Lecture 11
Topic 8: Data Transformations

Assumptions of the Analysis of Variance

1. Independence of errors

The ε_{ij} (i.e. the *errors* or *residuals*) are statistically independent from one another.
Failure to meet this assumption is often the result of failing to randomize properly.

One ordinarily does not need to do more than a true randomization to satisfy this requirement because the process of randomly allocating the treatments to the experimental units usually ensures that the ε_{ij} will be independent.

2. Normal distribution of errors

The ε_{ij} (i.e. the *errors* or *residuals*) are normally distributed.
This assumption has the least influence on the F test.

For small to medium sample sizes, check normality using the Shapiro-Wilk test. For sample sizes ≥ 2000, the Kolmogorov-Smirnov statistic is recommended (R's ks.test() function).

3. Homogeneity of Variance

The average magnitude of the ε_{ij} (i.e. the *errors* or *residuals*) is the same for all treatments.
This assumption has the most influence on the F test.

MSE is a pooled variance

Two basic scenarios when this assumption is not satisfied:

1. Variances depend on the treatment means.
2. Variances are unequal but exhibit no apparent relation to the treatment means.

Example: Effect of the lack of variance homogeneity (homoscedasticity).

	Replicate							
Treatment	**1**	**2**	**3**	**4**	**5**	**Total**	**Mean**	**s^2**
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	21.5	15.5	12.5	18.5	9.5	77.5	15.5	22.5

The result of the ANOVA:

Source of variation	**df**	**SS**	**MS**	**F**
Treatments	3	428.4	142.8	11.4 ***
Error	16	200	12.5	

The Tukey minimum significant difference ($\alpha = 0.05$) is 6.3974.

A versus B = NS
C versus D = Significant

When analyzed *separately*, however, (i.e. with two separate ANOVAs):

Source of variation	**df**	**SS**	**MS**	**F**	**p**
Treatments A B	1	22.5	22.5	9.00 *	0.02
Error	8	20.0	2.5		
Source of variation	**df**	**SS**	**MS**	**F**	
Treatments C D	1	105.6	105.6	4.69 NS	0.06
Error	8	180	22.5		

A versus B = Significant
C versus D = NS

Moderate heterogeneity of variances may not affect the overall test of significance but may have a huge effect on single df comparisons.

Among the different tests for homogeneity of variances, Bartlett's Test and Levene's Test are the most widely used.

4. Additive effects (blocked designs with one rep per cell)

Block and treatment effects are additive.
Violation of this assumption indicates potential issues with your blocking variable.

The treatment effects are constant across blocks and the block effects are constant across treatments.

Use 1 df from the error to perform Tukey's 1-df Test for Nonadditivity to determine if the non-additive effects are significantly larger than the true experimental error.

Transformations

Your data do not meet the assumptions of the analysis. What to do?

1. Carry out a different analysis which does not require the rejected assumptions:
    a. Non-parametric tests (e.g. Durbin Test; Friedman RCBD Rank-Sum Test)
    b. Variance-weighted Welch's one-way ANOVA [R's oneway.test() function]
2. Transform the data

What is the justification for transforming data?

The common linear (or arithmetic) scale is arbitrary.

Logarithms:	pH values
Square roots:	Surface areas of organisms
Reciprocals:	Microbiological titrations

Since *the scale of measurement is arbitrary*, it is valid to transform a variable into a scale that best satisfies the assumptions of the analysis.

Transformations very often correct for several departures from the ANOVA assumptions simultaneously.

1. The logarithmic transformation
2. The square root transformation
3. The angular or arcsine transformation
4. The power transformation.

The log transformation

General indicators that a log transformation may be appropriate:

1. The **standard deviations** are roughly proportional to the means
2. There is evidence of multiplicative effects (significant Tukey's test, smiling residuals)
3. Right-skewed frequency distribution

Logarithms to any base can be used, but common logarithms (base 10, base e) are generally the most convenient.

Data with values ≤ 0 cannot be transformed using logarithms.

If there are zeros, a 1 may be added to all datapoints before transforming.

