My Data Aren’t Normal: Now What?

Dr. Machelle Wilson
October 9 & 16, 2019
What to Do with Non-Normal Data

We are video recording this seminar so please hold questions until the end.

Thanks
Outline

- Why do we care?
- When do we not care?
- How can we tell?
- What to do?
  - Transformations
  - Non-parametric Tests
- SAS code and output
Why Do We Care if Our Data are Normal?

• Most of the common statistical methods you are familiar with assume that they are.
• Our inference is only as good as our model.
• If our data are too far from the normal model we are using, then our inference may be faulty. That is, our p-values may be wrong.
Example: Why Do We Care?

• One example where the data fail to be normal is that they are log normal.
• This is common for data that can’t be negative, have small means and large standard deviations.
• Examples include hospital length of stay, income, lengths of latent periods for infectious diseases, and plasma triglyceride concentrations.
Example: Why Do We Care?

(a) Normal distribution

(b) Log-normal distribution
Why Do We Care?

Figure 3.3: Alpha and Beta Errors
Why Do We Care?

![Graph showing the relationship between effect size and sample size, with different tests and error rates.]
Why Do We Care?

![Graph showing type-I error vs. sample size for different tests.](chart.png)
So, Why Do We Care?

• We want to be able to detect differences between treatment and placebo in a reliable manner, with known power and confidence.

• That is, we want our statistical test to do what we designed it to do.
When Do We Not Care?

• At large sample sizes:
  the power and confidence levels of the naïve t test are quite close to what they should be, even for non-normal data.

• This is generally true for statistical analyses –
  • the larger the sample size, the closer the distribution of the mean (or other parameter estimates such as regression coefficients) is to normal.
When Do We Not Care?

• Just how large the sample size needs to be depends on the severity of the non-normality of the data.

• There is no easy or hard and fast way to know when the sample size is large enough.

• 

https://www.youtube.com/watch?v=dlbkaurTAUg
How to Tell if Your Data are Not Normal?
OK, What to do with Small Sample Sizes?

- There are three main approaches to handling non-normal data:
  - Transform the data from continuous to categorical
  - Transform the data to achieve normality,
  - Or use a non-parametric test.
What to Do?

• The first type of transformation is to convert the continuous data to categorical. For example:
  • HLOS (days) \(\rightarrow\) categorical:
    • < 7 days,
    • 7 – 30 days,
    • > 30 days.
  • This is a good option if there are natural, intuitive, or clinically meaningful categories.
What to Do With Non-Normal Data

- Find a transformation that makes the data normal.
  - For example, taking the natural or base 10 log.
  - Taking the square root.
  - There are many others.
- We will discuss the log transformation at length.
What to Do?

• Use a non-parametric test that does not require the assumption of normality.

• We will discuss:
  • For independent samples:
    • Wilcoxon rank sum test.
      • Kruskal-Wall/Mann-Whitney (SAA).
  • For paired data:
    • Signed rank test
What to Do?

- **Comparison of Means:**
  - In a simple comparison of means, it is easiest to simply use a non-parametric test rather than trying to find the right transformation.
  - The exception might be taking the log if the data are clearly log-normal.
  - For both log-transformed and non-parametric approaches, the comparison becomes between a comparison of the medians rather than the mean.
What to Do?

- **Regression Models:**
  - Find the right transformation (can be very tedious and frustrating).
  - Do a non-parametric regression (but they involve more advanced techniques).
  - Find a statistician.
What to Do? More Transformations
How to Check: *SAS* Code & Output

```sas
proc sort data=hlos;
  by treatment; /* sort by treatment */
run;

proc univariate data=hlos;
  var hlos;
  by treatment; /* view histograms for each treatment, separately. */
  histogram;
run;
```

