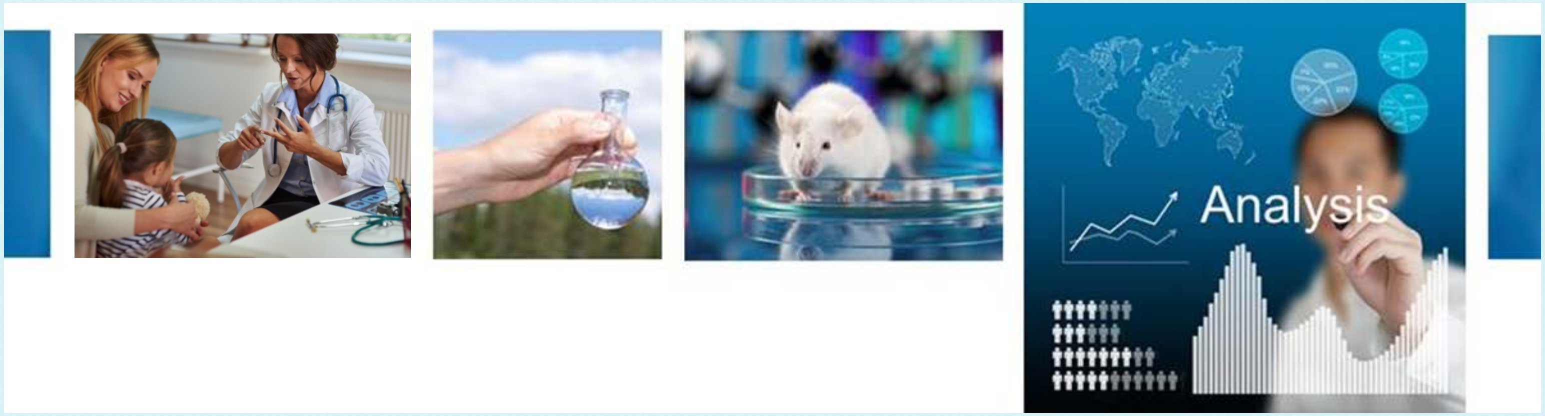


My Data Aren’t Normal: Now What?

Dr. Machelle Wilson

October 9 & 16, 2019

UCDAVIS HEALTH

- UC Davis Health Clinical and Translational Science Center
- UC Davis Health Mind Institute
- UC Davis Health Comprehensive Cancer Center
- UC Davis Environment Health Sciences Center

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

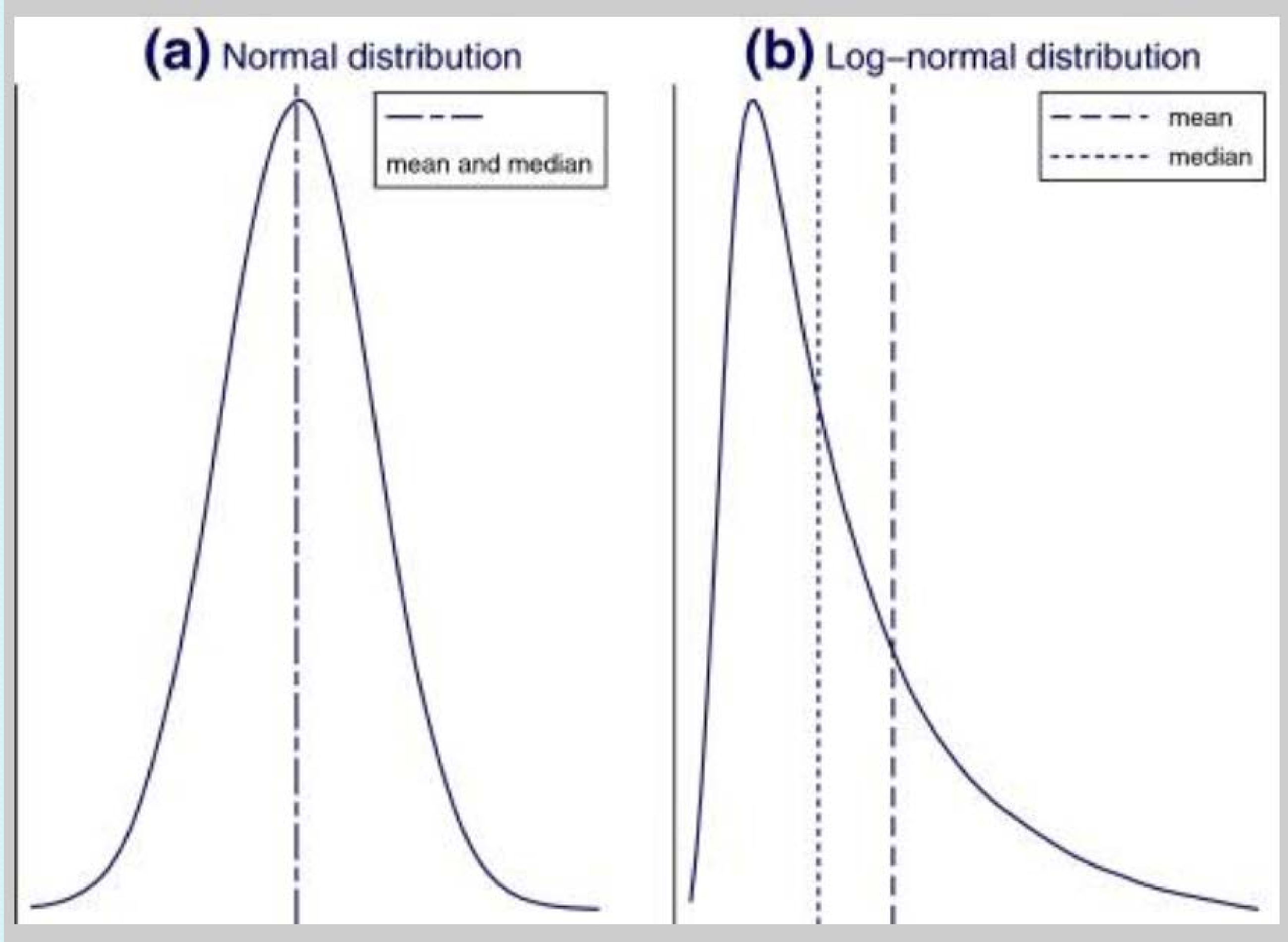
Why Do We Care if Our Data are Normal?

- 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?

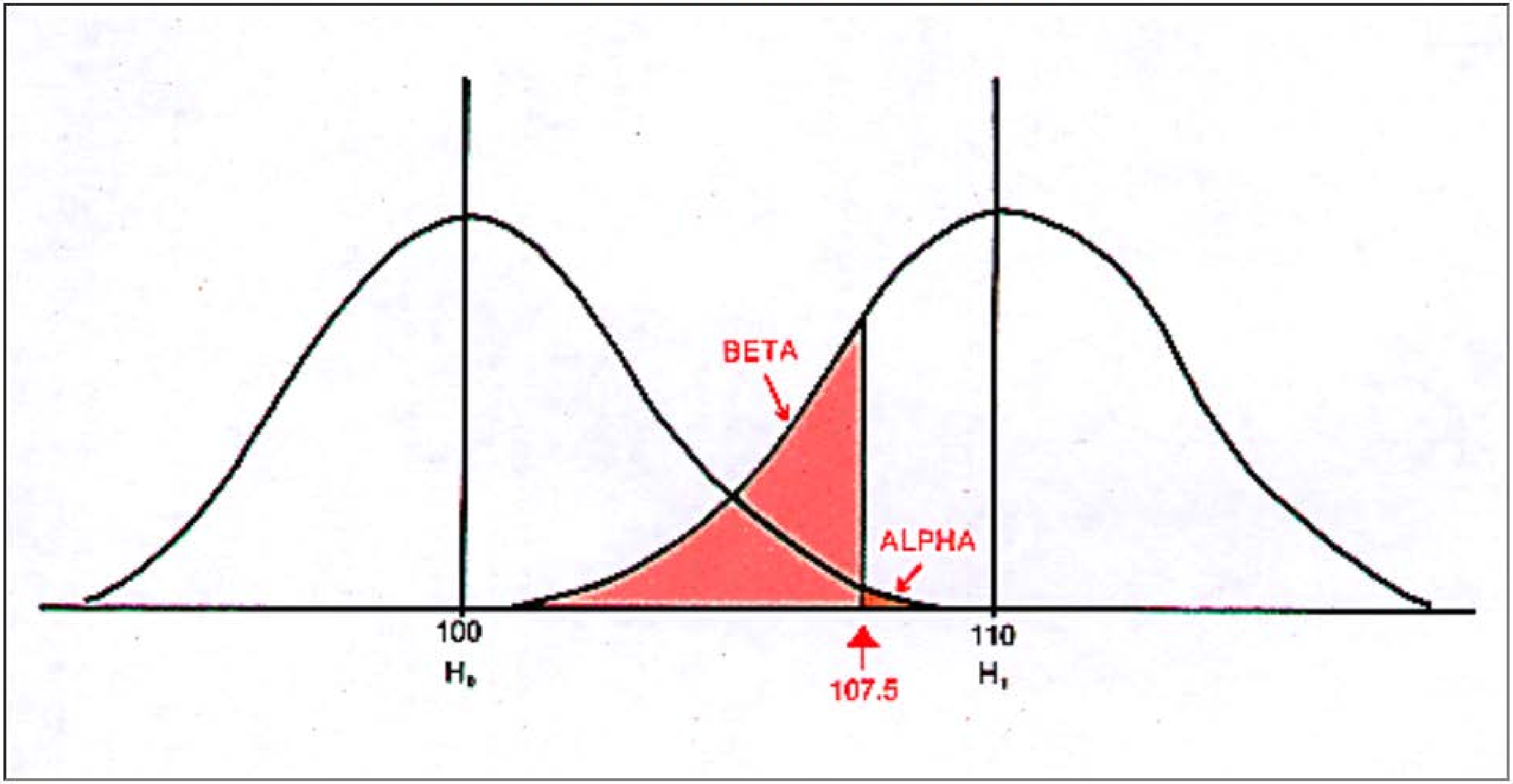
- One example where the data fail to be normal is that they are ***log*** normal.
- This is common for data that can’t be negative, have small means and large standard deviations.
- Examples include hospital length of stay, income, lengths of latent periods for infectious diseases, and plasma triglyceride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

Example: Why Do We Care?

