

## Topic 8: Transformations of the data [S&T Ch. 9.16]

### 8. 1. The assumptions of ANOVA

#### 8. 1. 1. Additive effects

The treatment ( $\tau_i$ ), block ( $\beta_j$ ), **and** error terms  $\varepsilon_{ij}$  are *added*

"The treatment effects remain constant over blocks (or replications) and the block effects remain constant over treatments."

Tukey's test for RCBD with one observation per cell (Topic 6.4.1)

Uses **1 df** from the **error** to test if the **multiplicative effects** are significantly larger than the rest of the error.

#### 8. 1. 2. Independence of errors

This says that the values of the  $\varepsilon_{ij}$  are statistically independent. Failure of this assumption is often caused by a failure to properly randomize the plots:

Rep	I	II	III
1	A	2 A	3 A
4	B	5 B	6 B
7	C	8 C	9 C

- The self-similarity of experimental units adjacent in space or time is called **positive autocorrelation**. Regular alternation of positive and negative errors is a manifestation of **negative autocorrelation**.
- Independence of errors in a sequence of continuous variates may be tested using a test based on the differences between adjacent values (Biometry p394 – 395).
- However, the physical process of randomly allocating the treatments to the experimental units ensures that the  $\varepsilon_{ij}$  will be independent.

#### 8. 1. 3. Normally distributed errors

Violation of normality means that the  $\varepsilon_{ij}$  are not normally distributed.

This is the least influential assumption on the  $F$  test.

Normality can be checked using

- A plot of the residuals produced using the OUTPUT statement
- **Shapiro-Wilk** (ST&Dp.567) using PROC UNIVARIATE normal
  - Reject if  $W$  very different from 1, and  $P < 0.05$

### 8. 1. 4. Homogeneity of Variance

- This is usually the most important assumption.
- Lack of homogeneity may be due to
  - Variances depending on the treatment mean
  - Unequal variances with no apparent relation to treatment means.

*Example* of the effect of lack of variance homogeneity:

Treatment	Replicate					Total	Mean	s <sup>2</sup>
	1	2	3	4	5			
A	3	1	5	4	2	15	3	2.5
B	6	8	7	4	5	30	6	2.5
C	12	6	9	3	15	45	9	22.5
D	20	14	11	17	8	70	14	22.5

The *F* statistic of the ANOVA is significant. *LSD* for  $\alpha$  0.05 = 4.74.

Difference between means A and B = not significant

Difference between means C and D = significant.

Analyzing A and B separately from C and D yields the opposite result.

Source of variation	df	SS	MS	F
Treatments A B	1	22.5	22.5	9*
Error	8	20.0	2.5	

Source of variation	df	SS	MS	F
Treatments C D	1	62.5	62.5	2.8 NS
Error	8	180	22.5	

The difference between A and B is significant while that between C and D is not, when analyzed separately. The reason is that the variance for treatments C and D is much larger than for A and B.

Moderate heterogeneity of variances may not affect too much the overall test of significance but **may have a huge effect on the significance of mean comparisons**

Among the different tests for homogeneity of variances, **Bartlett's** (S&T, p. 471) and **Levene's** test are widely used.

## 8. 2. Transformations

If the ANOVA assumptions cannot be maintained:

- 1.- Carry out a different test not requiring the rejected assumptions, such as:
  - Non-parametric test
  - Variance-weighted Welch's ANOVA: **means** variable / **Welch**;
- 2.- **Transform the data** so the variable to be analyzed meet the ANOVA assumptions. Then perform the analysis on the transformed variables.

When a statistical test may be made significant after transformation of a set of data, people may feel suspicious...

### What is the justification for transforming the data?

There is really no scientific necessity to employ the common **linear or arithmetic scale** to which we are accustomed. If a relationship is **multiplicative on a linear scale**, it may make much more sense to think of it as **additive on a logarithmic scale**. Examples:

- **Logarithms**: pH values
- **Square roots**: The square root of the surface area of an organism may be a more appropriate measure of the biological variable subjected to physiological and evolutionary forces than is the area.
- **Reciprocals**: microbiological titrations.

Since the scale of measure is arbitrary is accepted it is valid to transform the data and **select** the transformation that most **closely satisfies the assumptions** of the ANOVA.

Transformations very often improves simultaneously several departures from the assumptions of ANOVA.

Four transformations will be discussed:

- the **logarithmic** transformation,
- the **square root** transformation,
- the **angular** or arcsine transformation
- the **power** transformation.