<table>
<thead>
<tr>
<th>Extreme Observations</th>
<th>Lowest</th>
<th>Highest</th>
<th>Obs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>(trt = 0)</strong></td>
<td>Value</td>
<td>Value</td>
<td></td>
</tr>
<tr>
<td>4.68540</td>
<td>2</td>
<td>205.629</td>
<td>13</td>
</tr>
<tr>
<td>5.94221</td>
<td>7</td>
<td>247.320</td>
<td>3</td>
</tr>
<tr>
<td>8.67996</td>
<td>6</td>
<td>547.812</td>
<td>23</td>
</tr>
<tr>
<td>19.27006</td>
<td>20</td>
<td>570.694</td>
<td>24</td>
</tr>
<tr>
<td>21.57602</td>
<td>22</td>
<td>941.425</td>
<td>8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Extreme Observations</th>
<th>Lowest</th>
<th>Highest</th>
<th>Obs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>(trt = 1)</strong></td>
<td>Value</td>
<td>Value</td>
<td></td>
</tr>
<tr>
<td>6.32026</td>
<td>43</td>
<td>136.658</td>
<td>38</td>
</tr>
<tr>
<td>8.12841</td>
<td>36</td>
<td>154.181</td>
<td>32</td>
</tr>
<tr>
<td>10.28102</td>
<td>40</td>
<td>206.994</td>
<td>42</td>
</tr>
<tr>
<td>11.99024</td>
<td>37</td>
<td>283.190</td>
<td>44</td>
</tr>
<tr>
<td>15.61117</td>
<td>34</td>
<td>1079.286</td>
<td>45</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Extreme Observations</th>
<th>Lowest</th>
<th>Highest</th>
<th>Obs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>(Trt=2)</strong></td>
<td>Value</td>
<td>Value</td>
<td></td>
</tr>
<tr>
<td>0.779388</td>
<td>81</td>
<td>118.302</td>
<td>78</td>
</tr>
<tr>
<td>1.004175</td>
<td>62</td>
<td>158.992</td>
<td>57</td>
</tr>
<tr>
<td>1.699970</td>
<td>74</td>
<td>161.207</td>
<td>67</td>
</tr>
<tr>
<td>4.229360</td>
<td>70</td>
<td>559.306</td>
<td>68</td>
</tr>
<tr>
<td>8.781664</td>
<td>69</td>
<td>751.933</td>
<td>66</td>
</tr>
</tbody>
</table>
SAS Output

- Histogram of HLOS for Treatment 0:
SAS Output

- Histogram of HLOS for Treatment 1:
SAS Output

- Histogram of HLOS for Treatment 2:
The histograms show that the data have an approximately log normal distribution. So we will take the natural log and then see if the histograms are improved.

Now we repeat proc univariate using the log transformed variable

```sas
data hlos; /* using data step to add to the data */
set hlos;
logHLOS = log(hlos); /* taking the natural log */
run;
```

```sas
proc univariate data=hlos;
var loghlos;
by treatment;
histogram;
run;
```
SAS Output

- Histogram log(HLOS) for Treatment 0:
SAS Output

• Histogram for Log(HLOS) for Treatment 1:
SAS Output

- Histogram of log(HLOS) for Treatment 2:
Now that the data are approximately normal we can perform a normal ANOVA.

```
proc anova data=hlos;
class treatment;
model loghlos = treatment;
means treatment;
run;
quit;
```
SAS Output: Raw

Distribution of HLOS

- F: 1.15
- Prob > F: 0.3226

HLOS

- 0
- 1
- 2

Treatment

- 45
- 8
- 66
- 68
SAS Output: Log transformed
Non-parametric Tests

• **Comparison of Means**
  - For comparing means from independent samples that are not normal we can also use the SAS procedure `npar1way`.
  - This procedure will fit the Wilcoxon rank sum test for 2 sample designs and the Kruskal-Wallis test for designs with 3 or more.
  - This works well if transforming the data isn’t working.
  - It’s also very common to use these tests for Likert-Scale-type data.
SAS Code

```
proc sort data=hlos;
   by treatment; /* sort by treatment */
run;

proc means data=hlos n median min q1 q3 max;
   /* Use proc means to get medians and IQRs. */
   var hlos;
   by treatment;
run;

proc npar1way data=hlos wilcoxon;
   /*Always specify Wilcoxon or you'll get a 100 pages of output.*/
   class treatment;
   var hlos;
run;
```
### SAS Output