To avoid negative logarithms, it is also legitimate to multiply all data points by a constant.

	Block					
Species—Treatment	**I**	**II**	**III**	**IV**	**mean**	**s**
Mice—control	0.18	0.30	0.28	0.44	0.3	0.11
Mice—vitamin	0.32	0.40	0.42	0.46	0.4	0.06
Subtotals	**0.5**	**0.7**	**0.7**	**0.9**	**0.35**	**0.08**
Chickens—control	2.0	3.0	1.8	2.8	2.40	0.58
Chickens—vitamin	2.5	3.3	2.5	3.3	2.90	0.46
Subtotals	**4.5**	**6.3**	**4.3**	**6.1**	**2.65**	**0.52**
Sheep—control	108.0	140.0	135.0	165.0	137.0	23.3
Sheep—vitamin	127.0	153.0	148.0	176.0	151.0	20.6
Subtotals	**235.0**	**293.0**	**283.0**	**341.0**	**144.0**	**22.0**

Means and standard deviations are proportional ($\overline{Y} / s =$ 4.4, 5.1, 6.5).

```
#read in, re-classify, and inspect the data
#as_data.frame, as.factor
str(log_dat)

#The ANOVA
log_mod<-lm(gain ~ trtmt + block, log_dat)
anova(log_mod)

#Need to assign contrast coefficients
#str() tells us that R orders the Trtmt levels this way: chick_cont,
chick_vit, mice_cont, ...
# Our desired contrasts:
# Contrast ‘Mam vs. Bird’         2,2,-1,-1,-1,-1
# Contrast ‘Mouse vs. Sheep'      0,0,1,1,-1,-1
# Contrast ‘Vit’                  1,-1,1,-1,1,-1
# Contrast ‘MamBird*Vit’          2,-2,-1,1,-1,1
# Contrast ‘MouShe*Vit’           0,0,1,-1,-1,1

contrastmatrix<-cbind(c(2,2,-1,-1,-1,-1),c(0,0,1,1,-1,-1),
      c(1,-1,1,-1,1,-1),c(2,-2,-1,1,-1,1),c(0,0,1,-1,-1,1))
contrasts(log_dat$trtmt)<-contrastmatrix

log_contrast_mod<-aov(gain ~ trtmt + block, log_dat)
summary(log_contrast_mod, split = list(trtmt = list("MvsB" = 1, "MvsS" = 2,
      "Vit" = 3, "MB*Vit" = 4, "MS*Vit" = 5)))

#TESTING ASSUMPTIONS
#Generate residual and predicted values
log_dat$resids <- residuals(log_mod)
log_dat$preds <- predict(log_mod)
log_dat$sq_preds <- log_dat$preds^2

#Look at a plot of residual vs. predicted values
plot(resids ~ preds, data = log_dat,
      xlab = "Predicted Values",
      ylab = "Residuals")

#Perform a Shapiro-Wilk test for normality of residuals
shapiro.test(log_dat$resids)

#Perform Levene's Test for homogenity of variances
library(car)
leveneTest(gain ~ trtmt, data = log_dat)

#Perform a Tukey 1-df Test for Non-additivity
log_1df_mod<-lm(gain ~ trtmt + block + sq_preds, log_dat)
anova(log_1df_mod)
```

The results:

	Original data	
	F	**p**
Treatment F test	174.4	< 0.0001
"Vitamin" Contrast	142.1	0.3025
"B/M*Vit" Interaction	57.2	0.5084
"M/S*Vit" Interaction	1.55	0.2322

