean	15.36119	Std Deviation	2.04152
Median	15.45330	Variance	4.16781
Mode	.	Range	7.43700
		Interquartile Range	2.33750

The UNIVARIATE Procedure
Variable: length3

species=Perch

Moments			
N	30	Sum Weights	30
Mean	31.0033333	Sum Observations	930.1
Std Deviation	9.79213208	Variance	95.8858506
Skewness	0.00115348	Kurtosis	-0.7335947
Uncorrected SS	31616.89	Corrected SS	2780.68967
Coeff Variation	31.5841267	Std Error Mean	1.78779054

Basic Statistical Measures			
Location		Variability	
Mean	31.00333	Std Deviation	9.79213
Median	29.15000	Variance	95.88585
Mode	23.50000	Range	37.80000
		Interquartile Range	15.90000

The UNIVARIATE Procedure
Variable: height

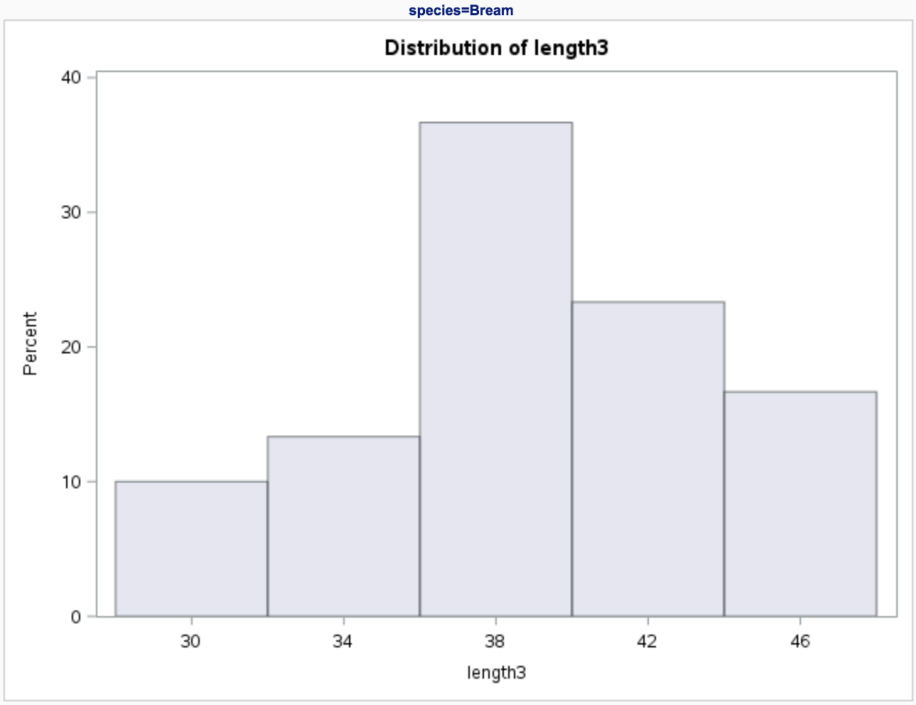
species=Perch

Moments			
N	30	Sum Weights	30
Mean	8.22613333	Sum Observations	246.784
Std Deviation	2.83292407	Variance	8.02545878
Skewness	0.07718395	Kurtosis	-1.0202433
Uncorrected SS	2262.81639	Corrected SS	232.738305
Coeff Variation	34.4381006	Std Error Mean	0.51721881

Basic Statistical Measures			
Location		Variability	
Mean	8.226133	Std Deviation	2.83292
Median	7.376200	Variance	8.02546
Mode	5.692500	Range	10.49200
		Interquartile Range	4.77100