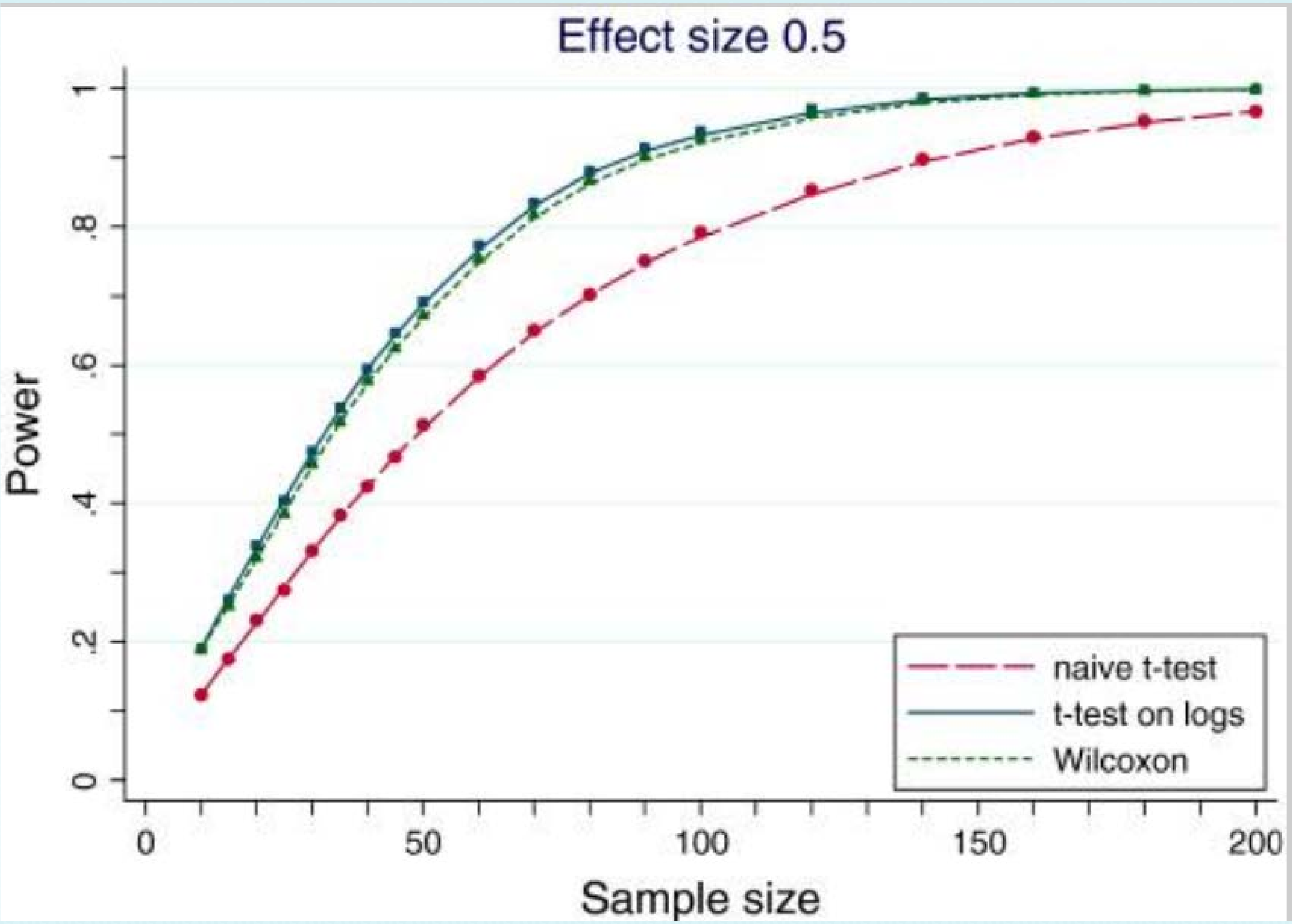


Why Do We Care?

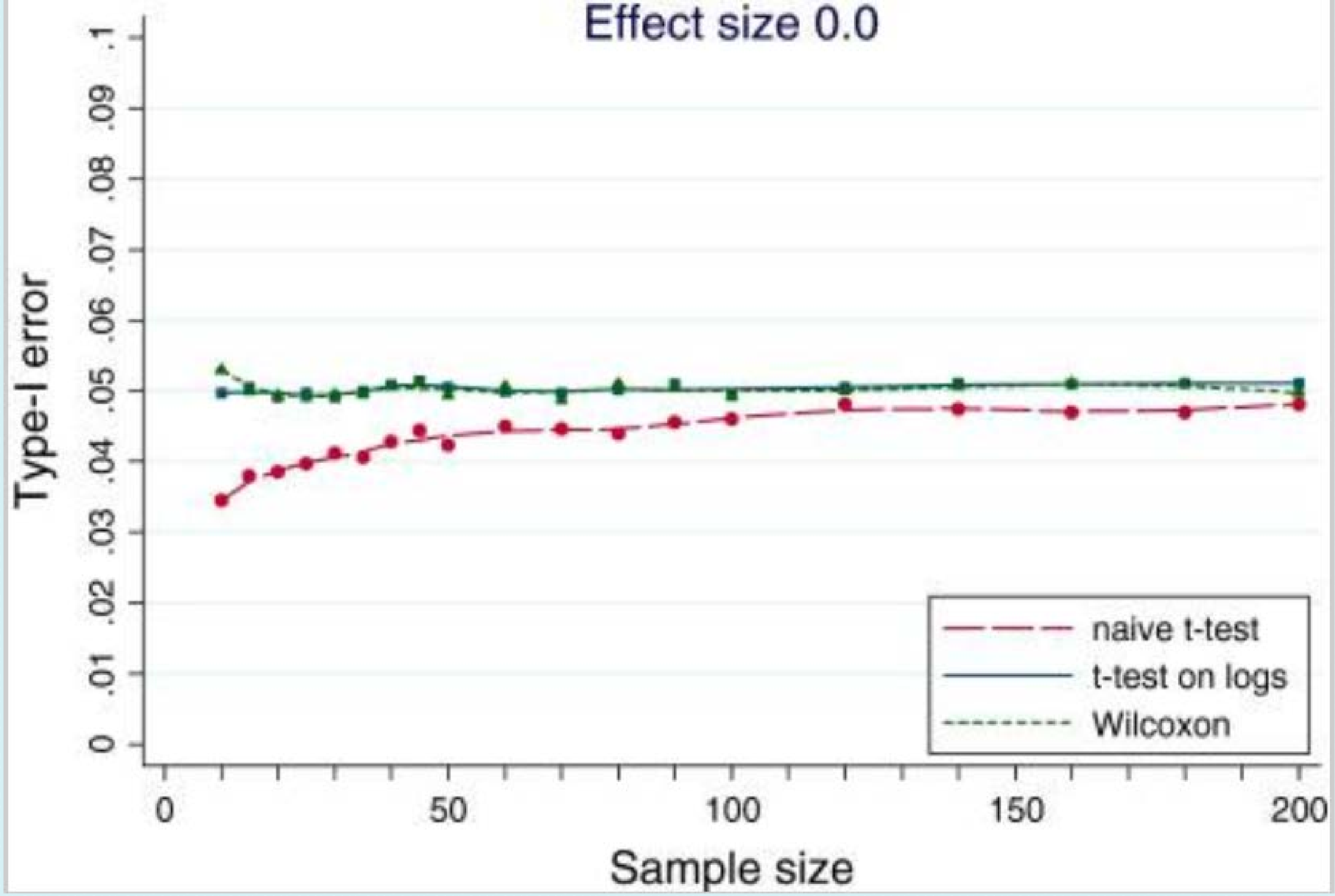
Figure 3.3: Alpha and Beta Errors



Why Do We Care?



Why Do We Care?



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:

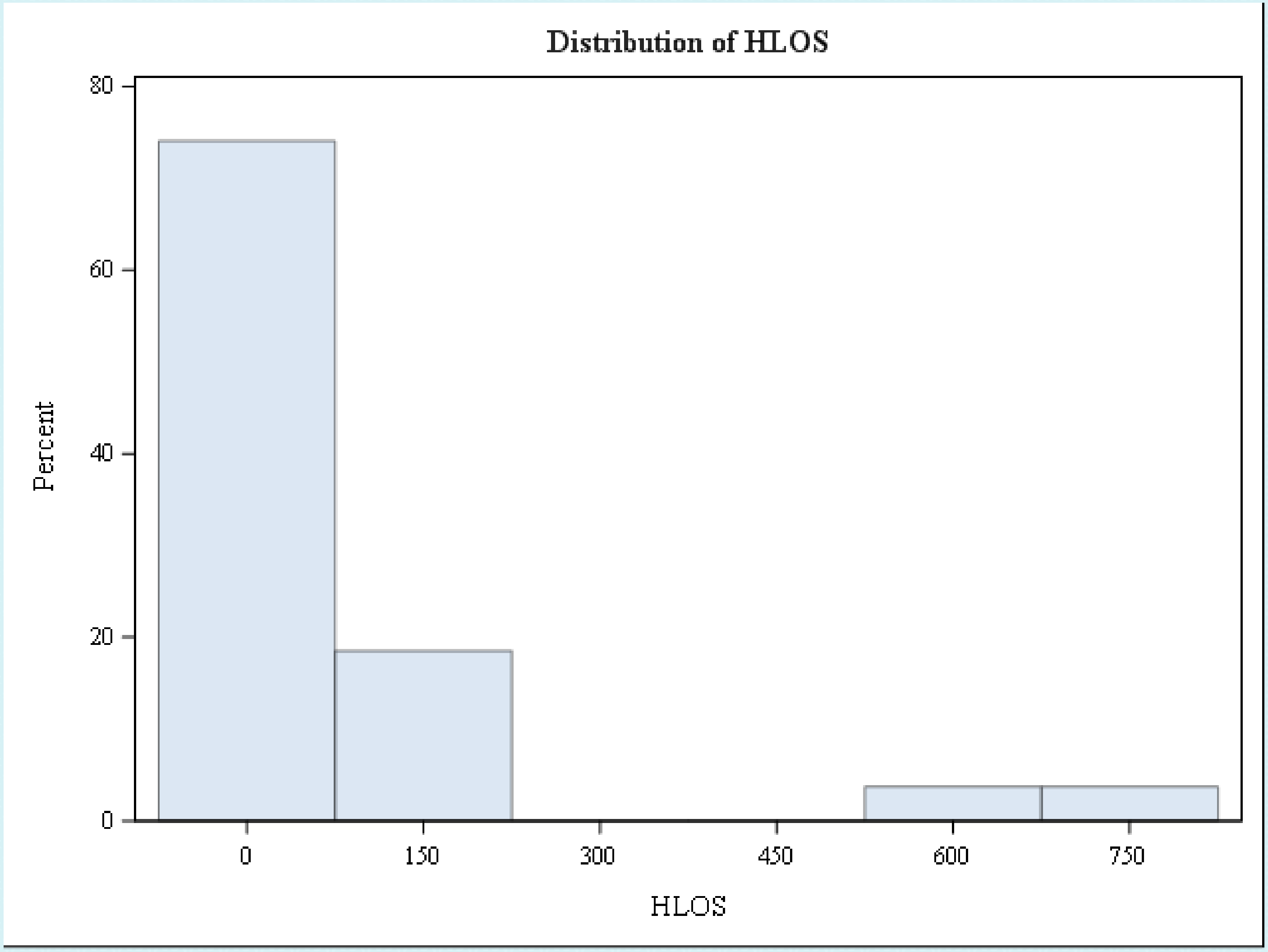
  the power and confidence levels of the naïve t test are quite close to what they should be, even for non-normal data.

- 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) → 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**

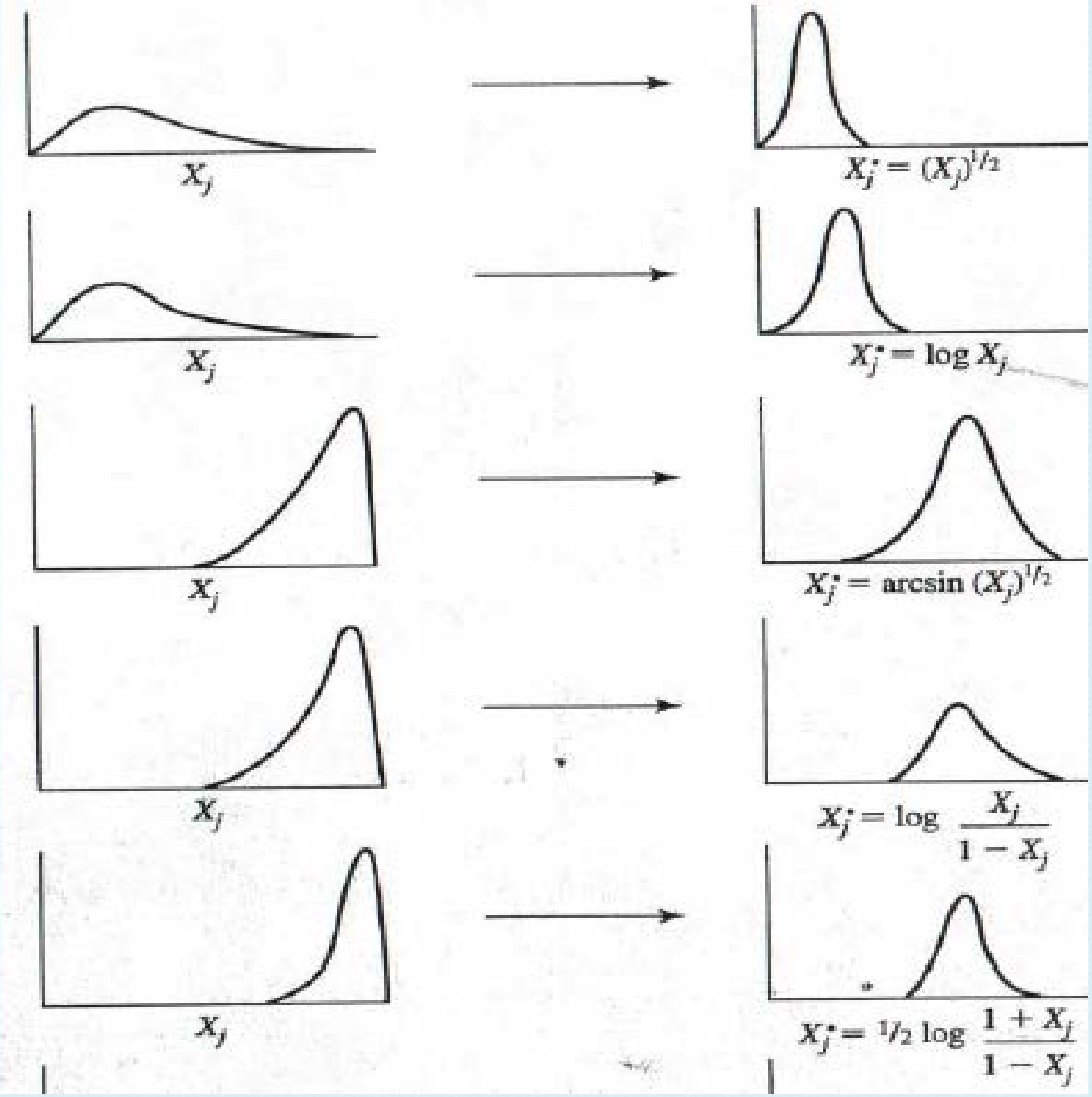
What to Do?