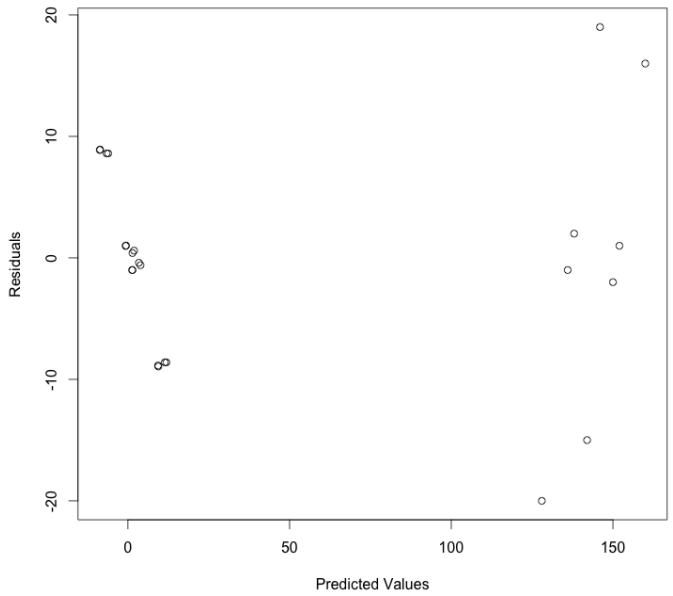
Results of analysis using the original data suggest *something is very wrong with the ANOVA*:

1. No significant effect of the vitamin treatment?
2. No significant interaction between vitamin and species?

The problem? The data do not satisfy the assumptions of the analysis:

	Original data	
	F	**p**
Shapiro-Wilk	W = 0.95	0.32
Levene's Test	3.37	**0.03**
Tukey's Nonadditivity Test	545.5	**< 0.0001**

This is reflected in the "tea-leaves" – it's smiling at you.



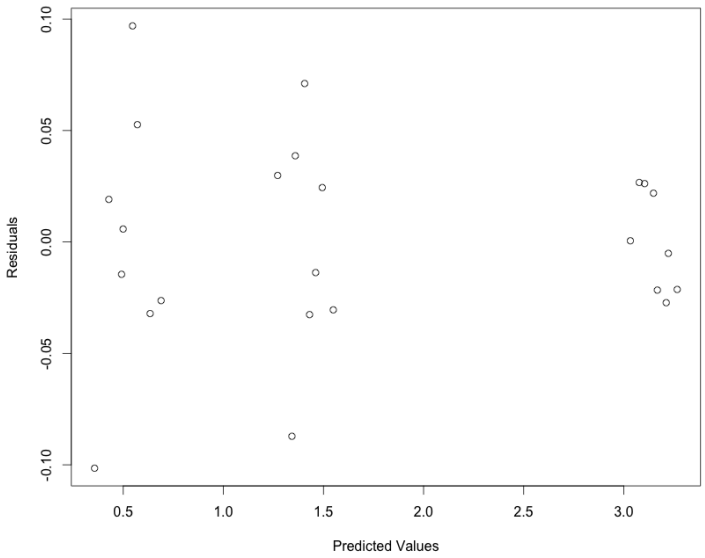
Residual vs. Predicted values **(Original data)**

Log-transforming the data and re-analyzing...

```
#Create log-transformed variable
log_dat$trans_gain<-log10(10*log_dat$gain)

#The ANOVA
trans_log_mod<-lm(trans_gain ~ trtmt + block, log_dat)
anova(trans_log_mod)

trans_log_contrast_mod<-aov(trans_gain ~ trtmt + block, log_dat)
summary(trans_log_contrast_mod, split = list(trtmt = list("MvsB" = 1,
      "MvsS" = 2, "Vit" = 3, "MB*Vit" = 4, "MS*Vit" = 5)))

#TESTING ASSUMPTIONS
#Generate residual and predicted values
log_dat$trans_resids <- residuals(trans_log_mod)
log_dat$trans_preds <- predict(trans_log_mod)
log_dat$sq_trans_preds <- log_dat$trans_preds^2

#Look at a plot of residual vs. predicted values
plot(trans_resids ~ trans_preds, data = log_dat,
     xlab = "Predicted Values",
     ylab = "Residuals")

#Perform a Shapiro-Wilk test for normality of residuals
shapiro.test(log_dat$trans_resids)

#Perform Levene's Test for homogenity of variances
leveneTest(trans_gain ~ trtmt, data = log_dat)

#Perform a Tukey 1-df Test for Non-additivity
trans_log_1df_mod<-lm(trans_gain ~ trtmt + block + sq_trans_preds, log_dat)
anova(trans_log_1df_mod)
```

