Partitioning Variance in your Data - ANOVAs and GLMMs

A Michelle Edwards, Ph.D., MLIS

November 5, 2019

Table of Contents

Note: This is part of my Computer setup - you will NOT see this!! Please ignore

```
## SAS found at C:/Program Files/SASHome/SASFoundation/9.4/sas.exe

## sas, saslog, sashtml, sashtml5, and sashtmllog & sashtml5log engines

##    are now ready to use.
```

End of Note

Reading and cleaning the data

Download the Excel file from the OACStats blog post and save it in the directory you chose to use as your working directory for this workshop.

Let's work together and read the default worksheet (sheet1) from the RCBD_excel Excel file and called it **rcbd**

Choose a method to enter the data into SAS. You can copy and paste it into the Editor window as I've done here. Or you can save it as a CSV and use the INFILE statement.

Take the next few mintues now to enter this data into your SAS program.

```
Data rcbd;
  input block trmt Nitrogen Weed Bin_weed;
  datalines;
1 1 34.98 81 0.81
2 1 41.22 87 0.87
3 1 36.94 89 0.89
4 1 39.97 79 0.79
1 2 40.89 88 0.88
2 2 46.69 85 0.85
3 2 46.65 99 0.99
4 2 41.90 86 0.86
1 3 42.07 54 0.54
2 3 49.42 23 0.23
3 3 52.68 11 0.11
4 3 42.91 16 0.16
1 4 37.18 10 0.10
2 4 45.85 23 0.23
3 4 40.23 10 0.10
4 4 39.20 69 0.69
1 5 37.99 61 0.61
2 5 41.99 06 0.06
3 5 37.61 02 0.02
4 5 40.45 75 0.75
1 6 34.89 21 0.21
2 6 50.15 08 0.08
3 6 44.57 27 0.27
4 6 43.29 03 0.03
;
Run;
```

Review the contents of the data you just loaded. Anything unusual about it?

Use the PROC PRINT to review the data:

```
Proc print data=rcbd;
Run;
```

Obs	block	trmt	Nitrogen	Weed	Bin_weed
1	1	1	34.98	81	0.81
2	2	1	41.22	87	0.87
3	3	1	36.94	89	0.89
4	4	1	39.97	79	0.79
5	1	2	40.89	88	0.88
6	2	2	46.69	85	0.85
7	3	2	46.65	99	0.99
8	4	2	41.90	86	0.86
9	1	3	42.07	54	0.54
10	2	3	49.42	23	0.23
11	3	3	52.68	11	0.11
12	4	3	42.91	16	0.16
13	1	4	37.18	10	0.10
14	2	4	45.85	23	0.23
15	3	4	40.23	10	0.10
16	4	4	39.20	69	0.69
17	1	5	37.99	61	0.61
18	2	5	41.99	6	0.06
19	3	5	37.61	2	0.02
20	4	5	40.45	75	0.75
21	1	6	34.89	21	0.21
22	2	6	50.15	8	0.08
23	3	6	44.57	27	0.27
24	4	6	43.29	3	0.03

Now that our data has been entered let's start to review our analysis.

Analyzing our data - Paritioning of the variation - Analysis of variation (ANOVA) - Are there treatment differences?

With many of our research projects, we are implementing "some" treatment. We do this to answer a specific research question. Maybe there are beliefs that a pariticular treatment will provide more fruit than another. Or maybe a particular treatment will reduce the severity of a disease. Just remember that we are doing this to answer a specific research question.

For many research projects, we want to determine whether there are differences between the treatments we impose. To test for these differences, we design an experiment. Ideally, we design the experiment in a way that will allow us to answer our research question.

Let's review a few aspects of an experimental design:

- Experimental Unit: The unit to which the treatment is applied. Sometimes we lose track of what we are measuring and what we are testing. This is a key component to any analysis.
- Identify and be clear about the measurements you are taking - keep them as similar as possible - ensure we are measuring the same thing every time!
- Treatments - make sure you are applying them the same way to every experimental unit

In an ideal and perfect world, our experimental units would be identical, our treatments would be applied identically, our measures would be perfect, leading to the only differences to be seen attributed to the treatments. Is this possible? NOPE - why?

- There is natural variation between experimental units
- There will be variability in the measurements we take
- We just cannot replicate our treatments exactly
- Some of our experimental units might react differently to the same treatement
- Other factors that may play a role - weather, lighting, etc…

All of these are what we refer to as sources of experimental error.

Our goal with an experimental design is to control the experimental error - we want to be able to explain the variation or difference we see in our measures in a way that will lead us to saying Yes there are differences between our treatments or NO there are not - while knowing we did the best we could containing that random error.

Goal of ANOVA - as we traditionally come to know it - is to partition the variation in our outcome measures. In other words, to be able to explain the variation in our outcome measures and conclude whether our treatments were similar or not.

RCBD Example

Let's review the data collected from a small RCBD trial. There were 4 blocks, where 6 treatments were randomly assigned to each. The statistical model for this experimental design is:

$Nitrogen_{ij} = \mu + block_i + trmt_j + e_{ij}$

Where:

- $Nitrogen_{ij}$ = nitrogen measure taken on $plot_{ij}$
- μ = overall mean of nitrogen
- $block_i$ = random effect of $block_i$
- $trmt_j$ = fixed effect of treatment $_j$
- e_{ij} = random experimental error

Fixed vs random effects

Fixed effects are something you want to study - you set out the levels that you are interested in. You "fix" the levels. The results from your experiment can only talk about the levels you studied.