- **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.

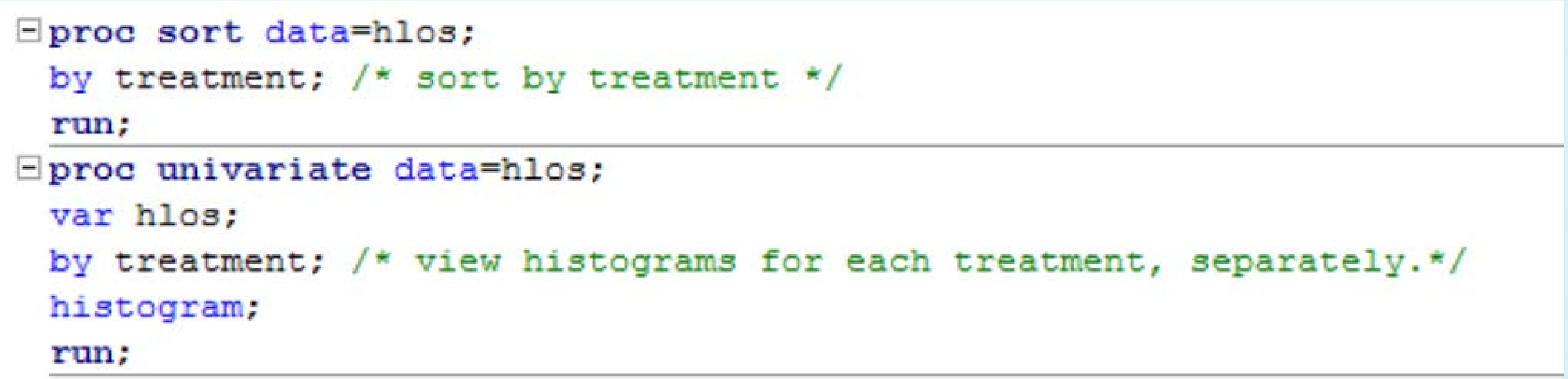
What to Do?

- **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


X_j
$X_j^* = (X_j)^{1/2}$
X_j
$X_j^* = \log X_j$
X_j
$X_j^* = \arcsin (X_j)^{1/2}$
X_j
$X_j^* = \log \frac{X_j}{1 - X_j}$
X_j
$X_j^* = {}^1/_2 \log \frac{1 + X_j}{1 - X_j}$

How to Check: SAS Code & Output



```
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;
```

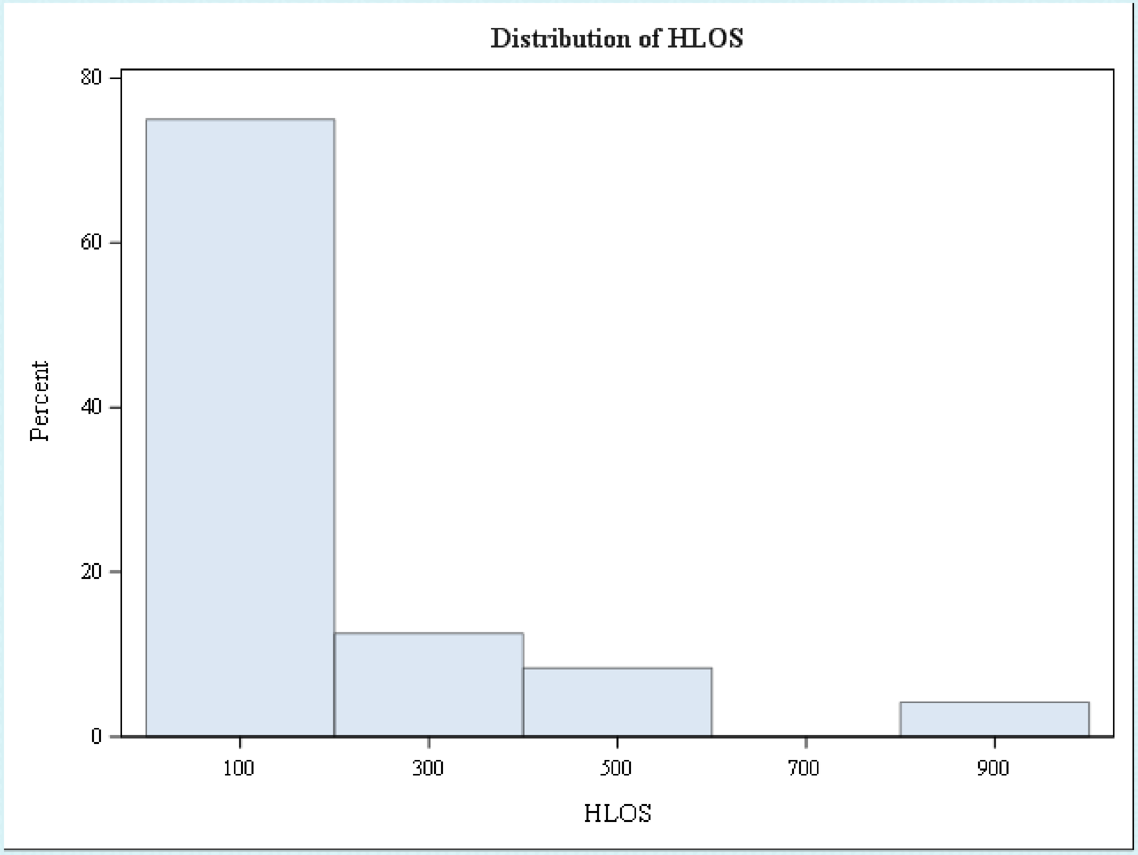
Extreme Observations (trt = 0)			
Lowest		Highest	
Value	Obs	Value	Obs
4.68540	2	205.629	13
5.94221	7	247.320	3
8.67996	6	547.812	23
19.27006	20	570.694	24
21.57602	22	941.425	8

Extreme Observations (trt = 1)			
Lowest		Highest	
Value	Obs	Value	Obs
6.32026	43	136.658	38
8.12841	36	154.181	32
10.28102	40	206.994	42
11.99024	37	283.190	44
15.61117	34	1079.286	45

Extreme Observations (Trt=2)			
Lowest		Highest	
Value	Obs	Value	Obs
0.779388	81	118.302	78
1.004175	62	158.992	57
1.699970	74	161.207	67
4.229360	70	559.306	68
8.781664	69	751.933	66

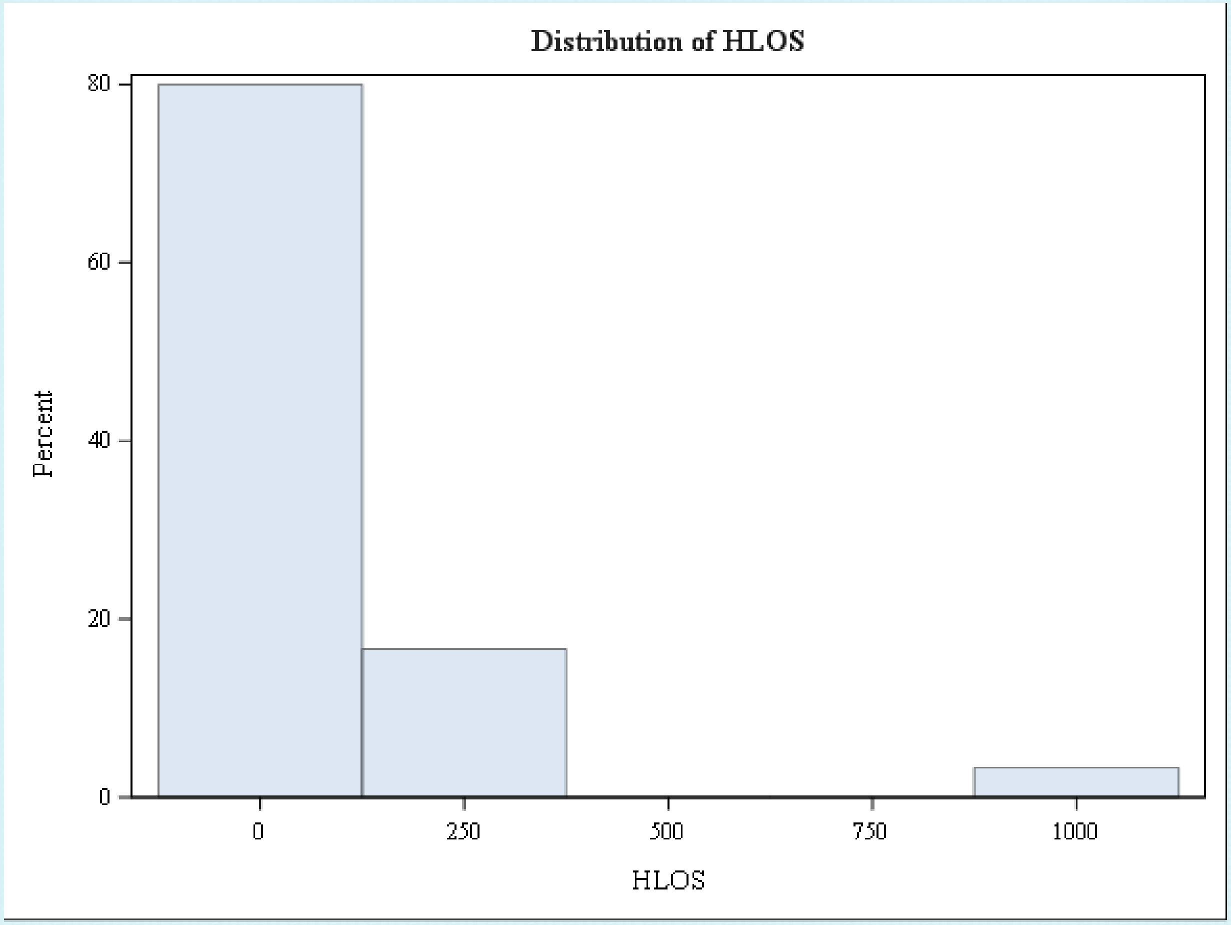
SAS Output

- Histogram of HLOS for Treatment 0:



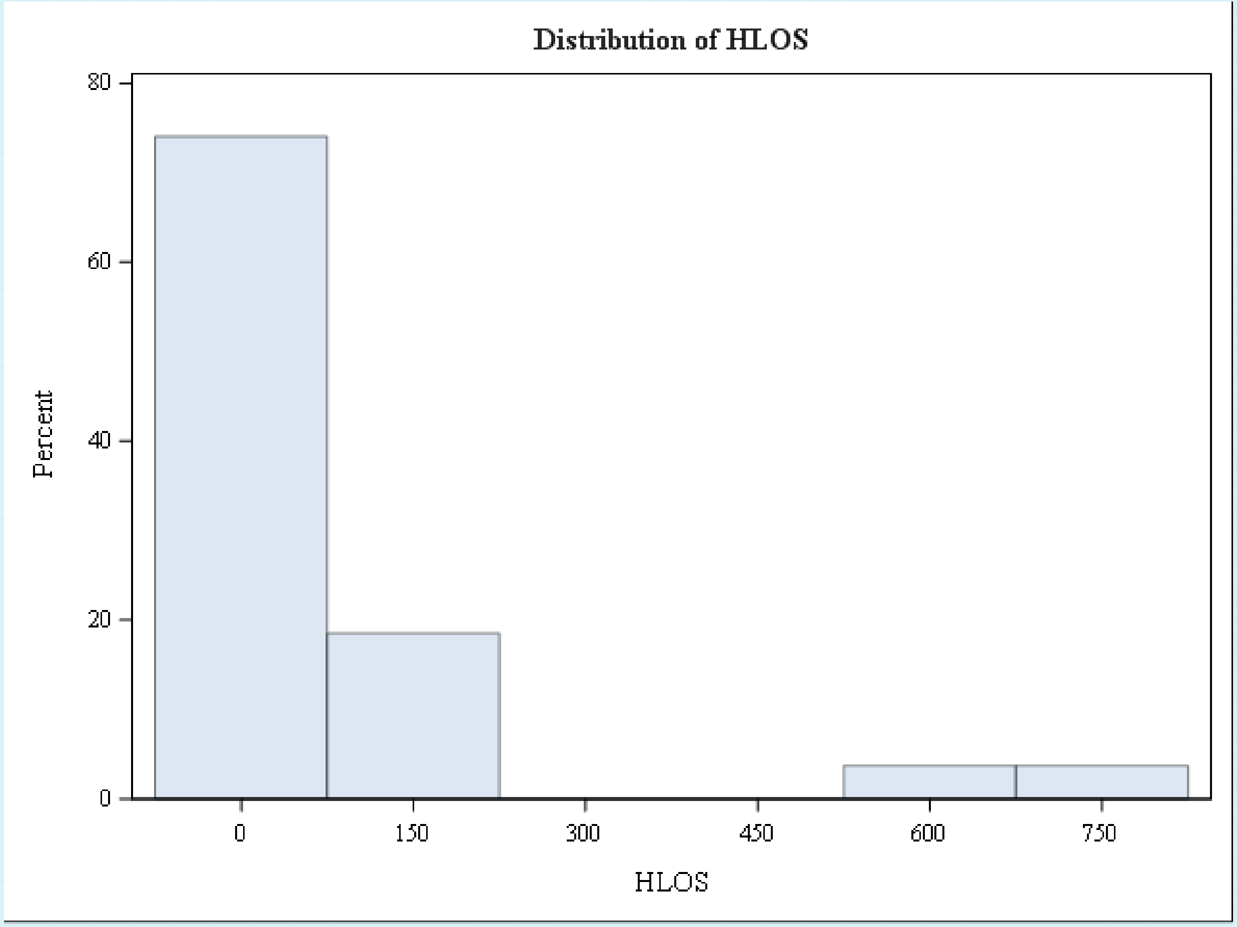
SAS Output

- Histogram of HLOS for Treatment 1:



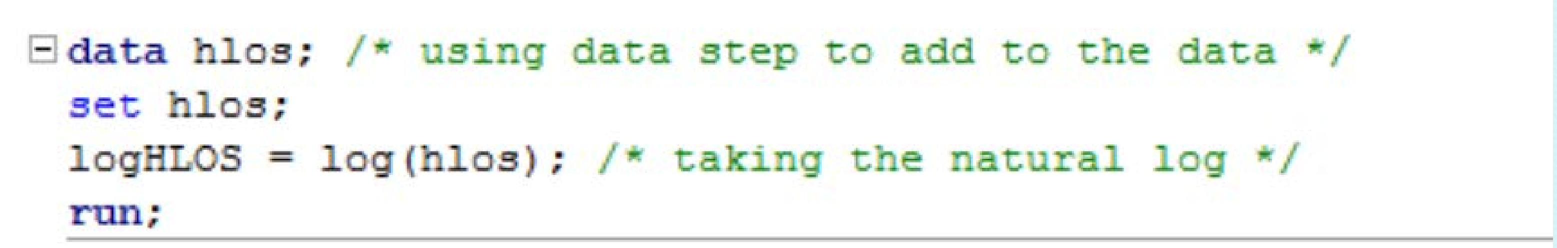
SAS Output

- Histogram of HLOS for Treatment 2:



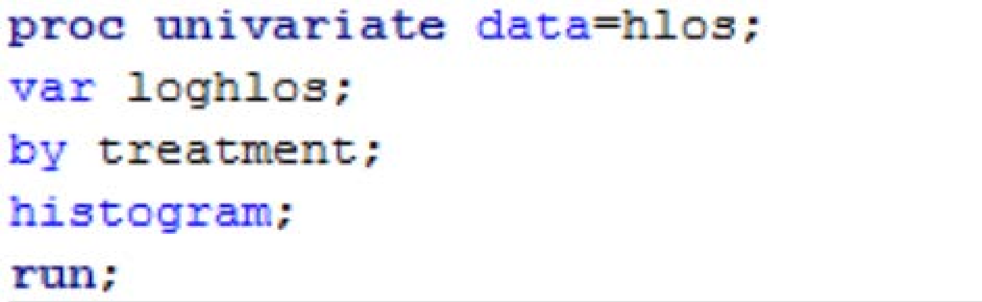
SAS Code

- 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.



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

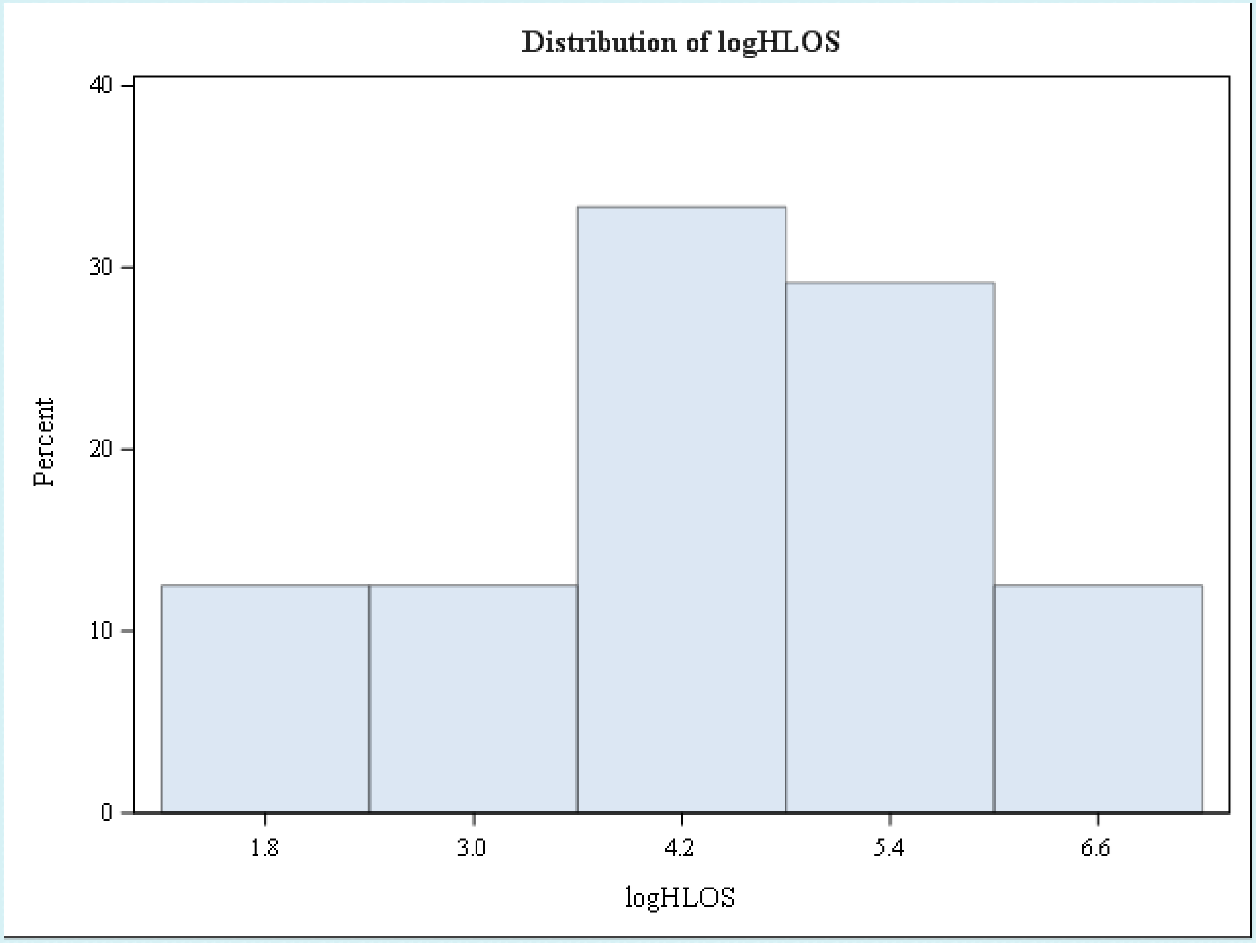
- Now we repeat proc univariate using the log transformed variable



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

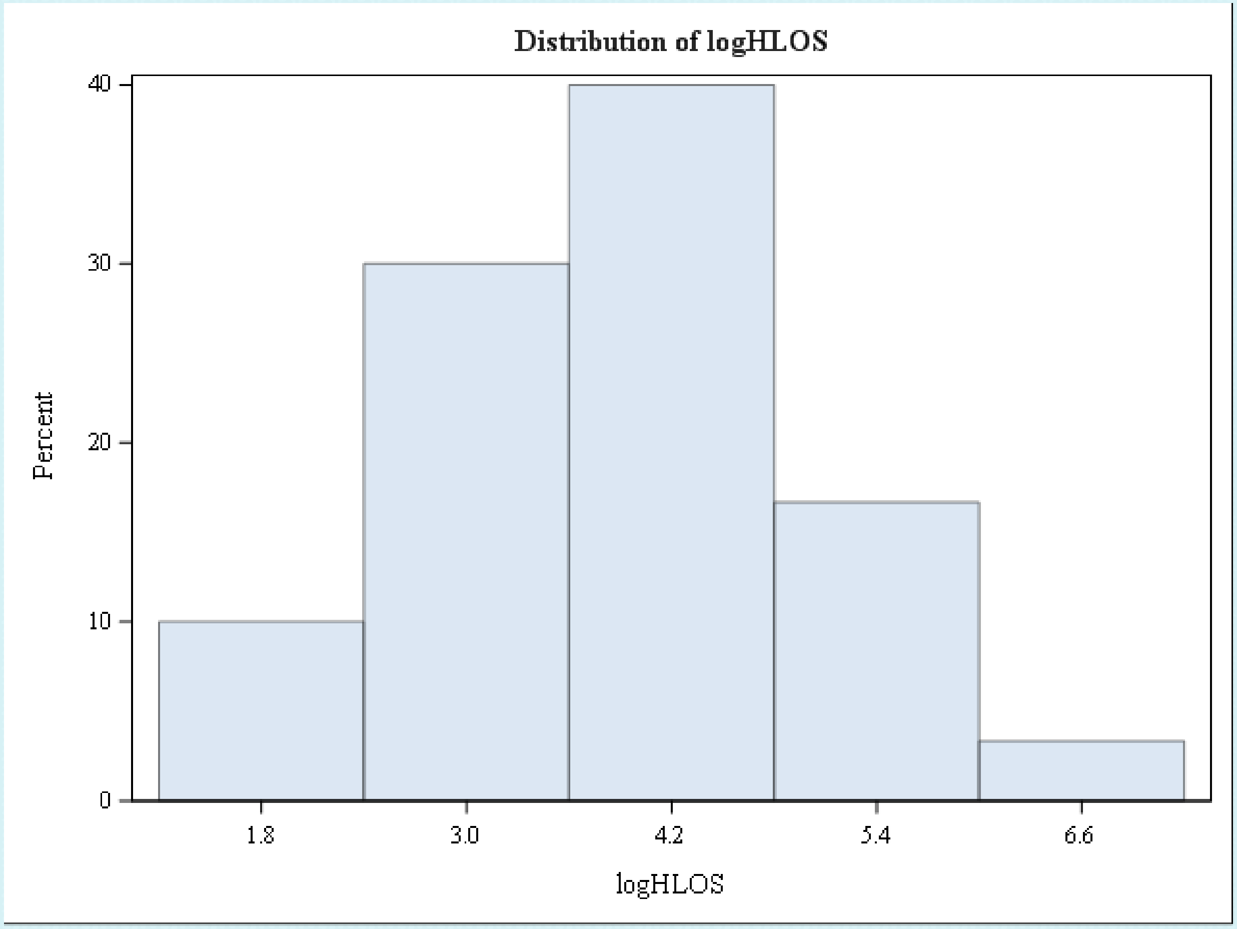
SAS Output

- Histogram log(HLOS) for Treatment 0:



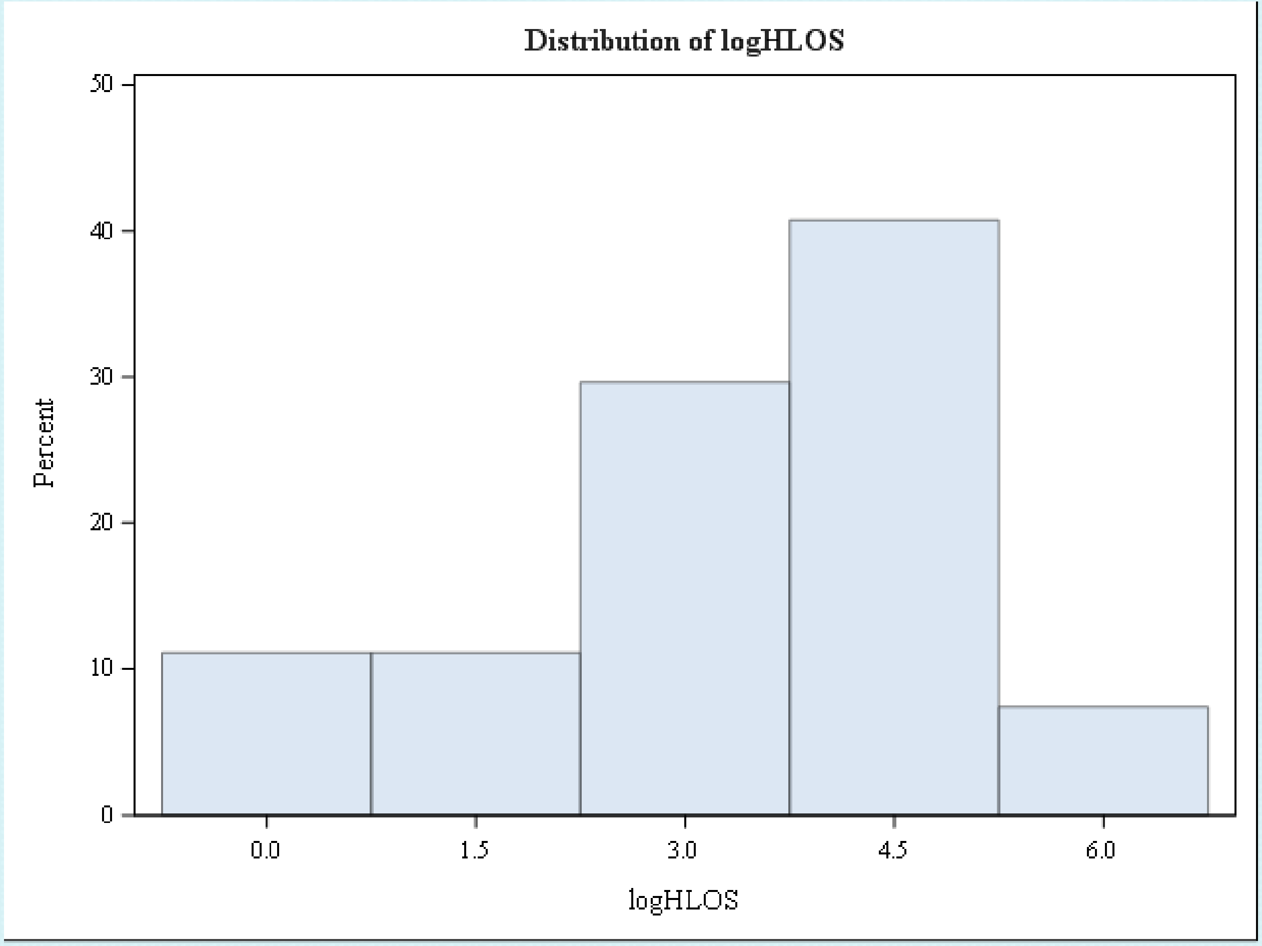
SAS Output

- Histogram for Log(HLOS) for Treatment 1:



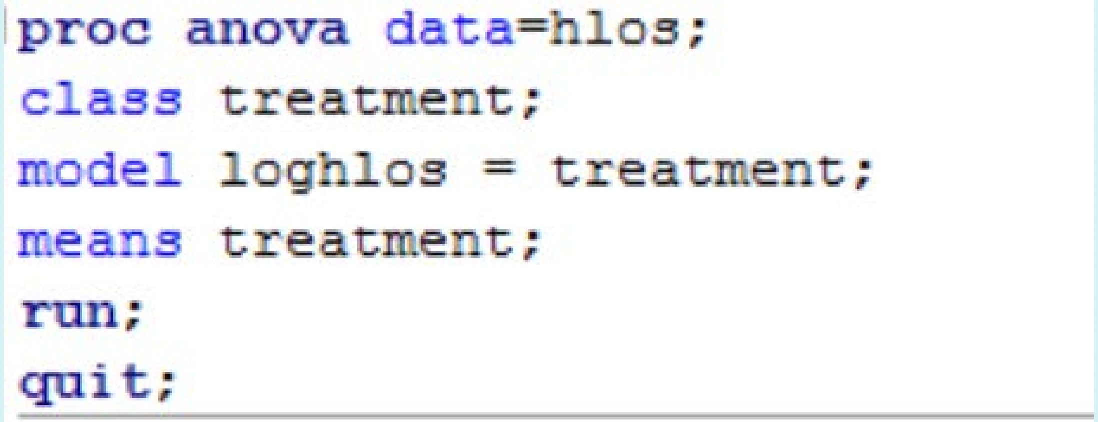
SAS Output

- Histogram of log(HLOS) for Treatment 2:



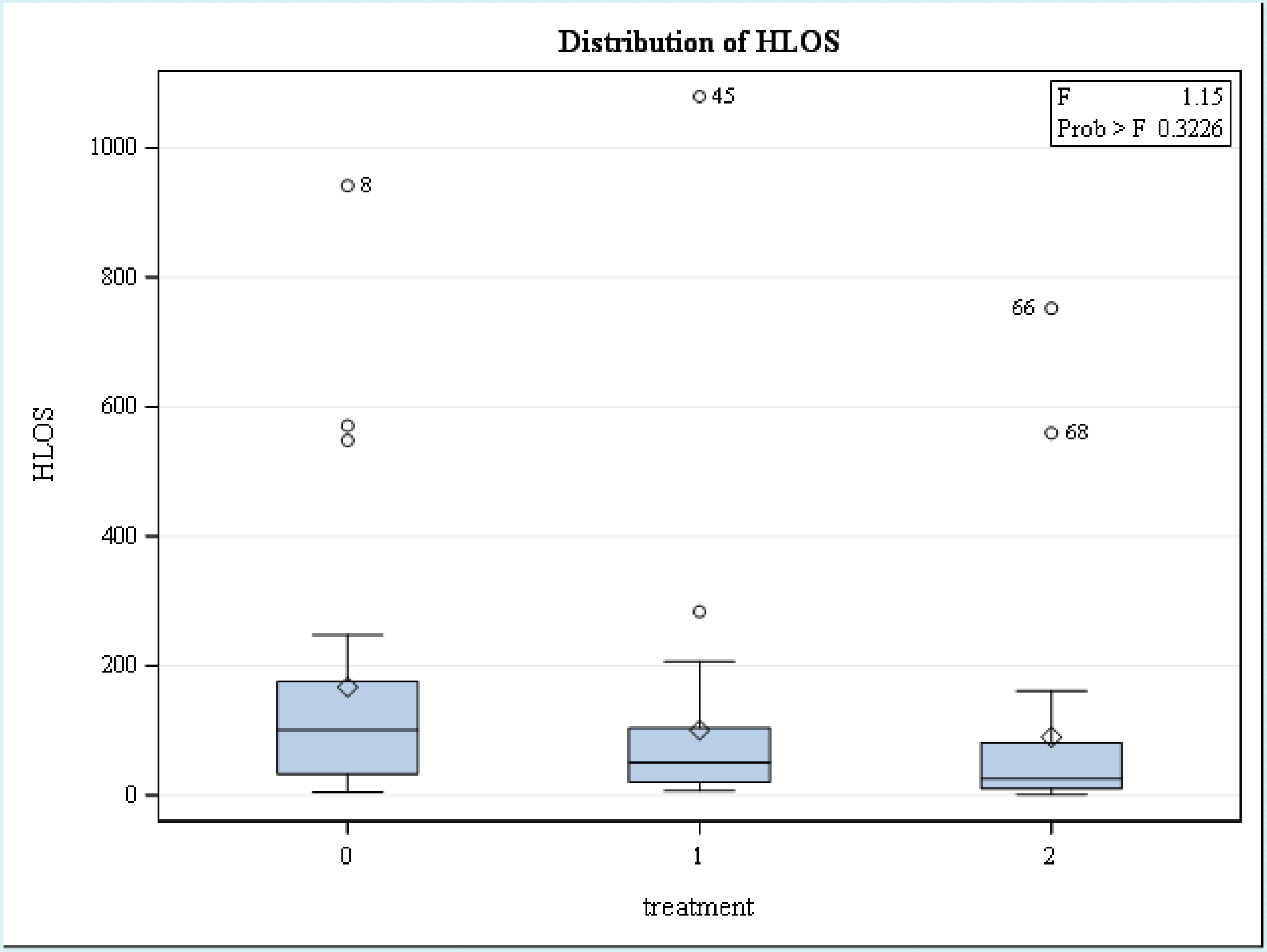
SAS Code

- Now that the data are approximately normal we can perform a normal ANOVA.