#### treatment=0

<table>
<thead>
<tr>
<th>Analysis Variable : HLOS</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>24</td>
<td>24</td>
<td>24</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td>Median</td>
<td>100.4051498</td>
<td>100.4051498</td>
<td>100.4051498</td>
<td>100.4051498</td>
<td>100.4051498</td>
</tr>
<tr>
<td>Minimum</td>
<td>4.6853989</td>
<td>4.6853989</td>
<td>4.6853989</td>
<td>4.6853989</td>
<td>4.6853989</td>
</tr>
<tr>
<td>Lower Quartile</td>
<td>32.7586077</td>
<td>32.7586077</td>
<td>32.7586077</td>
<td>32.7586077</td>
<td>32.7586077</td>
</tr>
<tr>
<td>Upper Quartile</td>
<td>175.1775236</td>
<td>175.1775236</td>
<td>175.1775236</td>
<td>175.1775236</td>
<td>175.1775236</td>
</tr>
<tr>
<td>Maximum</td>
<td>941.4254456</td>
<td>941.4254456</td>
<td>941.4254456</td>
<td>941.4254456</td>
<td>941.4254456</td>
</tr>
</tbody>
</table>

#### treatment=1

<table>
<thead>
<tr>
<th>Analysis Variable : HLOS</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>Median</td>
<td>50.2618567</td>
<td>50.2618567</td>
<td>50.2618567</td>
<td>50.2618567</td>
<td>50.2618567</td>
</tr>
<tr>
<td>Upper Quartile</td>
<td>103.6998143</td>
<td>103.6998143</td>
<td>103.6998143</td>
<td>103.6998143</td>
<td>103.6998143</td>
</tr>
<tr>
<td>Maximum</td>
<td>1079.29</td>
<td>1079.29</td>
<td>1079.29</td>
<td>1079.29</td>
<td>1079.29</td>
</tr>
</tbody>
</table>

#### treatment=2

<table>
<thead>
<tr>
<th>Analysis Variable : HLOS</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>27</td>
<td>27</td>
<td>27</td>
<td>27</td>
<td>27</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.7793879</td>
<td>0.7793879</td>
<td>0.7793879</td>
<td>0.7793879</td>
<td>0.7793879</td>
</tr>
<tr>
<td>Upper Quartile</td>
<td>80.9398000</td>
<td>80.9398000</td>
<td>80.9398000</td>
<td>80.9398000</td>
<td>80.9398000</td>
</tr>
<tr>
<td>Maximum</td>
<td>751.9330514</td>
<td>751.9330514</td>
<td>751.9330514</td>
<td>751.9330514</td>
<td>751.9330514</td>
</tr>
</tbody>
</table>

#### Kruskal-Wallis Test

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-Square</td>
<td>6.1983</td>
</tr>
<tr>
<td>DF</td>
<td>2</td>
</tr>
<tr>
<td>Pr &gt; Chi-Square</td>
<td>0.0451</td>
</tr>
</tbody>
</table>
Non-Parametric Tests

• Comparison of Paired Means
  • For paired means, we need a test appropriate for *dependent* data (analog to the paired $t$ test). The Wilcoxon test is *not* appropriate.
  • So, we first calculate the difference between the pre and post means for each patient.
  • Then use the one sample Wilcoxon signed rank test.
**SAS Code**

```sas
data paired; /* data step to calculate differences */
  set paired;
  delta = post - pre;
run;

proc means data=paired n median q1 q3; /* To get medians and IQR */
  var pre post;
run;

proc univariate data=paired;
  var delta; /* to get statistics and Signed Rank Test for differences */
run;
```
### SAS Output