## 8. 2. 1 The log transformation

Criteria for deciding on the log transformation:

- **Standard deviations** (not the variances) of samples are roughly **proportional to the means**
- **Multiplicative** rather than additive effects (significant Tukey's test).
- **Skewed** frequency distributions to the right

Data with negative values cannot be transformed using logarithms. In cases with 0s, a 1 must be added to each data point before transforming.

### Effects of the log transformation

The dependent variable is weight in pounds of vitamin-treated and control animals, in a RCBD.

Species— Treatment	Block				Total	Mean	s <sub>i</sub>	M/s <sub>i</sub>
	I	II	III	IV				
Mice—control	0.18	0.30	0.28	0.44	1.2	0.3	0.11	
Mice—vitamin	0.32	0.40	0.42	0.46	1.6	0.4	0.06	
Subtotals	<b>0.50</b>	<b>0.70</b>	<b>0.70</b>	<b>0.90</b>	<b>2.8</b>	<b>0.35</b>	<b>0.08</b>	<b>4.4</b>
Chickens—control	2.0	3.0	1.8	2.8	9.6	2.40	0.58	
Chickens—vitamin	2.5	3.3	2.5	3.3	11.6	2.90	0.46	
Subtotals	<b>4.5</b>	<b>6.3</b>	<b>4.3</b>	<b>6.1</b>	<b>21.2</b>	<b>2.65</b>	<b>0.52</b>	<b>5.1</b>
Sheep—control	108.0	140.0	135.0	165.0	548.0	137.0	23.3	
Sheep—vitamin	127.0	153.0	148.0	176.0	604.0	151.0	20.6	
Subtotals	<b>235.0</b>	<b>293.0</b>	<b>283.0</b>	<b>341.0</b>	<b>1152.0</b>	<b>144.0</b>	<b>22.0</b>	<b>6.5</b>

Means and s<sub>i</sub> are proportional

Vitamin effect: **P= 0.30**

Vitamin x Species: **P= 0.39**

### Problems of the original data suggest something wrong with the ANOVA

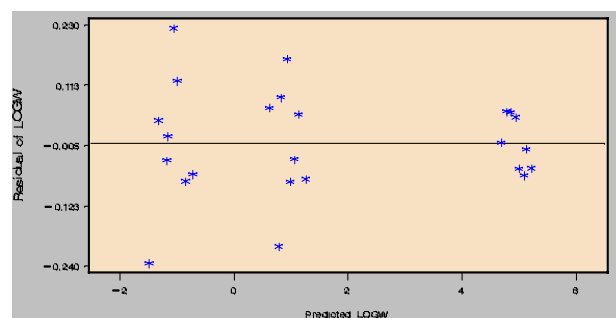
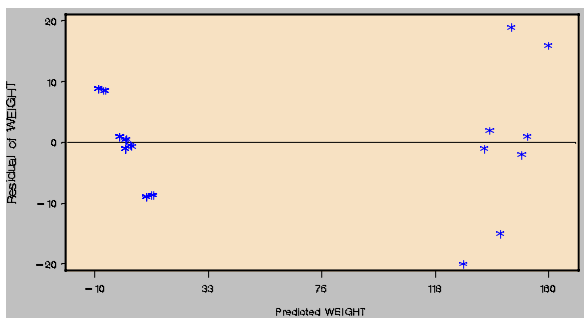
- No significant differences were detected between vitamin and control though every animal in every replicate receiving the vitamin showed a greater weight than the corresponding control animal.
- No significant differences were detected for the interaction between vitamin effects and species though the apparent response to vitamins is so different in the different species.

## Comparison of transformed and the original data.

	Original data	Transformed data
Shapiro-Wilks test ~N or res.	W= 0.95, p= 0.32	W= 0.97, p= 0.56
Tuckey's non-additivity test	F= 545, <b>p= 0.0001</b>	F= 1.74, <b>p= 0.21</b>
Levene's test ( $s^2$ homogeneity)	F= 2.5, <b>p= 0.07</b>	F= 1.78, <b>p= 0.17</b>
Species effect	F= 434.5, p= 0.0001	F= 4883.0, p= 0.0001
Vitamin effect	F= 1.14, <b>p= 0.30</b>	F= 16.62, <b>p= 0.011</b>
Species * Vitamin	F= 1.0, p= 0.39	F= 1.58, <b>p= 0.24</b>

Significant Tukey and Levene's test demonstrate that there is no additivity and that there is a significant heterogeneity of variances.