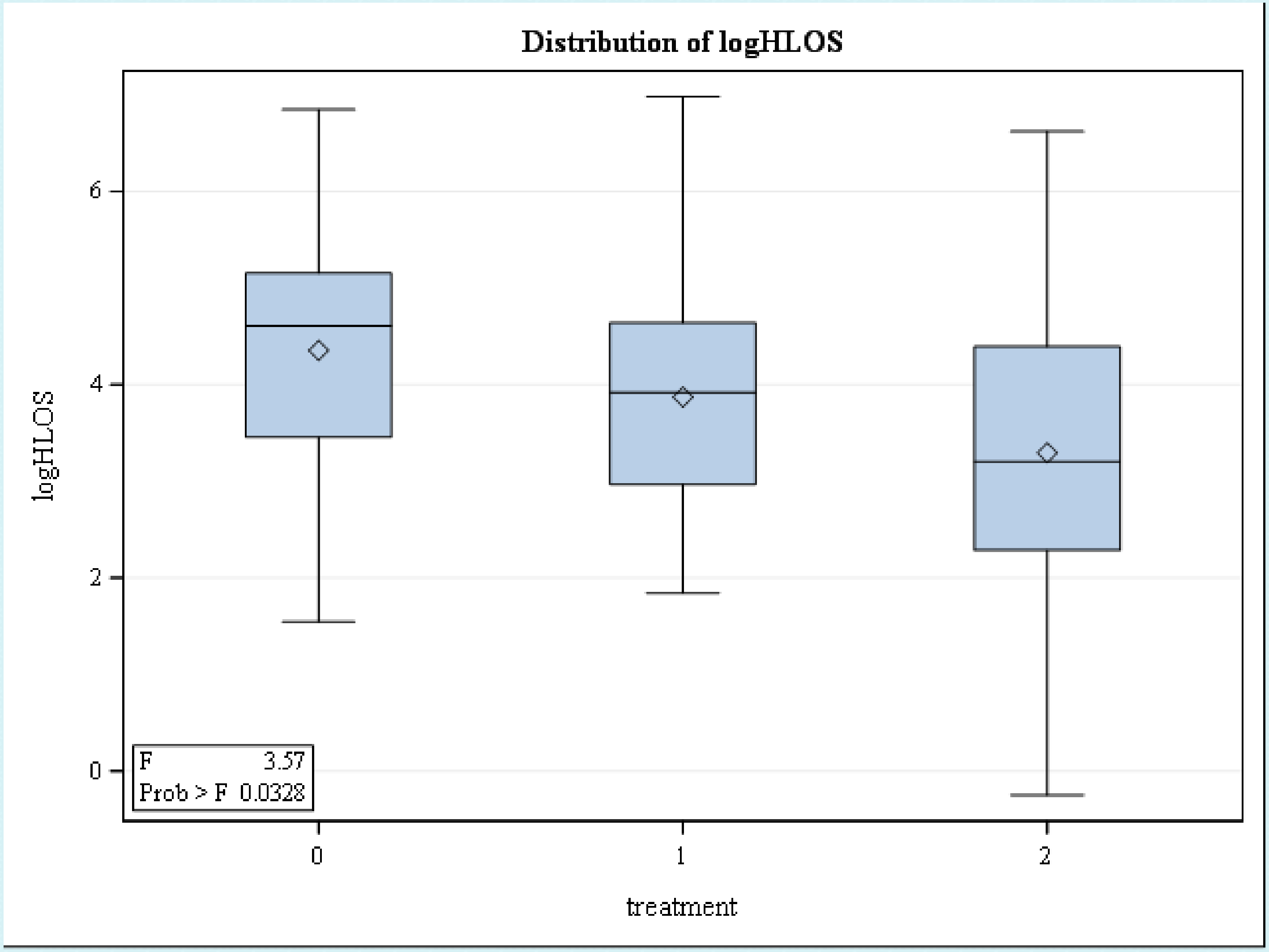


```
proc anova data=hlos;
class treatment;
model loghlos = treatment;
means treatment;
run;
quit;
```

SAS Output: Raw



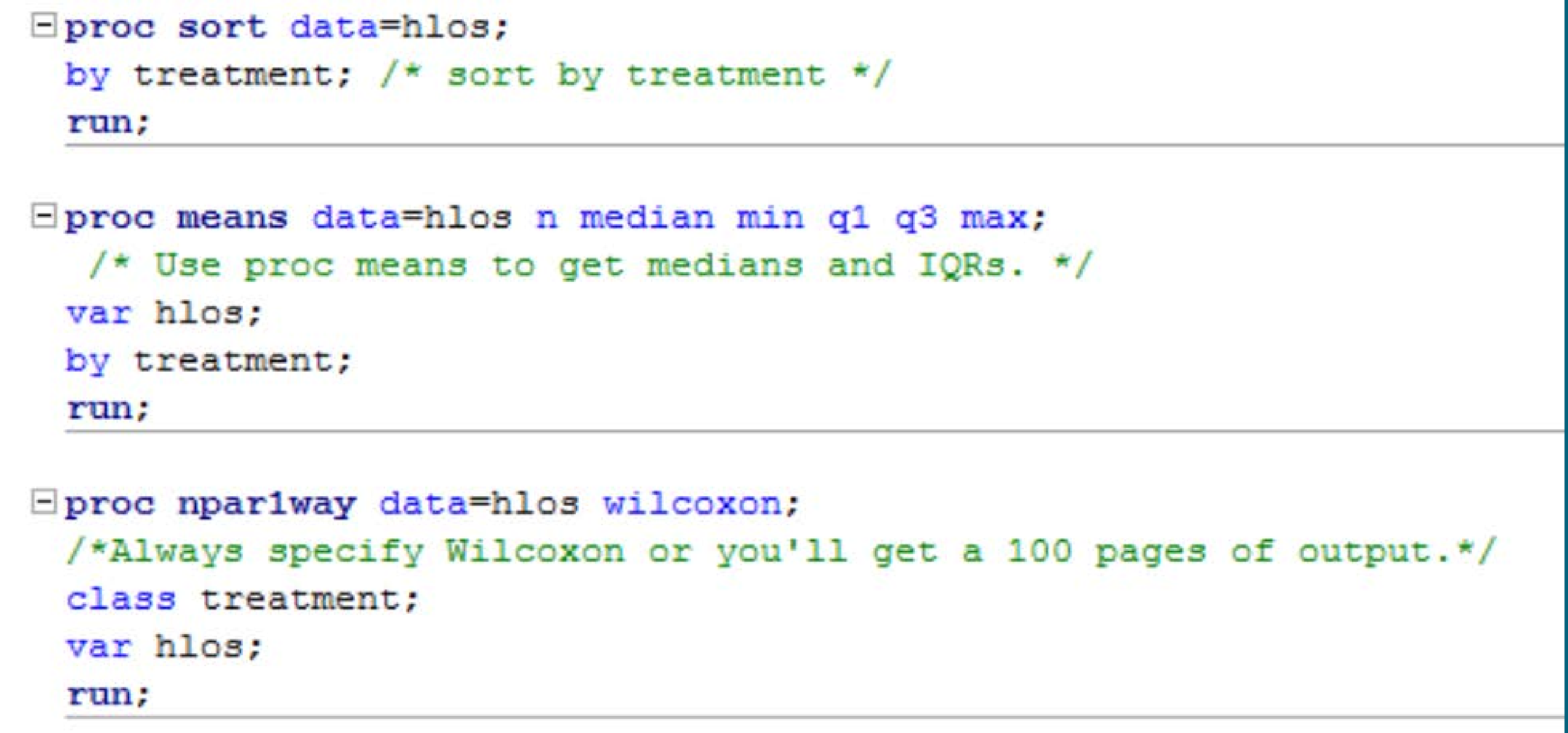
SAS Output: Log transformed



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

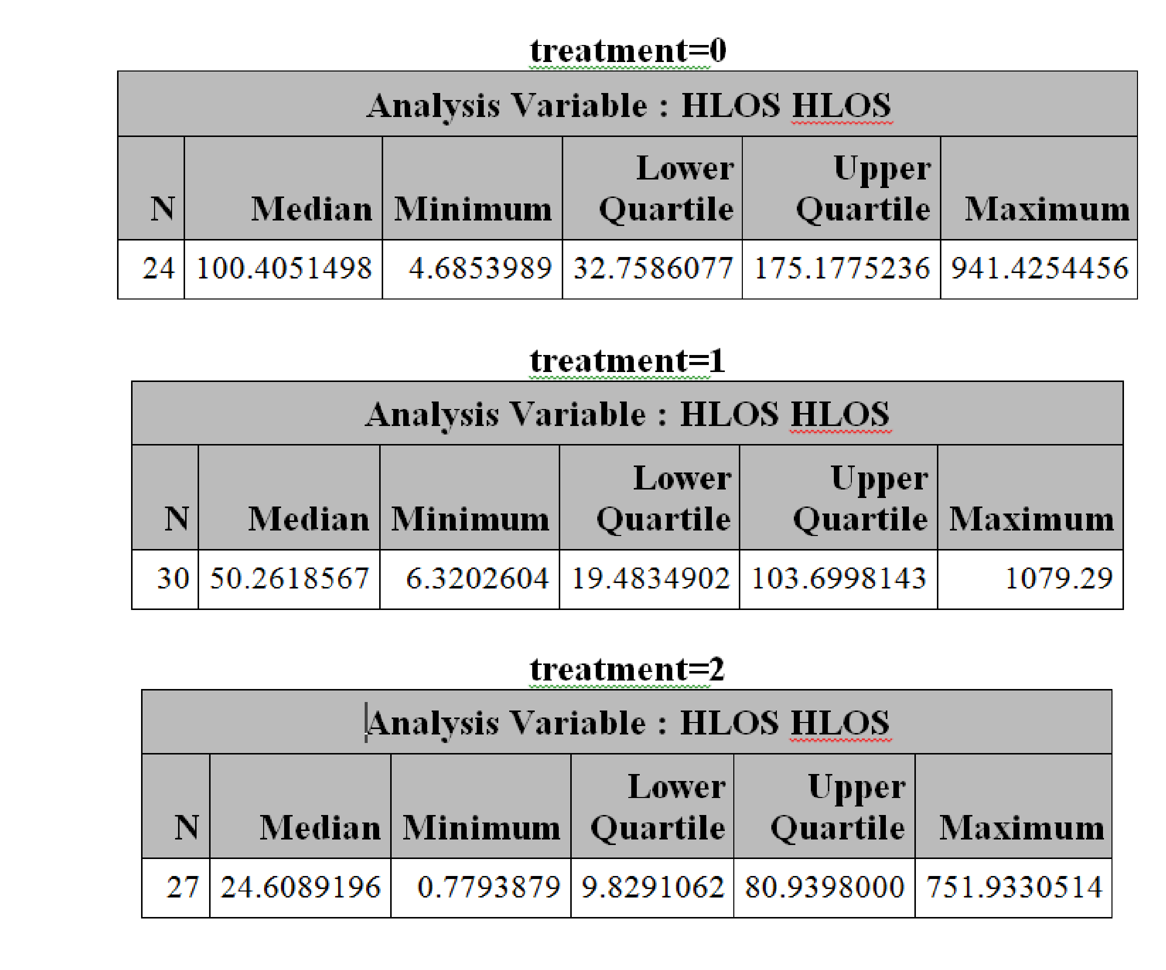


```
proc sort data=hlos;
by treatment; /* sort by treatment */
run;

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

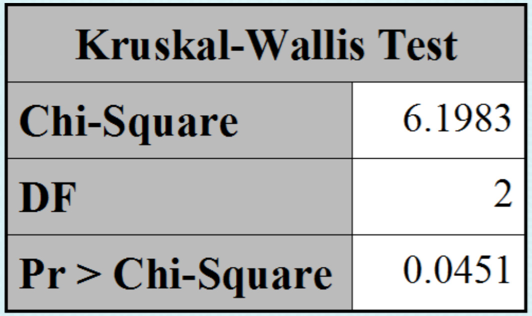
Analysis Variable : HLOS HLOS					
N	Median	Minimum	Lower Quartile	Upper Quartile	Maximum
24	100.4051498	4.6853989	32.7586077	175.1775236	941.4254456

treatment=1

Analysis Variable : HLOS HLOS					
N	Median	Minimum	Lower Quartile	Upper Quartile	Maximum
30	50.2618567	6.3202604	19.4834902	103.6998143	1079.29

treatment=2

Analysis Variable : HLOS HLOS					
N	Median	Minimum	Lower Quartile	Upper Quartile	Maximum
27	24.6089196	0.7793879	9.8291062	80.9398000	751.9330514

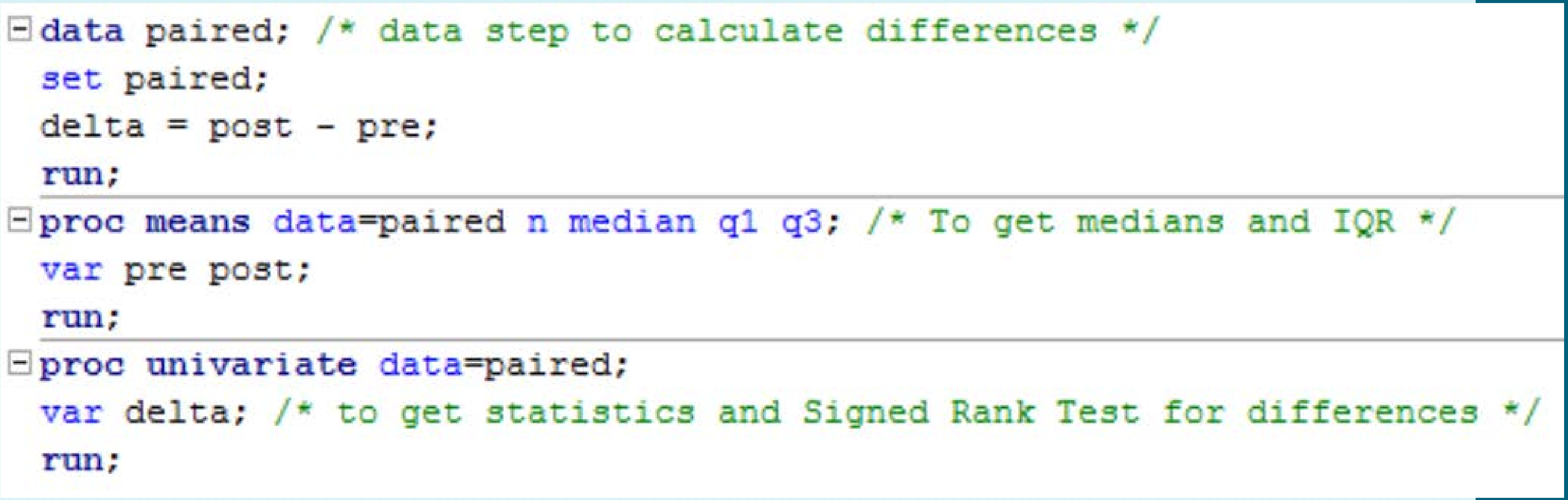


Kruskal-Wallis Test	
Chi-Square	6.1983
DF	2
Pr > Chi-Square	0.0451

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