<table>
<thead>
<tr>
<th>Variable</th>
<th>Label</th>
<th>N</th>
<th>Median</th>
<th>Lower Quartile</th>
<th>Upper Quartile</th>
</tr>
</thead>
<tbody>
<tr>
<td>pre</td>
<td>pre</td>
<td>20</td>
<td>3.50000000</td>
<td>2.00000000</td>
<td>4.00000000</td>
</tr>
<tr>
<td>post</td>
<td>post</td>
<td>20</td>
<td>4.00000000</td>
<td>3.00000000</td>
<td>5.00000000</td>
</tr>
</tbody>
</table>

### Basic Statistical Measures

<table>
<thead>
<tr>
<th>Location</th>
<th>Variability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1.100000</td>
</tr>
<tr>
<td>Median</td>
<td>1.000000</td>
</tr>
<tr>
<td>Mode</td>
<td>1.000000</td>
</tr>
<tr>
<td>Std Deviation</td>
<td>1.07115</td>
</tr>
<tr>
<td>Variance</td>
<td>1.14737</td>
</tr>
<tr>
<td>Range</td>
<td>4.00000</td>
</tr>
<tr>
<td>Interquartile Range</td>
<td>2.00000</td>
</tr>
</tbody>
</table>

### Tests for Location: Mu0=0

<table>
<thead>
<tr>
<th>Test</th>
<th>Statistic</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student's t</td>
<td>t</td>
<td>4.592575</td>
</tr>
<tr>
<td></td>
<td>Pr &gt;</td>
<td>t</td>
</tr>
<tr>
<td>Sign</td>
<td>M</td>
<td>6.5</td>
</tr>
<tr>
<td></td>
<td>Pr &gt;=</td>
<td>M</td>
</tr>
<tr>
<td>Signed Rank</td>
<td>S</td>
<td>55.5</td>
</tr>
<tr>
<td></td>
<td>Pr &gt;=</td>
<td>S</td>
</tr>
</tbody>
</table>
Non-Normal Data: correlation

• Pearson’s correlation measures the strength of the *linear* relationship between two variables.
• It ranges between -1 and +1, where values further from 0 indicate stronger correlation.
• When the data are not normal, continuous, or linearly related, Pearson’s correlation is not appropriate.
• **Spearman’s** correlation also measures the strength of the association and ranges between -1 and +1.
• However, it does not make assumptions of continuity, normality, or linearity.
• Spearman’s correlation only assumes that the relationship is *monotone*. 
Non-Normal Correlation

• SAS Code

```sas
proc sgplot data=hlos;
scatter x=hgbalc y=hlos;
run;
proc sgplot data=hlos;
scatter x=hgbalc y=loghlos;
run;
proc corr data=hlos spearman pearson;
var hlos loghlos;
with hgbalc;
run;
```
Non-Normal Correlation

- Scatter plot of HLOS by Hgb A1c
Non-Normal Correlation

- Scatter plot of log(HLOS) by Hgb A1c
## Non-Normal Correlation

### SAS Output

#### Pearson Correlation Coefficients, \( N = 81 \)

<table>
<thead>
<tr>
<th></th>
<th>HLOS</th>
<th>logHLOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>HgbA1c</td>
<td>0.14611</td>
<td>0.26932</td>
</tr>
<tr>
<td>HgbA1c</td>
<td>0.1931</td>
<td>0.0150</td>
</tr>
</tbody>
</table>

#### Spearman Correlation Coefficients, \( N = 81 \)

<table>
<thead>
<tr>
<th></th>
<th>HLOS</th>
<th>logHLOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>HgbA1c</td>
<td>0.28485</td>
<td>0.28485</td>
</tr>
<tr>
<td>HgbA1c</td>
<td>0.01000</td>
<td>0.01000</td>
</tr>
</tbody>
</table>
Non-Normal: Regression

- SAS code for the transformed HLOS

```sas
proc glm data=hlos plots=diagnostic;
model loghlos = hgbalc; /* log transformed HLOS as endpoint */
run;
quit;
```
Non-normal Regression