The effect of the transformation:

	Original data		**Transformed data**	
	F	**p**	**F**	**p**
Shapiro-Wilk	W = 0.95	0.32	W = 0.97	0.56
Levene's Test	2.5	**0.07**	1.01	**0.44**
Tukey's Nonadditivity Test	545.5	**< 0.0001**	1.74	**0.21**
Treatment F test	174.4	< 0.0001	1859.6	< 0.0001
"Vitamin" Contrast	142.1	**0.3025**	16.4	**0.0011**
"B/M*Vit" Interaction	57.2	0.5084	0.01	0.9135
"M/S*Vit" Interaction	1.55	0.2322	3.14	0.0967



Residual vs. Predicted values **(Transformed data)**

Interactions: Problems of Interpretation

With the original data, the interaction question was:
"Does the amount of change in weight due to vitamins vary from species to species?"

With the log-transformed data, the question is
"Does the ***proportion*** change in weight due to vitamins vary from species to species?"

Log transformed data showing perfect additivity (NO INTERACTION):

	A0	A1		
B0	2.301	2.477	*0.176*	**Effect of A = + 0.176**
B1	2.602	2.778	*0.176*	
	0.301	*0.301*		
	Effect of B = + 0.301			

Underlying original data:

	A0	A1		
B0	200	300	*100*	**Effect of A = + 50%**
B1	400	600	*200*	
	200	*300*		
	Effect of B = + 100%			

A non-significant interaction in log-transformed data indicates a *constant percent or proportional difference* in the original data!

The effect of log transformation on means and variances

	Control					Mean	Var
Y	20	40	50	60	80	50	**500**
ln(Y)	2.9957	3.6889	3.9120	4.0943	3.820	3.8146	**0.2740**

	Treatment (Control + 20)					Mean	Var
Y	40	60	70	80	100	70	**500**
ln(Y)	3.6889	4.0943	4.2485	4.3820	4.6052	4.2038	0.1180

	Treatment (Control*1.5)					Mean	Var
Y	30	60	75	90	120	75	1125
ln(Y)	3.4012	4.0943	4.3175	4.4998	4.7875	4.2201	**0.2740**

Geometric mean $G = (Y_1 * Y_2 * \ldots * Y_n)^{1/n}$

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

$e^{3.8146} = 45.3586$

The square root transformation

General indicators that a square root transformation may be appropriate:

1. The **variances** are roughly proportional to the means
2. The data are counts of rare events
3. The data exhibit a Poisson frequency distribution

Data of this kind can be made more nearly normal and the correlation between variances and means can be reduced by using a square root transformation. For values near to or less than 10, it is best to use:

$$Y_{transformed} = \sqrt{Y_{original} + \frac{1}{2}}$$

Example: Number of lygus bugs per 50 sweeps with a net. The experiment is an RCBD, testing 10 insecticides (A – J) and a control (K).

Treatment	Block					
	I	II	III	IV	mean	s^2
A	7	5	4	1	4.25	6.25
B	6	1	2	1	2.50	5.67
C	6	2	1	0	2.25	6.92
D	0	1	2	0	0.75	0.92
E	1	0	1	2	1.00	0.67
F	5	14	9	15	10.75	21.58
G	8	6	3	6	5.75	4.25
H	3	0	5	9	4.25	14.25
I	4	10	13	5	8.00	18.00
J	6	11	5	2	6.00	14.00
K	8	11	2	6	6.75	14.25