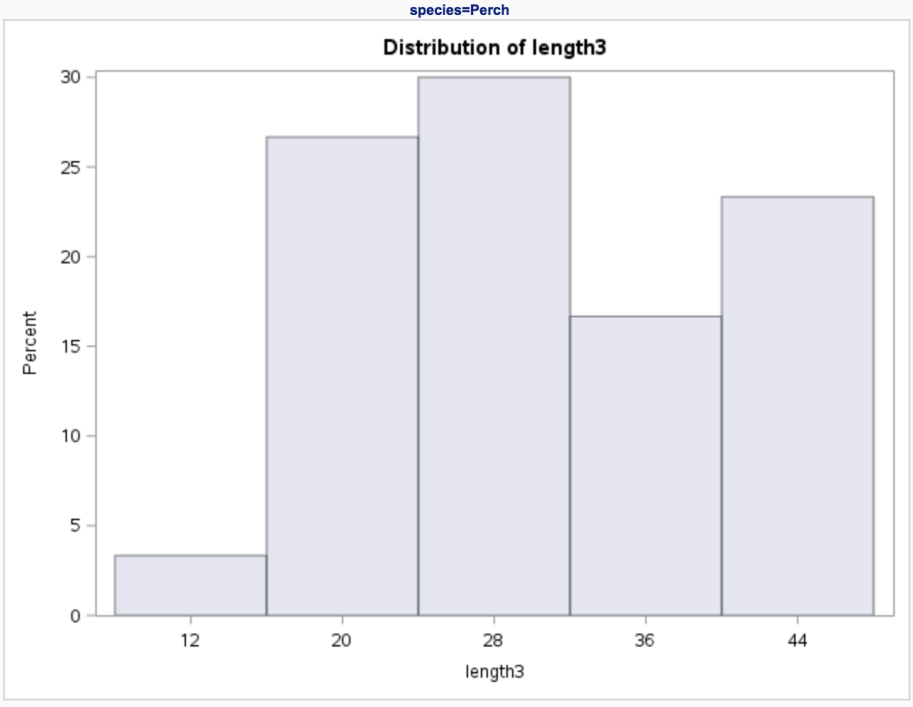
The UNIVARIATE Procedure

species=Bream



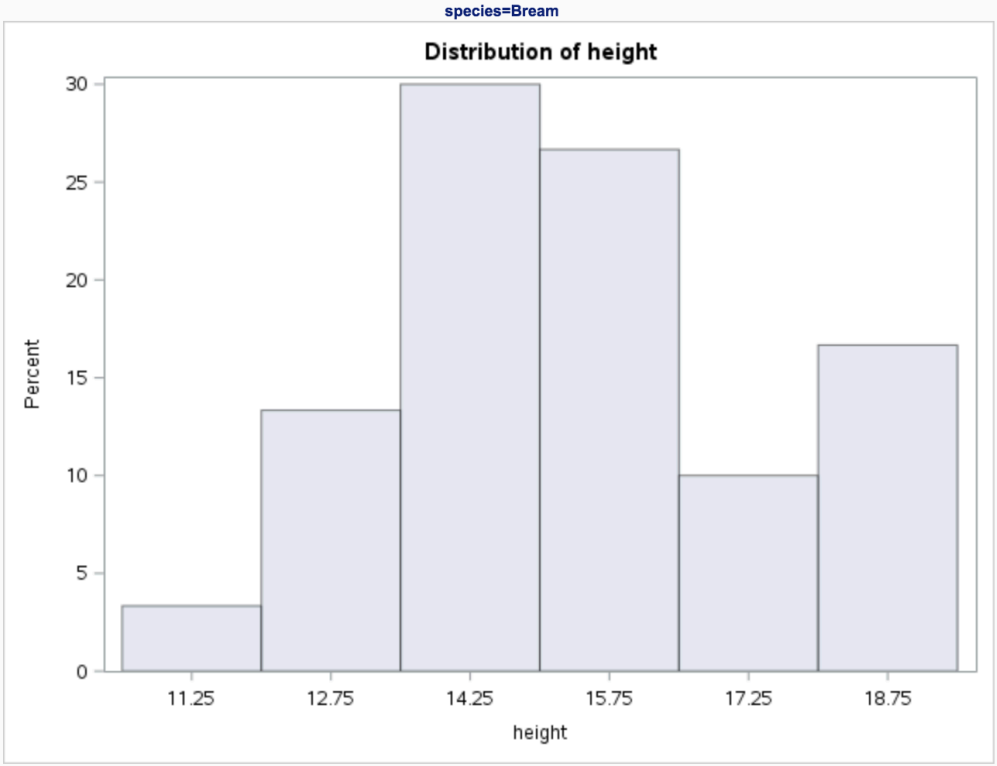
The UNIVARIATE Procedure

species=Perch



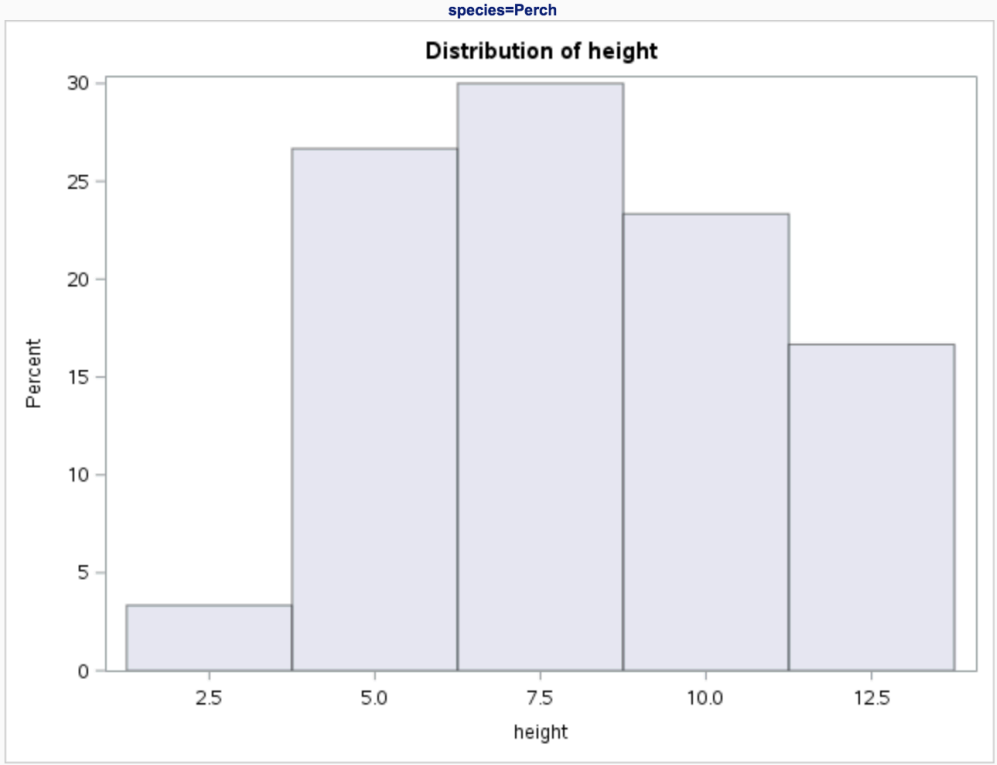
The UNIVARIATE Procedure

species=Bream



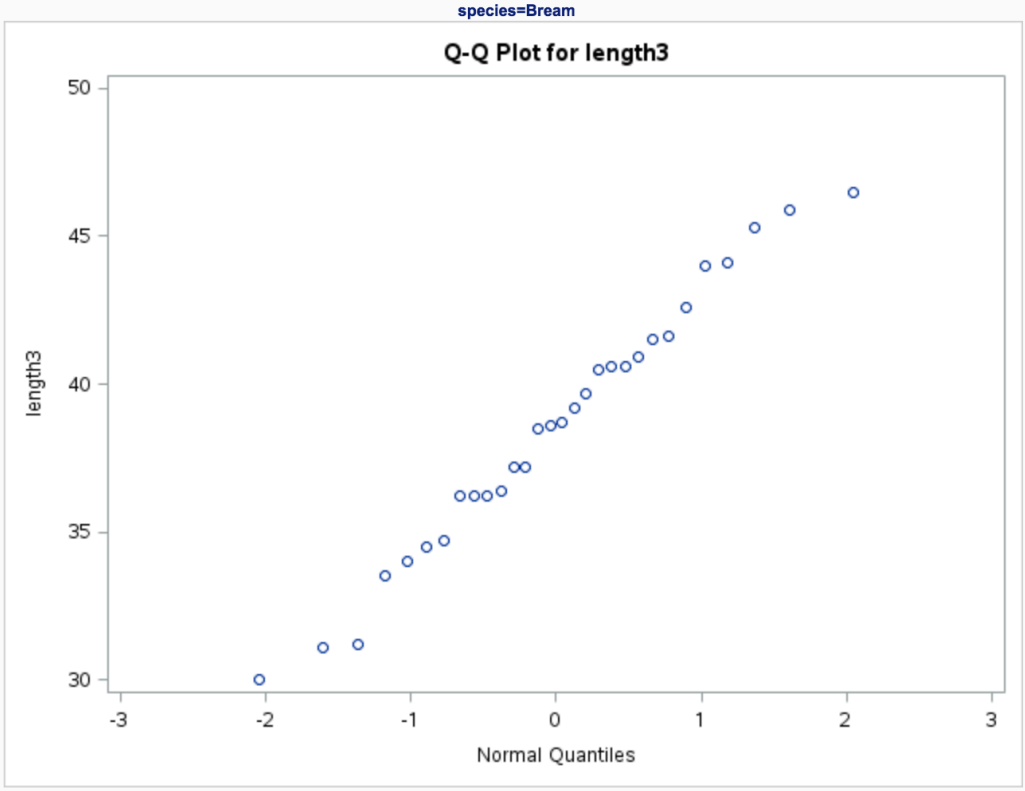
The UNIVARIATE Procedure

species=Perch