```
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

Variable	Label	N	Median	Lower Quartile	Upper Quartile
pre	pre	20	3.5000000	2.0000000	4.0000000
post	post	20	4.0000000	3.0000000	5.0000000

Basic Statistical Measures			
Location		Variability	
Mean	1.100000	Std Deviation	1.07115
Median	1.000000	Variance	1.14737
Mode	1.000000	Range	4.00000
		Interquartile Range	2.00000

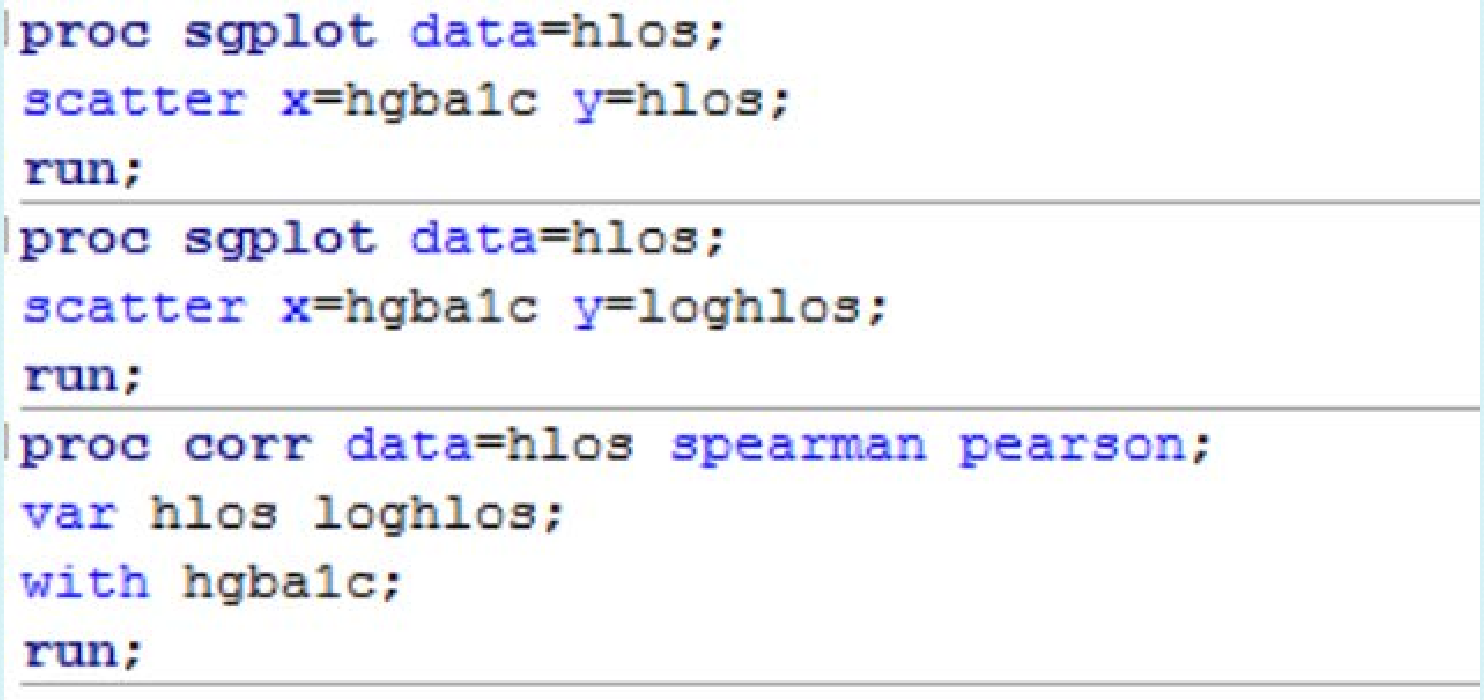
Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	4.592575	Pr > |t|	0.0002
Sign	M	6.5	Pr >= |M|	0.0010
Signed Rank	S	55.5	Pr >= |S|	0.0005

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

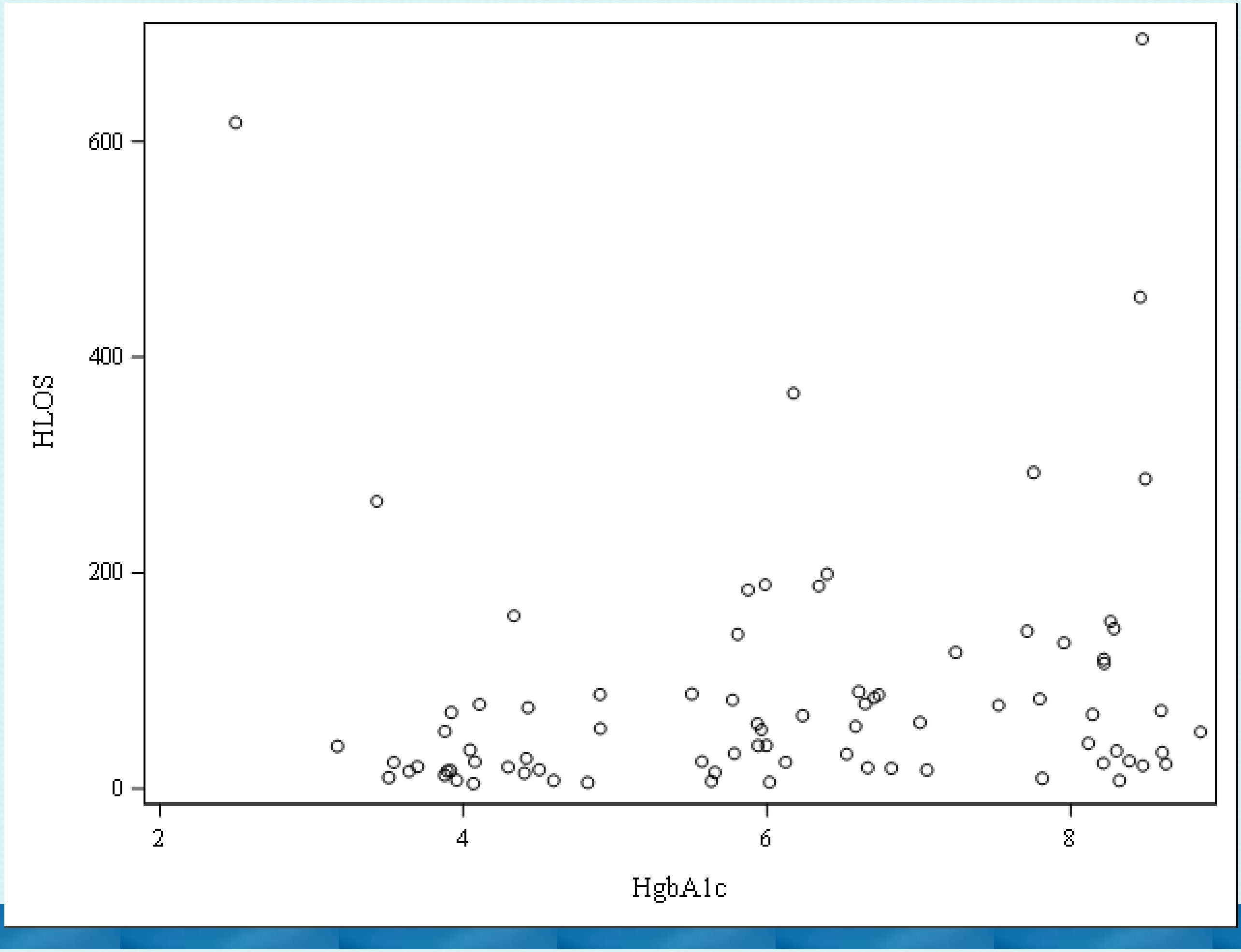
- SAS Code



```
proc sgplot data=hlos;
scatter x=hgba1c y=hlos;
run;
proc sgplot data=hlos;
scatter x=hgba1c y=loghlos;
run;
proc corr data=hlos spearman pearson;
var hlos loghlos;
with hgba1c;
run;
```