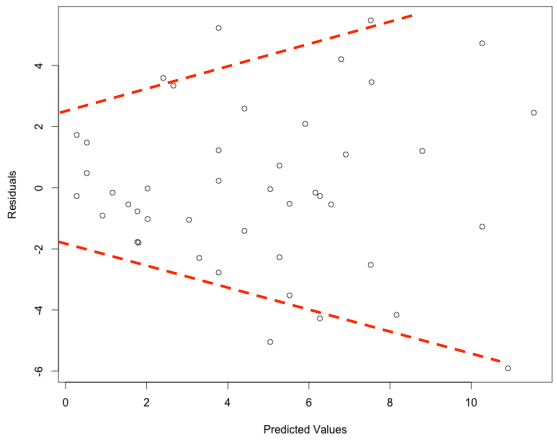
Means and variance are roughly proportional ($\bar{Y} / s^2 \approx 0.7$).

```
#Create sqrt-tranformed variable
bug_dat$trans_count<-sqrt(0.5+bug_dat$count)
```

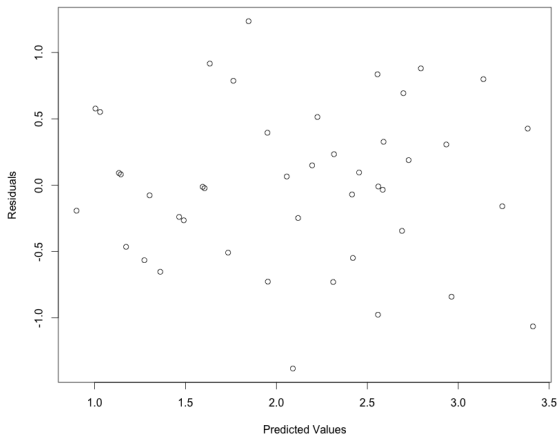
A comparative summary of the results:

	Original data		Transformed data	
	F	p	F	p
Shapiro-Wilk	W = 0.98	0.8204	W = 0.99	0.9790
Levene's Test	1.5	0.1729	0.6	0.8259
Tukey's Nonadditivity Test	0.6	0.4351	0.1	0.7892
Treatment F test	3.7	0.0026	4.0	0.0014

The two analyses are not that different. Both show a highly significant treatment effect.



Res vs. Pred (Original data)
Correlation mean-variance **r = 0.89 ****



Res vs. Pred (Transformed data)
Correlation mean-variance **r = 0.37**

The differences appear when we start separating means.

Original Data Tukey Grouping		Means	N	Trtmt
	A	10.750	4	**F**
B	A	8.000	4	**I**
B	A	6.750	4	K
B	A	6.000	4	J
B	A	5.750	4	G
B	A	4.250	4	H
B	A	4.250	4	A
B		2.500	4	**B**
B		2.250	4	C
B		1.000	4	E
B		0.750	4	**D**

Transformed Data Tukey Grouping			*Detransformed* Means	N	Trtmt
	A		10.3445	4	**F**
B	A		7.5957	4	**I**
B	A	C	6.3084	4	K
B	A	C	5.6073	4	J
B	A	C	5.5851	4	G
B	A	C	3.9416	4	H
B	A	C	3.5052	4	A
B	A	C	2.2060	4	**B**
B		C	1.7970	4	C
B		C	0.9028	4	E
		C	0.6130	4	**D**

Weighted means are obtained by *detransforming* the means of the transformed data.

For example, the mean of the transformed data for Treatment F was 3.293076.
The detransformed mean is therefore:

$$3.293076^2 - 0.5 = 10.3445$$

The detransformed means are *less than* the means of the original data
because more weight is given to smaller numbers.

The general effect of the square root transformation is to increase the precision with which we measure differences between *small* means.

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

Data requiring the square root transformation do not violate the ANOVA assumptions as drastically as data requiring a log transformation.

Consequently, the changes in the analysis brought about by the transformation are not nearly as spectacular.

The arcsine or angular transformation

General indicators that an arcsine transformation may be appropriate:

1. The data are counts expressed as a proportion of the total sample.
2. The data exhibit a binomial frequency distribution (smaller variance at the two ends of the range of values (0, 100%) but larger in the middle (50%).