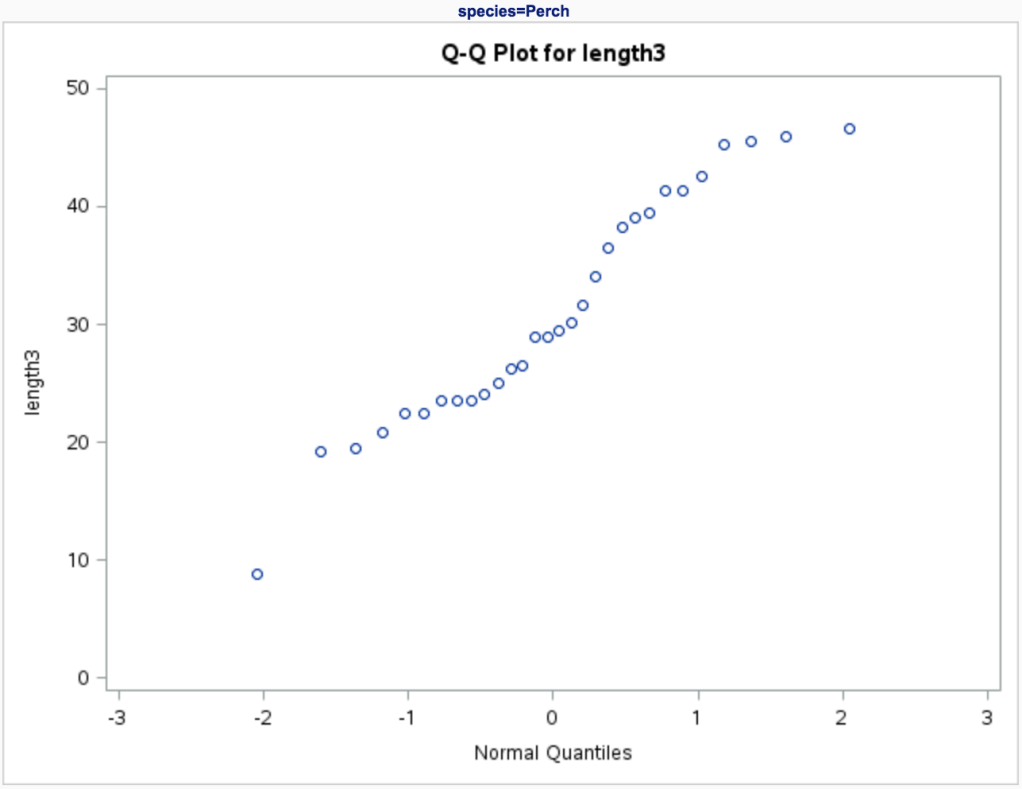
The UNIVARIATE Procedure

species=Bream



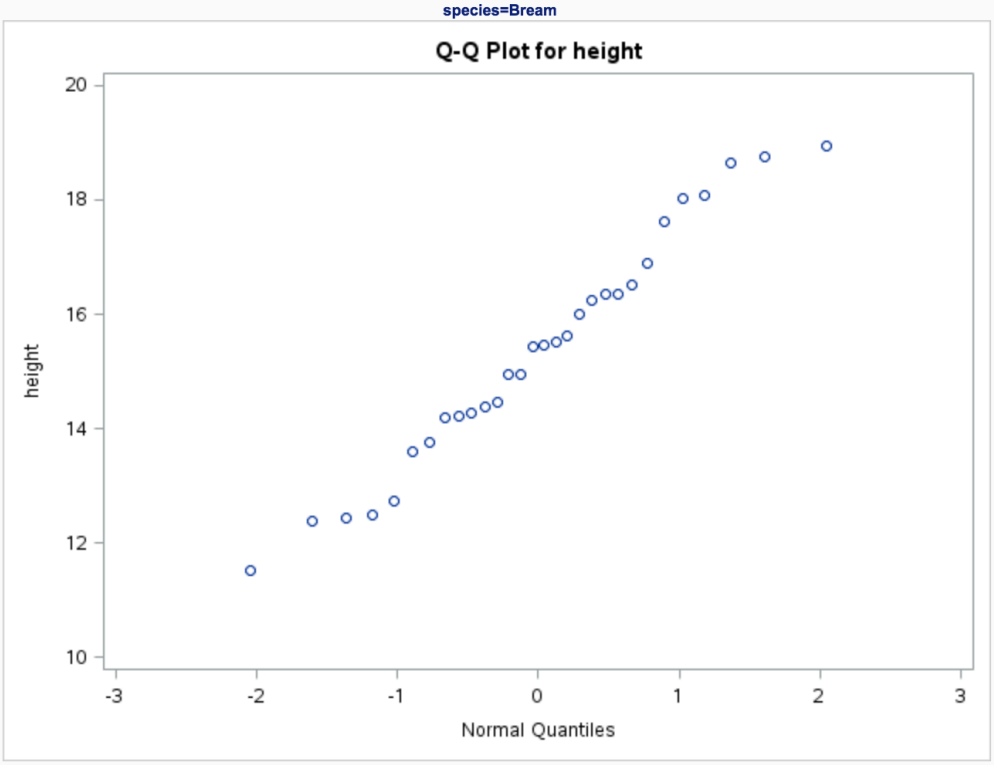
The UNIVARIATE Procedure

species=Perch



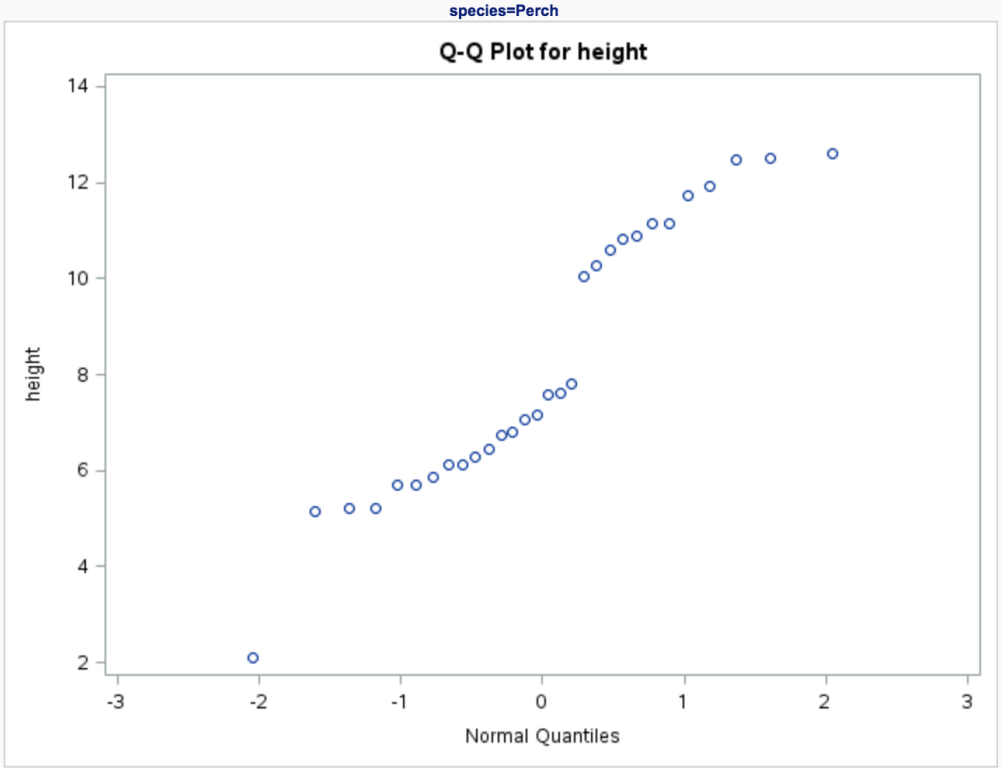
The UNIVARIATE Procedure

species=Bream



The UNIVARIATE Procedure

species=Perch



Problem 1.(h)

Here is the code that produces hypothesis tests and confidence intervals of **length1:**

```
proc univariate data=fishy mu0=30 normaltest cibasic
alpha=0.05;
 var length1;
 ods select TestsForLocation TestsForNormality
BasicIntervals;
run;
proc ttest data = fishy h0=30;
  var length1;
  ods select ConfLimits TTests;
run;
```

From the goodness of fit tests, p-values > 0.05 indicates the normality of the variable length1. Altogether with the result from the previous QQ plot and histogram, t-test is appropriate for determining the true center. From the result of t-test, p-value of 0.1736 >0.05 indicates that it is reasonable to say that the true central length1 among all fishes is 30 cm.

