Log Transforms example on Rainseed Data

PROC IMPORT OUT= MTH567.rainseed;
   DATAFILE= "E:\Excel\CASE0301.XLS"
   DBMS=EXCEL2000 REPLACE;
   GETNAMES=YES;
RUN;

proc univariate data=mth567.rainseed normal plots;
   class treatment;
   var rainfall;
   QQplot / normal (mu=est sigma=est color=blue);
   histogram;
run;

<table>
<thead>
<tr>
<th>Test</th>
<th>Statistic</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shapiro-Wilk</td>
<td>W 0.656264</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Kolmogorov-Smirnov</td>
<td>D 0.296602</td>
<td>&lt;0.0100</td>
</tr>
<tr>
<td>Cramer-von Mises</td>
<td>WSq 0.646116</td>
<td>&lt;0.0050</td>
</tr>
<tr>
<td>Anderson-Darling</td>
<td>ASq 3.39107</td>
<td>&lt;0.0050</td>
</tr>
</tbody>
</table>
Tests for Normality

<table>
<thead>
<tr>
<th>Test</th>
<th>Statistic</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shapiro-Wilk</td>
<td>W 0.602159</td>
<td>Pr &lt; W &lt;0.0001</td>
</tr>
<tr>
<td>Kolmogorov-Smirnov</td>
<td>D 0.290989</td>
<td>Pr &gt; D &lt;0.0100</td>
</tr>
<tr>
<td>Cramer-von Mises</td>
<td>W-Sq 0.705389</td>
<td>Pr &gt; W-Sq &lt;0.0050</td>
</tr>
<tr>
<td>Anderson-Darling</td>
<td>A-Sq 3.801896</td>
<td>Pr &gt; A-Sq &lt;0.0050</td>
</tr>
</tbody>
</table>

$H_0$: the data is Normally Distributed
$H_1$: the data is not Normally Distributed

We see that both the Seeded and Unseeded Cloud rainfall data fails the Normality tests because the P-values are below 0.05.

*yes – note that ln is log in SAS, common log (log) is log10 in SAS;

data mth567.rainseed2;
set mth567.rainseed;
lnrain=log(rainfall);
run;

*proc univariate data=mth567.rainseed2 normal plots;
*class treatment;
*var lnrain;
*QQplot / normal (mu=est sigma=est color=blue); histogram;
*run;
Tests for Normality

<table>
<thead>
<tr>
<th>Test</th>
<th>Statistic</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shapiro-Wilk</td>
<td>W</td>
<td>0.965906</td>
</tr>
<tr>
<td>Kolmogorov-Smirnov</td>
<td>D</td>
<td>0.135264</td>
</tr>
<tr>
<td>Cramer-von Mises</td>
<td>W-Sq</td>
<td>0.067552</td>
</tr>
<tr>
<td>Anderson-Darling</td>
<td>A-Sq</td>
<td>0.370332</td>
</tr>
</tbody>
</table>

After the transformation, both the Seeded and Unseeded rainfall follows a Normal distribution.

The TTEST Procedure

Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>TREATMENT</th>
<th>Lower CL</th>
<th>Upper CL</th>
<th>Lower CL</th>
<th>Upper CL</th>
</tr>
</thead>
<tbody>
<tr>
<td>lnrain</td>
<td>SEEDED</td>
<td>4.4881</td>
<td>5.1342</td>
<td>5.7802</td>
<td>1.2544</td>
</tr>
<tr>
<td></td>
<td>UNSEEDED</td>
<td>3.3272</td>
<td>3.9904</td>
<td>4.6536</td>
<td>1.2876</td>
</tr>
<tr>
<td>lnrain</td>
<td>Diff (1-2)</td>
<td>0.2409</td>
<td>1.1438</td>
<td>2.0467</td>
<td>1.3562</td>
</tr>
</tbody>
</table>

T-Tests

| Variable | Method     | Variiances | DF  | t Value | Pr > |t| |
|----------|------------|------------|-----|---------|------|---|
| lnrain   | Pooled     | Equal      | 50  | 2.54    | 0.0141|
| lnrain   | Satterthwaite | Unequal    | 50  | 2.54    | 0.0141|

We see that there is a significant difference in the means of the log transformed values. The 95% confidence interval for the difference in means of the logged values is (.2409, 2.0467). When we exponentiate these values, we obtain:

\[
(e^{.2409}, e^{2.0467}) = (1.27, 7.74) .
\]

Thus we are 95% sure that the true ratio of how much more rainfall comes from seeded clouds over unseeded clouds is between 1.27 and 7.74. Our best guess for how much more rainfall occurs with seeding is \( e^{1.1438} = 3.14 \).

For practice, perform a log transform on the salary data and construct a 95% confidence interval for the ratio of how much more men were paid than women.