**The residuals confirm the violations of the ANOVA assumptions.**



Residuals vs. predicted. **Original data**

Residual vs. predicted **Log10(X)**

**Logarithm transformation** suggested by

- Multiplicative effects
- Rough proportionality between standard deviations and means

After logarithm transformation the assumptions of ANOVA are satisfied. The new analysis is valid and the P values are correct

## Problems of interpretation

- Lineal relationship in transformed data, not necessarily linear in the original
- The interaction in original data: “Does the amount of change in weight due to the addition of vitamins vary from species to species?”
- The interaction with log transformed data: “Does the **proportion or percent** change in weight due to vitamins vary from species to species?”

## Problems with interpretation of interactions in log transformed data

### Additive **log transformed** data

Log (Y)	No A	A	
No B	2.30	2.47	2.47 - 2.30 = 0.17
B	2.60	2.77	2.77 - 2.65 = 0.17
	2.60 - 2.30 = 0.30	2.77 - 2.47 = 0.30	<b>No interaction</b>

### Original data

Original	No A	A	
No B	200	300	A increases 50%
B	400	600	A increases 50%
	B increases 100%	B increases 100%	

**A non-significant interaction of log transformed data indicates a constant % difference in the original data!**

## Detransforming means calculated from transformed data

### Logarithmic transformation (ST&D 242)

Influence of a multiplicative treatment effect on the variance

	Control					Mean	Variance
Y	20	40	50	60	80	50	500
ln(Y)	2.9957	3.6889	3.9120	4.0943	4.3820	3.8146	0.2740
<b>Treatment = control + 20</b>							
Y	40	60	70	80	100	70	500
ln(Y)	3.6889	4.0943	4.2485	4.3820	4.6052	4.2038	0.1180
<b>Treatment = control * 1.5</b>							
Y	30	60	75	90	120	75	1125
ln(Y)	3.4012	4.0943	4.3175	4.4998	4.7875	4.2201	0.2740

Detransforming means of the ln to the original scale = **Geometric means**

$e^{3.8146} = 45.3586$  and NOT the arithmetic mean = 50

**Geometric mean G** =  $(20 \cdot 40 \cdot 50 \cdot 60 \cdot 80)^{1/5} = 45.3586$

$G = (Y_1 \cdot Y_2 \cdot \dots \cdot Y_n)^{1/n}$  and that  $\log G = (1/n) \sum \log Y_i$

### 8. 2. 2. The square root transformation

- Counts of rare events -insects on a leaf or blood cells in a hematocytometer.
- Any data with a **Poisson distribution**.
- Data with **variances roughly proportional to the means**. This violates the assumption that the variances and means are not correlated.

Data of this kind can be made more nearly normal and variances relatively independent of the means by transforming them to square roots.

Actually, it is better to use:

$$\sqrt{Y + \frac{1}{2}}$$

if there are counts under 10 [SAS: SQRT(variable+0.5)].

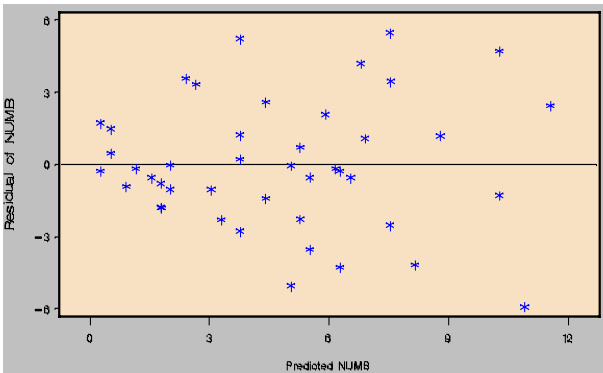
**Example.** Number of lygus per 50 sweeps. In each plot of an experiment testing 10 insecticides and a check treatment, replicated four times in a RCBD

Treatment	Block				Total	Mean	s <sub>i</sub> <sup>2</sup>
	I	II	III	IV			
A	7	5	4	1	17	4.25	6.25
B	6	1	2	1	10	2.50	5.67
C	6	2	1	0	9	2.25	6.92
D	0	1	2	0	3	0.75	0.92
E	1	0	1	2	4	1.00	0.67
F	5	14	9	15	43	10.75	21.58
G	8	6	3	6	23	5.75	4.25
H	3	0	5	9	17	4.25	14.25
I	4	10	13	5	32	8.00	18.00
J	6	11	5	2	24	6.00	14.00
K	8	11	2	6	27	6.75	14.25