The UNIVARIATE Procedure
Variable: length1

Basic Confidence Limits Assuming Normality			
Parameter	Estimate	95% Confidence Limits	
Mean	28.76333	26.96668	30.55998
Std Deviation	6.95494	5.89524	8.48267
Variance	48.37118	34.75389	71.95576

Tests for Location: Mu0=30				
Test	Statistic		p Value	
Student's t	t	-1.37732	Pr > |t|	0.1736
Sign	M	-1	Pr >= |M|	0.8974
Signed Rank	S	-136.5	Pr >= |S|	0.3190

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.976422	Pr < W	0.2964
Kolmogorov-Smirnov	D	0.076355	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.050996	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.324666	Pr > A-Sq	>0.2500

The TTEST Procedure

Variable: length1

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
28.7633	26.9667	30.5600	6.9549	5.8952	8.4827

DF	t Value	Pr > |t|
59	-1.38	0.1736

Problem 1.(i)

Here is the code that produces hypothesis tests and confidence intervals of **length3** by species**:**

```
proc univariate data= fishy normaltest;
 var length3;
 by species;
 ods select TestsForLocation TestsForNormality
BasicIntervals;
run;
proc ttest data=fishy;
class species;
var length3;
ods select TTests Equality ConfLimits;
run;
```

First, from the goodness of fit tests, p-value >0.05 for variable length3 from both species indicate that two classes are normal. Then, from the t-test, p-value =0.0003 indicates that lenght3 between the two species are different. Since length3 of Bream has a lower bound of confidence interval(36.9) larger than the upper bound(34.7) of length3 of Perch, Bream has significantly greater length3 than Perch.

The TTEST Procedure

Variable: length3

species	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Bream		38.5733	36.9294	40.2173	4.4025	3.5062	5.9184
Perch		31.0033	27.3469	34.6598	9.7921	7.7985	13.1637
Diff (1-2)	Pooled	7.5700	3.6463	11.4937	7.5917	6.4266	9.2767
Diff (1-2)	Satterthwaite	7.5700	3.6092	11.5308			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	58	3.86	0.0003
Satterthwaite	Unequal	40.264	3.86	0.0004

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	29	29	4.95	<.0001

Problem 1.(j)

Here is the code that produces the log transformation and goodness of fit tests for **weight**:

```
data log_fishy;
 set fishy;
 lnweight = log(weight);
run;
proc univariate data= log_fishy;
 var lnweight;
 by species;
run;
```

After performing the log transformation, goodness of fit tests are used and the results show that p-value >0.05 for Bream indicates that **weight** of Bream is normal. However, since the p-value =0.0032 < 0.05 for Shapiro-Wilk test for Perch, the **weight** of Perch is not normal. Hence, using a log transformation of weight **does not** allow for a two sample t-test of the mean weight to be performed by species.

Variable: lnweight

species=Bream

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	95.43263	Pr > |t|	<.0001
Sign	M	15	Pr >= |M|	<.0001
Signed Rank	S	232.5	Pr >= |S|	<.0001

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.944478	Pr < W	0.1201
Kolmogorov-Smirnov	D	0.130407	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.073859	Pr > W-Sq	0.2436
Anderson-Darling	A-Sq	0.49411	Pr > A-Sq	0.2086

The UNIVARIATE Procedure
Variable: lnweight

species=Perch

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	27.20234	Pr > |t|	<.0001
Sign	M	15	Pr >= |M|	<.0001
Signed Rank	S	232.5	Pr >= |S|	<.0001

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.882537	Pr < W	0.0032
Kolmogorov-Smirnov	D	0.115726	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.085885	Pr > W-Sq	0.1717
Anderson-Darling	A-Sq	0.761639	Pr > A-Sq	0.0437