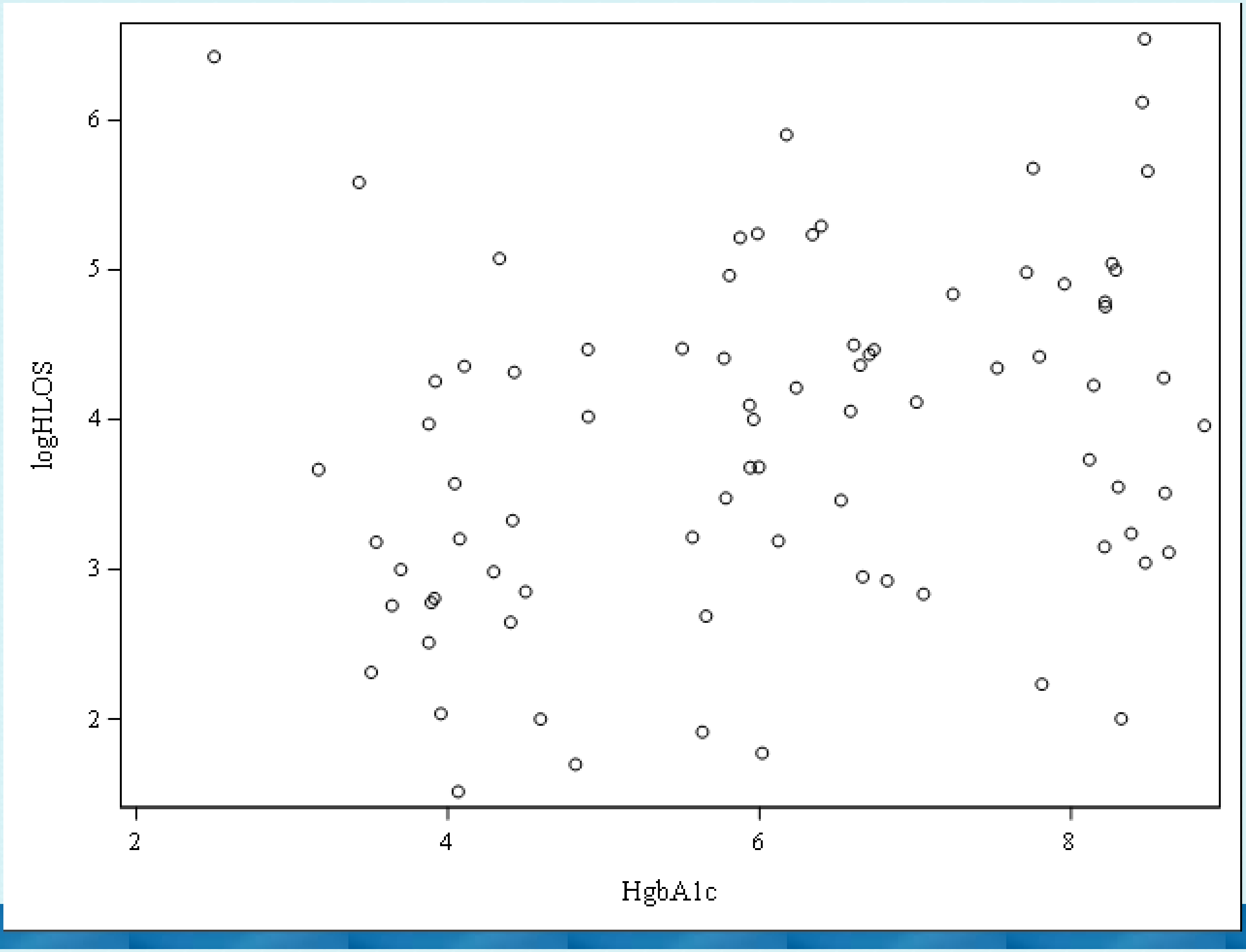
Non-Normal Correlation

- Scatter plot of HLOS by Hgb A1c



Non-Normal Correlation

- Scatter plot of log(HLOS) by Hgb A1c



Non-Normal Correlation

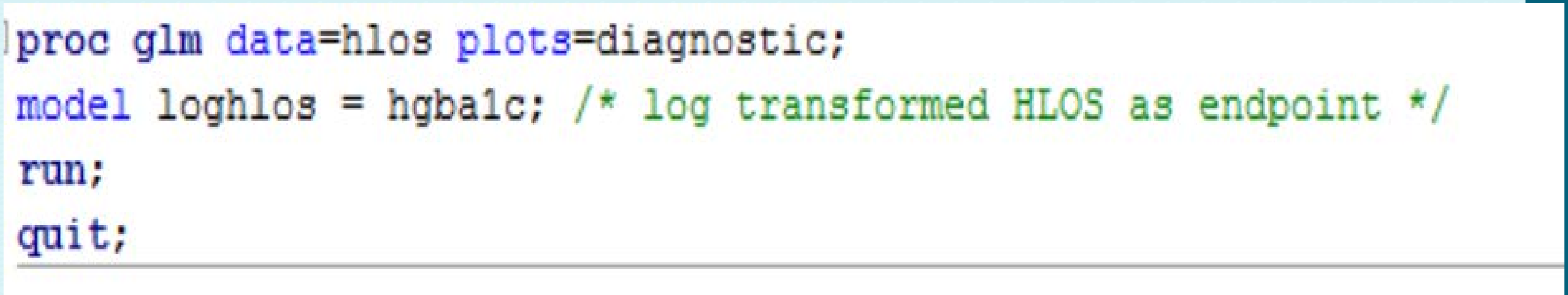
- SAS Output

Pearson Correlation Coefficients, N = 81 Prob > |r| under H0: Rho=0		
	HLOS	**logHLOS**
HgbA1c	0.14611	0.26932
HgbA1c	0.1931	0.0150

Spearman Correlation Coefficients, N = 81 Prob > |r| under H0: Rho=0		
	HLOS	**logHLOS**
HgbA1c	0.28485	0.28485
HgbA1c	0.0100	0.0100

Non-Normal: Regression

- SAS code for the transformed HLOS



```
proc glm data=hlos plots=diagnostic;
model loghlos = hgba1c; /* log transformed HLOS as endpoint */
run;
quit;
```

Non-normal Regression

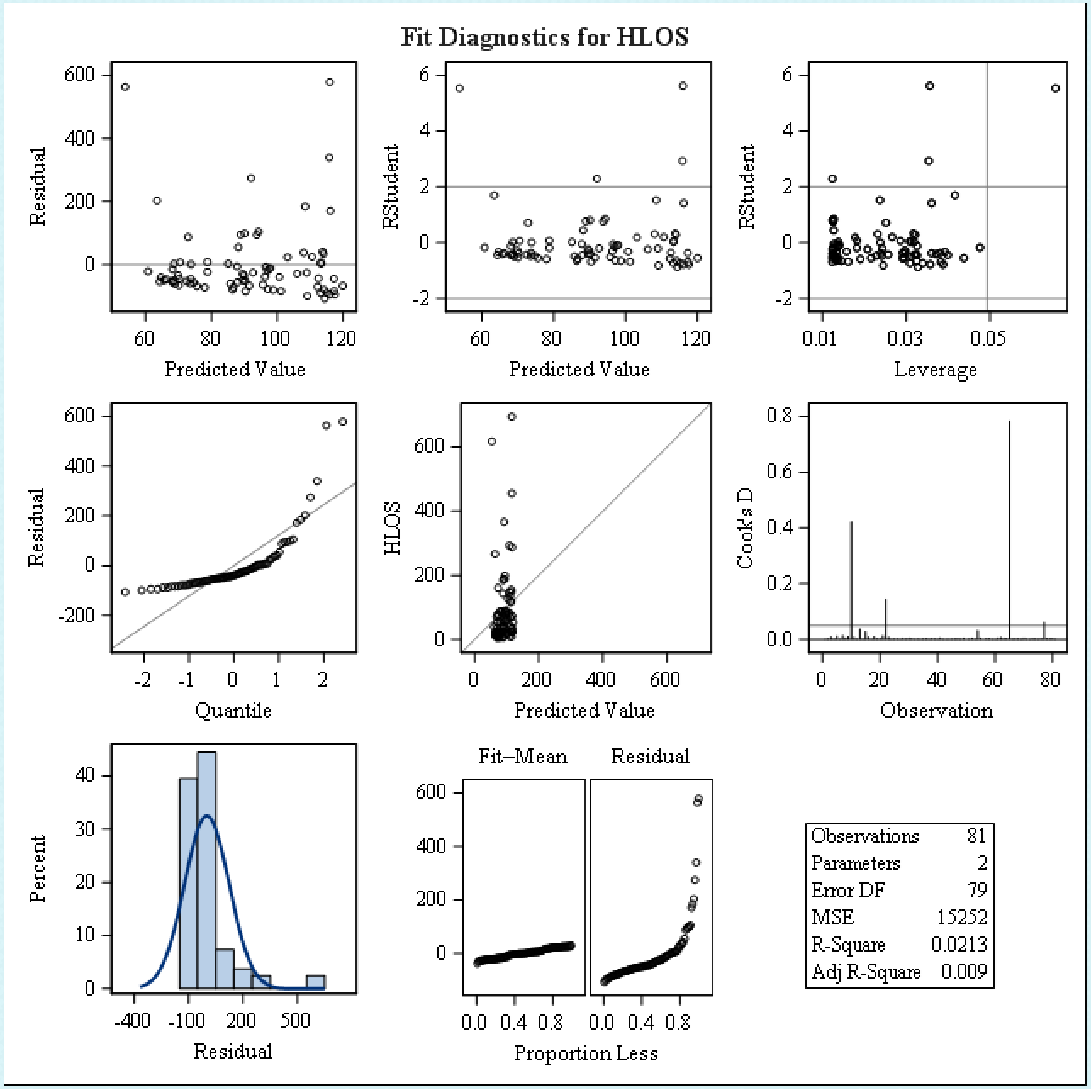
- Regression Results for Raw HLOS

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

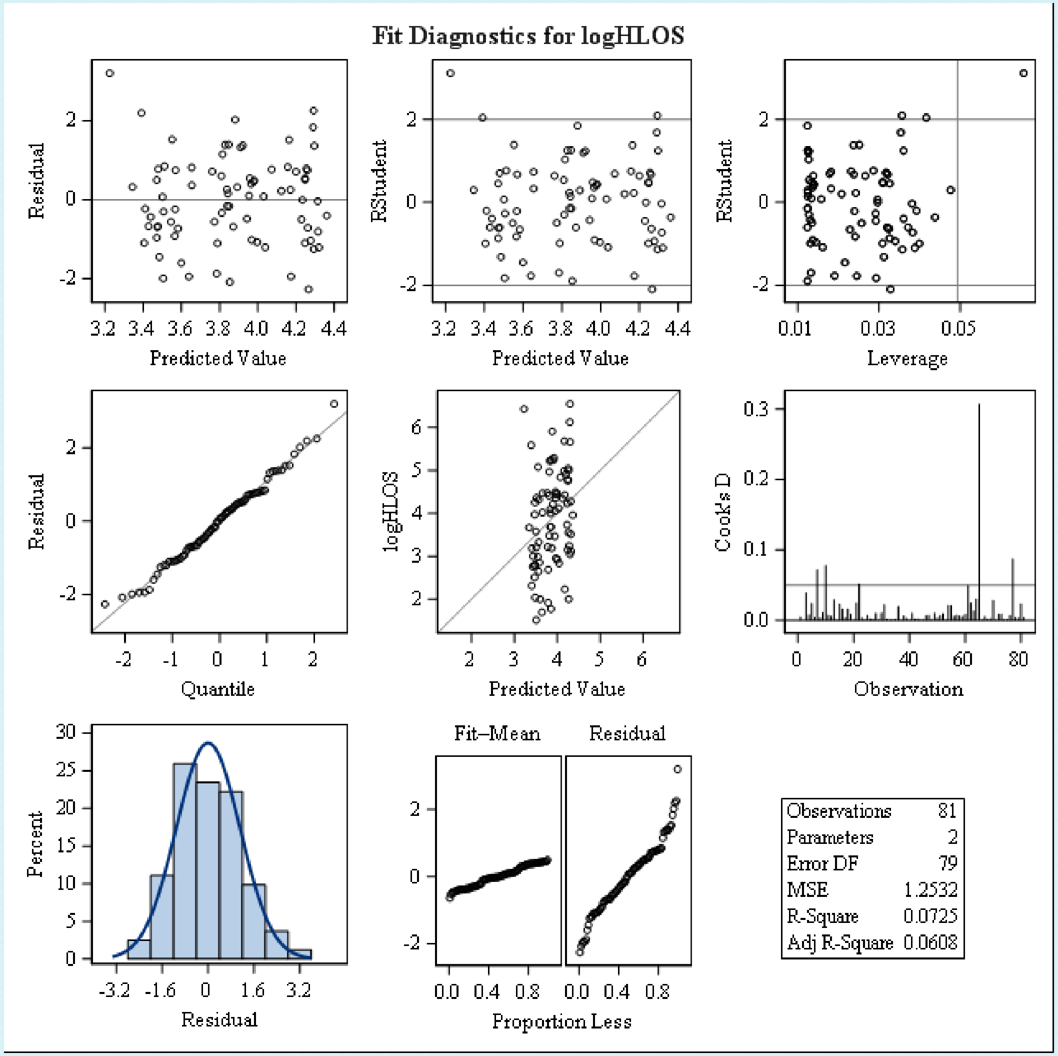
Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	2.778096911	0.45560075	6.10	<.0001
HgbA1c	0.178743597	0.07191073	2.49	0.0150

Diagnostic Plots: Raw HLOS



Observations	81
Parameters	2
Error DF	79
MSE	15252
R-Square	0.0213
Adj R-Square	0.009

Diagnostic Plots: Log HLOS



Fit Diagnostics for logHLOS

Observations	81
Parameters	2
Error DF	79
MSE	1.2532
R-Square	0.0725
Adj R-Square	0.0608

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/
  - MIND IDDRC - www.ucdmc.ucdavis.edu/mindinstitute/centers/iddrc/cores/bbrd.html
  - Cancer Center
  - https://health.ucdavis.edu/cancer/research/sharedresources/biostatistics.html
  - EHS Center - https://environmentalhealth.ucdavis.edu/core-resources

References

- Fayers, Peter (2011) “Alphas, Betas, and Skewy Distributions: two ways of getting the wrong answer, Adv Health Sci Edu, 16: 291-296
- Biostatistics for the Clinician, URL: https://www.uth.tmc.edu/uth_orgs/educ_dev/oser/L3_0.HTM