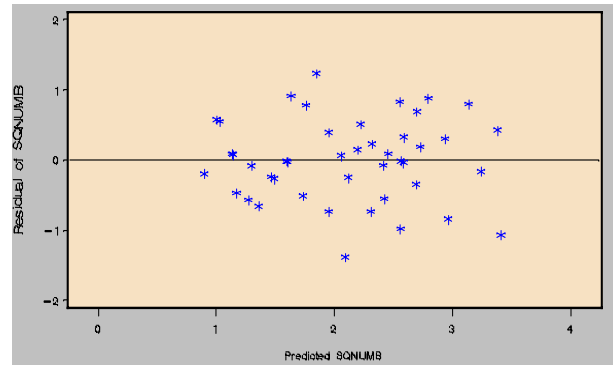
## ANOVA

	Original data	Transformed data
<b>Shapiro-Wilks W resid.</b>	W= 0.98, p= <b>0.73</b>	W= 0.987, p= <b>0.95</b>
<b>Tukey's non-additivity</b>	F= 0.63, p= 0.44	F= 0.07, p= 0.80
<b>Levene's test</b>	F= 1.61, p= 0.15	F= 0.90, p= 0.55
<b>Treatment effect</b>	F= 3.7, <b>p= 0.0026</b>	F= 4.04, <b>p= 0.0014</b>

The two analyses are not very different, since they both show a highly significant treatment effect. **The F value is about 10% higher after transformation.**



**Original. Residuals vs. Predicted**  
Correlation mean-var. **R=0.89 \*\***



**Transformed. Residuals vs. Predicted**  
Correlation mean-var. **R=0.37 NS**

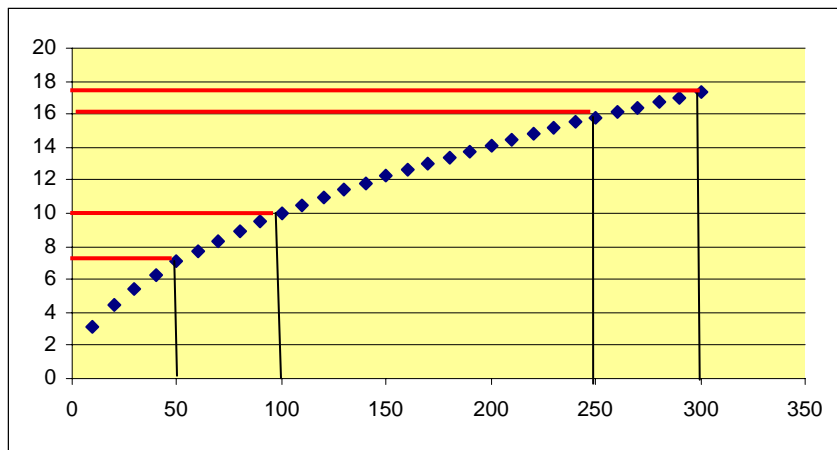
- **Heterogeneity of variances can produce differences in mean separation.**
- Transformed data: G and D, G and E, J and D, and J and E significantly different, whereas they were not in the raw data.

D	E	C	B	H	A	G	J	K	I	F
0.75	1.00	2.25	2.50	4.25	4.25	5.75	6.00	6.75	8.00	10.75
<b>ORIGINAL</b>										
_____										
_____										

D	E	C	B	H	A	G	J	K	I	F
0.62	0.89	1.81	2.22	3.50	3.95	5.50	5.60	6.31	7.57	10.32
<b>SQRT</b>										
_____										
_____										

## Interpretation of detransformed means in square root transformation

- **Weighted means** are obtained by “**detransforming**” the means of the transformed data.
- This is done by squaring the transformed means and subtracting one-half.
- These means are  $<$  than those from raw data because **more weight is given to smaller numbers**.
- Correct! Smaller numbers are measured with less sampling error than the larger ones.
- **The general effect of the square root transformation is to increase the precision with which we measure the differences between small means.**



This is highly desirable in insect control work, since we are generally not as interested in differences between two relatively ineffective treatments as we are in comparing treatments that give good control.

- The assumption of independence of means and variances was violated in the original data, and this was remedied by the transformation. Transformation reduced the amount of heterogeneity over that in the raw data.

**Data requiring the square root transformation do not violate the assumptions of the analysis of variance nearly as drastically as data requiring a log transformation. Consequently, changes in the analysis by the transformation are not so spectacular.**

### 8. 2. 3. The arcsine or angular transformation

Criteria for deciding on the angular transformation:

- Counts expressed as percentages or proportions of the total sample.
- Any other data with *binomial distribution* rather than a normal.