The appropriate transformation for this type of data is:

$$Y_{transformed} = \arcsin\left(\sqrt{\frac{Y_{original}}{Total\ \ Possible\ Counts}}\right)$$

Number of lettuce seeds germinating in samples of 50

Treatment	Blocks 1	2	3	Mean	s_i^2
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

In this case, since the total number of seeds in each replication is 50, the arcsine transformation would be:

$$Y_{transformed} = \arcsin\left(\sqrt{\frac{Y_{original}}{50}}\right)$$

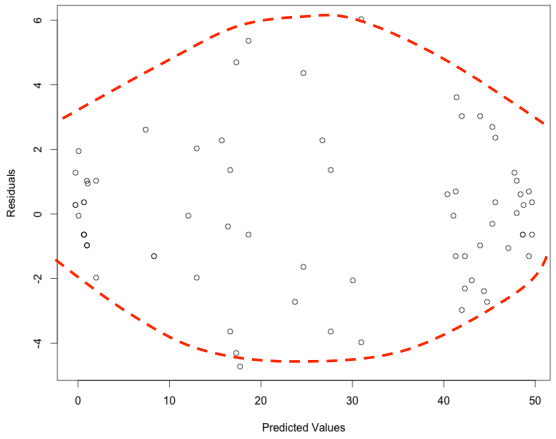
Note: When the proportions fall between 0.30 and 0.70, it is generally not necessary to apply the arcsine transformation.

In R, the transformation would be coded as follows:

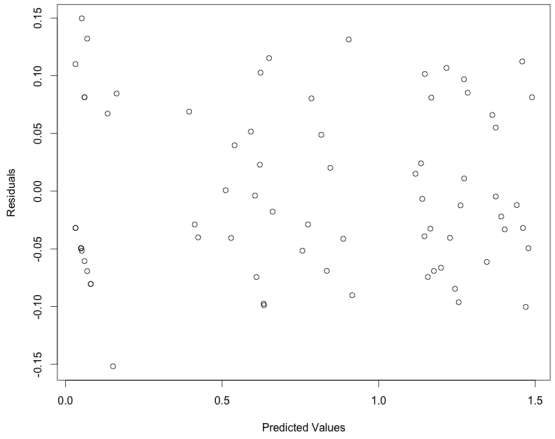
```
#Create arcsin-tranformed variable
germ_dat$trans_count<-asin(sqrt(germ_dat$count/50))
```

A comparative summary of the results:

	Original data		Transformed data	
	F	**p**	**F**	**p**
Levene's Test	2.64	**0.0023**	0.96	**0.5255**
Treatment F test	147.5	< 0.0001	98.59	< 0.0001



Res vs. Pred **(Original data)**



Res vs. Pred **(Transformed data)**

Important differences emerge in individual comparisons (e.g. contrasts and mean separations):

Original Data					Transformed Data				
			Original Means	Trtmt	Detransformed Means				
A			49.333	**24**	49.554	**A**			
A			49.000	**23**	49.348	**A**			
A			48.333	**22**	48.370	**A**	**B**		
A			47.667	**21**	47.827	**A**	**B**		
A			45.333	**20**	45.637	**A**	**B**		
A			45.000	**19**	45.302	**A**	**B**		
A			43.667	**18**	43.914	**A**	**B**		
A			42.000	**17**	42.134		**B**	**C**	
A			41.667	**16**	41.834		**B**	**C**	
A			41.000	**15**	41.015		**B**	**C**	
	B		30.667	**14**	30.789	**D**		**C**	
	B		27.333	**13**	27.341	**D**	**E**		
	B	**C**	24.333	**12**	24.333	**D**	**E**	**F**	
D		**C**	18.333	**11**	18.216	**D**	**E**	**F**	**G**
D		**C**	17.000	**10**	16.905	**D**	**E**	**F**	**G**
D	**E**	**C**	16.333	**9**	16.285		**E**	**F**	**G**
D	**E**		12.667	**8**	12.631			**F**	**G**
	E	**F**	8.000	**7**	7.952	**H**			**G**
		F	1.667	**6**	1.111	**H**	**I**		
		F	0.667	**5**	0.225		**I**		
		F	0.667	**4**	0.225		**I**		
		F	0.333	**1**	0.112		**I**		
		F	0.333	**3**	0.112		**I**		
		F	0.333	**2**	0.112		**I**		

Which set of conclusions should we accept? We accept the conclusions based on the more valid analysis.