Regression Results for Raw HLOS

| Parameter | Estimate     | Standard Error | t Value | Pr > |t| |
|-----------|--------------|----------------|---------|------|---|
| Intercept | 27.75683122  | 50.26277625    | 0.55    | 0.5823 |
| HgbA1c    | 10.41426033  | 7.93333375     | 1.31    | 0.1931 |

Regression Results for Transformed HLOS

| Parameter | Estimate     | Standard Error | t Value | Pr > |t| |
|-----------|--------------|----------------|---------|------|---|
| Intercept | 2.778096911  | 0.45560075     | 6.10    | <.0001 |
| HgbA1c    | 0.178743597  | 0.07191073     | 2.49    | 0.0150 |
Diagnostic Plots: Raw HLOS

- Residual vs Predicted Value
- Fit Diagnostics for HLOS
- Leverage
- Residual vs Quantile
- HLOS vs Predicted Value
- Cook's D vs Observation
- Percent vs Residual
- Fit-Mean vs Residual

Parameters: 2
Error DF: 79
MSE: 15252
R-Square: 0.3213
Adj R-Square: 0.009
Diagnostic Plots: Log HLOS

- Residual vs Predicted Value
- rStudent vs Predicted Value
- Leverage vs rStudent
- Quantile vs Residual
- LogHLOS vs Predicted Value
- Cook's D vs Observation
- Percent vs Residual
- Fit-Mean vs Residual

Observations: 81
Parameters: 2
Error DF: 79
MSE: 1.2532
R-Square: 0.0525
Adj R-Square: 0.0008
Interpreting the Coefficients in a Regression Model

• The correct interpretation of the coefficients of a regression model is that for every unit (whatever the units are) increase in the risk factor, the endpoint changes by beta units.

• For HLOS, pretending the model is correct, we have:
  • For every percent increase in HgbA1C, HLOS increases by 10.4 days. (HgbA1c is in units percent, HLOS in days.)

• Does this seem realistic?
Interpreting Coefficients of Log Transformed Regression Model

• But for log(HLOS) we no longer have units of days so how do we interpret the coefficients?
  • We back-transform (exponentiate) so we can once again have units that are understandable and clinically relevant.
  • We have that \( \exp(0.1787) = 1.196 \).
  • This is interpreted as the median HLOS (in days) increases by about 20% for every percent increase in HgbA1c.
Conclusion

- Non-parametric tests are the easiest solution for simple comparisons of means.
- Spearman’s correlation is easy to implement for non-linear, non-normal correlations.
- For regressions, a log (either natural or base 10) can often solve the problem, but requires a back-transformation to be interpretable.
- When in doubt, get help from a statistician.
Help is Available

- CTSC & Cancer Center Biostatistics Office Hours
  - Tuesdays from 12 – 1:30 in Sacramento
  - Sign up through the CTSC Biostatistics Website
- EHS Biostatistics Office Hours
  - Mondays from 2-4 in Davis. Sign up through EHS website
- Request Biostatistics Consultations
  - CTSC - [www.ucdmc.ucdavis.edu/ctsc/](http://www.ucdmc.ucdavis.edu/ctsc/)
  - MIND IDDRC - [www.ucdmc.ucdavis.edu/mindinstitute/centers/iddrc/cores/bbrd.html](http://www.ucdmc.ucdavis.edu/mindinstitute/centers/iddrc/cores/bbrd.html)
  - Cancer Center
    - [https://health.ucdavis.edu/cancer/research/sharedresources/biostatistics.html](https://health.ucdavis.edu/cancer/research/sharedresources/biostatistics.html)
  - EHS Center - [https://environmentalhealth.ucdavis.edu/core-resources](https://environmentalhealth.ucdavis.edu/core-resources)
References


• Biostatistics for the Clinician, URL: https://www.uth.tmc.edu/uth_orgs/educ_dev/oser/L3_o.HTM