In binomial data, variances tend to be small at the two ends of the range of values (close to zero and 100%), but larger in the middle (around 50%).

The appropriate transformation for data of this kind is obtained by finding: e

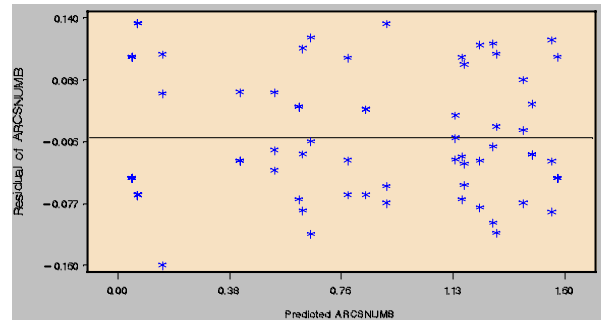
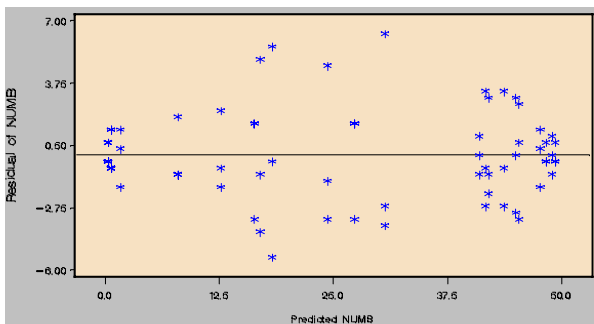
**arcsine(Y)** or  $\text{sine}^{-1}(Y)$  [in SAS **ARSIN(SQRT(proportion))**].

**Example:** Number of lettuce seeds germinating in samples of 50. CRD lettuce seed with 24 treatments, 3 replications. Treatments are arranged in order of magnitude.

Treatment	Blocks			Mean	s <sub>i</sub> <sup>2</sup>
	1	2	3		
1	0	0	1	0.33	0.33
2	0	1	0	0.33	0.33
3	0	0	1	0.33	0.33
4	0	2	0	0.67	1.33
5	2	0	0	0.67	1.33
6	0	2	3	1.67	2.33
7	7	10	7	8.00	3.00
8	11	12	15	12.67	4.33
9	13	18	18	16.33	8.33
10	22	16	13	17.00	21.00
11	24	13	18	18.33	30.33
12	23	21	29	24.33	17.33
13	24	29	29	27.33	8.33
14	37	28	27	30.67	30.33
15	42	41	40	41.00	1.00
16	39	41	45	41.67	9.33
17	41	45	40	42.00	7.00
18	47	41	43	43.67	9.33
19	45	42	48	45.00	9.00
20	46	42	48	45.33	9.33
21	49	46	48	47.67	2.33
22	48	49	48	48.33	0.33
23	50	49	48	49.00	1.00
24	49	49	50	49.33	0.33

Note that there is a strong tendency for the variances at the extremes to be smaller than those in the middle of the range.

	Original data	Transformed data
Levene's test	F= 2.43, <b>p= 0.0048</b>	F=0.99, <b>p= 0.49</b>
Treatment effect	F= 148.12, p= 0.0001	F= 100.14, p=0.0001



Residuals vs. predicted. **Original data**    Residual vs. pred. $\text{SIN}^{-1}(\text{SQRT}(\text{original} \cdot 2/100))$

The pattern of the residuals observable in the raw data is no longer apparent in the transferred data.

An analysis of variance of the transferred data does not seem to lead us to a different conclusion than the analysis of the raw data:

However, there are several **differences in mean separation**.

A Duncan's multiple range test shows that:

1. Five differences were declared significant before transformation but not after: 7–8, 8–11, 10–12, 11–12 and 12–14.
2. Five differences were declared significant after transformation but not before: 18–22, 19–23, 19–24, 20–23, and 20–24.

Which set of conclusions should we accept? The answer is simple: we should accept the conclusions based on the more valid analysis, in this case, the analysis of the transformed data.