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 *correct*.

The power transformation

> Hinz, PN and HA Eagles (1976) Estimation of a transformation for the analysis of some agronomic and genetic experiments. *Crop Science* **16**: 280-283.

Experiments are commonly conducted using **replicated trials over a broad range of locations and environmental conditions**. Often, the means and the residual variances differ markedly across environments.

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

The power transformation method provides a means of selecting an appropriate transformation from the broad class of power transformations by employing the dataset itself to estimate the exponent needed to transform the original measures.

$$Y_{transformed} = \left(Y_{original}\right)^{a}$$, where **a** is determined empirically

Generally, if the variances and means are positively correlated, **a** will be less than 1. If negatively correlated, **a** will be greater than 1. If **a** is found to be approximately equal to 0, the log transformation is recommended.

$Y = \sqrt{X}$ if **a** = ½
$Y = 1 / X$ if **a** = -1

$$Y_{transformed} = \left(Y_{original}\right)^{a} \text{ if } \mathbf{a} \neq 0$$
$$Y_{transformed} = \log\left(Y_{original}\right) \text{ if } \mathbf{a} \approx 0$$

By this method, **a** is determined by obtaining the slope **m** of the linear regression of $\log\left(s_i^2\right)$ versus $\log\left(\overline{Y}_i\right)$:

$$\log\left(s_i^2\right) = m * \log\left(\overline{Y}_i\right) + b$$

Example 1: Using the means and variances from the lygus bug dataset (see Square Root Transformation above), we can write a small program:

```
#Finding the exponent for a power transformation

means <- aggregate(bug_dat$count, list(bug_dat$trtmt), mean)
vars <- aggregate(bug_dat$count, list(bug_dat$trtmt), var)

logmeans <- log10(means$x)
logvars <- log10(vars$x)

power_mod<-lm(logvars ~ logmeans)
summary(power_mod)
```

This code performs a linear regression of log(Variance) on log(Mean) of the original data. The output includes an estimate (**m**) of the slope:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.1072     0.1281   0.837 0.424270
logmeans       1.2399     0.1932   6.418 0.000123 ***
```

$$\log\left(s_i^2\right) = 1.2399 * \log\left(\bar{Y}_i\right) + 0.1072$$

m = 1.2399
a = 1 – (1.2399/ 2) = **0.38**

The transformation would then be coded as follows:

```
#Create power-tranformed variable
bug_dat$trans_count<-(bug_dat$count)^0.38
```

Under this transformation, Levene's test improves (p = 0.81).

Example 2: Using the means and variances from the vitamin dataset (see Logarithmic Transformation), we obtain:

$$\log\left(s_i^2\right) = 1.8521 * \log\left(\bar{Y}_i\right) - 1.3393$$

m = 1.8521
a = 1 – (1.8521/ 2) = **0.074**

	Log Transformed data		Power Transformed data	
	F	**p**	**F**	**p**
Shapiro-Wilk	W = 0.97	0.56	W = 0.97	0.61
Levene's Test	1.01	0.44	0.60	0.70
Tukey's Nonadditivity Test	1.74	0.21	0.002	0.97

Reporting results from transformed data

Tests for significance are performed on the data that satisfy the assumptions.

Estimates of means should be given in the original scale.

Example: Using log transformed data, the mean of Treatment B is found to be 1.176. The Tukey MSD = 0.023. In the *transformed scale*, the confidence interval about the mean of Treatment B is symmetric:

$$\mu_B = \bar{Y}_B \pm MSD = 1.176 \pm 0.023 = \textbf{[1.153 , 1.199]}$$

But, the proper way to report this confidence interval is by *detransforming* these confidence limits:

$$\mu_B = [10^{1.153} , 10^{1.199}] = \textbf{[14.223 , 15.812]}$$

This is an asymmetric interval about the detransformed mean $10^{1.176} = 14.997$.