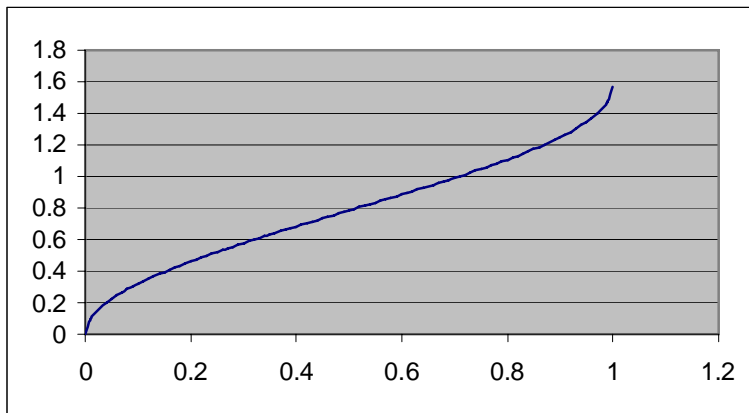
**We do not transform data to give us results more to our liking.**  
**We transform data so that the analysis will be *valid* and the conclusions and P values *correct*.**

## ARCSINE (SQRT(X))

The arcsine transformation finds  $Y = \arcsin(\sqrt{p})$  where **p is a proportion (0 to 1)**.

The term arcsine is synonymous with inverse sine or  $\text{sine}^{-1}$ .

ARCSINE: Returns the SINE of a number in radians in the range  $-\pi/2$  to  $+\pi/2$



**This transformation will spread the values at both ends of the distribution compared with the central part.**

When the percentages in the original data fall between 30 and 70%, it is generally not necessary to apply the arcsine transformation

**8. 2. 4. The power transformation** (Hinz, P. N. and H. A. Eagles. 1976. Crop Science 16: 280-283.)

**Field experiments** are often conducted using replicated trials over a broad range of locations and environmental conditions. **Often the means and the residual variances differ markedly among environments.**

The choice of an optimal transformation from the many possible alternatives is not always obvious especially if the functional relationship between variance and mean is not known.

The **power transformation** method provides a mean of selecting a transformation from a broad class of power transformations and employs the data themselves to estimate the exponent used in transforming the original measures.

The power transformation is to **transform X to Y empirically by “a”**,

$Y = X^a$	if $a \neq 0$
$Y = \log X$	if $a = 0$

Note that

$Y = \sqrt{X}$	if $a = 1/2$
$Y = \log X$	if $a = 0$
$Y = 1 / X$	if $a = -1$

- Generally (1) If variances  $\uparrow$  as means  $\uparrow$ , use  $a < 1$   
 (2) If variances  $\uparrow$  as means  $\downarrow$ , use  $a > 1$ .

The exact value of the power of the transformation ‘a’ can be estimated by obtaining the slope ‘b’ of the **regression** of  $\log s_i^2$  on  $\log \bar{x}_i$ , and then solving for ‘a’= 1-b/2

Use any program to calculate a lineal regression between the logarithms of the observed variances on the logarithms of the means. Solve the equation and find the least squared estimate (b) of  $\beta$ .

Then  $a = 1 - b / 2$ , which can be used to transform X to Y.

**Examples of power transformation**

Using means and variances from Table **Topic 8.2.2.**

$$\log s_i^2 = a + b \log \bar{x}_i \qquad \log s_i^2 = 0.1086 + 1.238 \bar{x}$$

$$b = 1.238 \qquad \text{and} \qquad a = 1 - (1.238 / 2) = 0.38$$

Using  $Y = X^{0.38}$ , Levene’s test is non significant (p=0.32)

## Calculation of the power transformation

```
Data Power;  
Input mean variance;  
    logmean=log10(mean);  
    logvar= log10(variance);  
Cards;  
274.2    6205.7  
153.0    1862.5  
 99.8     445.7  
237.8    2359.2  
151.2    1739.7  
 82.2     182.2  
;  
Proc reg;  
    Model logvar= logmean;  
Run; quit;
```

Source	DF	Squares	Square	F Value	Pr > F
Model	1	1.38674	1.38674	44.63	0.0026
Error	4	0.12429	0.03107		
Total	5	1.51103			
Root MSE		0.17628	R-Square	0.9177	
Dependent Mean		3.09762	Adj R-Sq	0.8972	
Coeff Var		5.69068			

### Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	-2.53529	0.84626	-3.00	0.0401
Logmean	1	<b>2.58143</b>	0.38642	6.68	0.0026

To calculate “a” and transform the data

- Find the slope **b= 2.58**
- Calculate **a= 1-(b/2)= -0.29**
- Add line to original SAS **trnsf\_lygus= lygus<sup>-0.29</sup>**
- Test if the transformed data satisfy the assumptions