- Example #1: I want to see whether 1st year students prefer Coke or Pepsi
- Example #2: I want to see the effect of 3 levels of fertilizer on my crop

Random effects are factors in your design that may contribute variation in your outcome measure, but you are not interested in it. You only want to account for it, before looking at your treatment effects.

- Example #1: I want to study the effect of fertilizer on my crop
- Example #2: Block effect, Weather, etc…

Let's first try running our data as a CRD or a Completely Randomized Design - no block effect. I want to do this to show you the differences between a fixed effects model and the RCBD model or a mixed effects model we will run in a moment.

To create our SAS code we are going to translate our statistical model into SAS.

Since we are going to analyze our data as a CRD - our statistical model would now be:

$Nitrogen_{ij} = \mu + trmt_j + e_{ij}$

Identify by underlining or highlighting the pieces of information we have in our dataset (I will list them here to be more visual):

- nitrogen
- trmt

These are the 2 pieces of information we have in our dataset that match our statistical model. The μ and e_{ij} will be calculated by SAS.

The next thing you need to identify is whether the variables on the right hand side of our model are FIXED or RANDOM.

One more thing - the variables listed on the right hand side of our statistical model - are they grouping variables or continuous measures?

Now to put it all together in SAS:

CLASS statement - you list all variables that are on the right side of our statistical model that are grouping variables - trmt in this dataset

MODEL statement - you rewrite the statistical model - but only use FIXED variables - **nitrogen** = **trmt**

That's it for this example. Let's try it out!

```
Proc glimmix data=rcbd;
  class trmt;
  model Nitrogen = trmt;
  title "Proc GLIMMIX Results";
  lsmeans trmt / pdiff adjust=tukey ilink lines;
  output out=second predicted=pred residual=resid residual(noblup)=mresid stu
dent=studentresid student(noblup)=smresid;
Run;
```

Proc GLIMMIX Results

The GLIMMIX Procedure

Model Information

Data Set	WORK.RCBD
Response Variable	Nitrogen
Response Distribution	Gaussian
Link Function	Identity
Variance Function	Default
Variance Matrix	Diagonal
Estimation Technique	Restricted Maximum Likelihood
Degrees of Freedom Method	Residual

Class Level Information

Class	Levels	Values
trmt	6	1 2 3 4 5 6

Number of Observations Read	24
Number of Observations Used	24

Dimensions

Covariance Parameters	1
Columns in X	7
Columns in Z	0
Subjects (Blocks in V)	1
Max Obs per Subject	24

Optimization Information

Optimization Technique	None
Parameters	7
Lower Boundaries	1
Upper Boundaries	0
Fixed Effects	Not Profiled

Fit Statistics

-2 Res Log Likelihood	110.34
AIC (smaller is better)	124.34
AICC (smaller is better)	135.54
BIC (smaller is better)	130.57
CAIC (smaller is better)	137.57
HQIC (smaller is better)	125.20
Pearson Chi-Square	305.01
Pearson Chi-Square / DF	16.95

Type III Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
trmt	5	18	2.38	0.0802

trmt Least Squares Means

trmt	Estimate	Standard Error	DF	t Value	Pr > |t|	Mean	Standard Error Mean
1	38.2775	2.0582	18	18.60	<.0001	38.2775	2.0582
2	44.0325	2.0582	18	21.39	<.0001	44.0325	2.0582
3	46.7700	2.0582	18	22.72	<.0001	46.7700	2.0582
4	40.6150	2.0582	18	19.73	<.0001	40.6150	2.0582
5	39.5100	2.0582	18	19.20	<.0001	39.5100	2.0582
6	43.2250	2.0582	18	21.00	<.0001	43.2250	2.0582

Differences of trmt Least Squares Means
Adjustment for Multiple Comparisons: Tukey

trmt	_trmt	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P
1	2	-5.7550	2.9108	18	-1.98	0.0635	0.3917
1	3	-8.4925	2.9108	18	-2.92	0.0092	0.0828
1	4	-2.3375	2.9108	18	-0.80	0.4324	0.9633
1	5	-1.2325	2.9108	18	-0.42	0.6770	0.9980
1	6	-4.9475	2.9108	18	-1.70	0.1064	0.5489
2	3	-2.7375	2.9108	18	-0.94	0.3594	0.9304
2	4	3.4175	2.9108	18	1.17	0.2557	0.8432
2	5	4.5225	2.9108	18	1.55	0.1377	0.6366
2	6	0.8075	2.9108	18	0.28	0.7846	0.9997
3	4	6.1550	2.9108	18	2.11	0.0487	0.3234
3	5	7.2600	2.9108	18	2.49	0.0226	0.1776
3	6	3.5450	2.9108	18	1.22	0.2390	0.8227
4	5	1.1050	2.9108	18	0.38	0.7087	0.9988
4	6	-2.6100	2.9108	18	-0.90	0.3817	0.9424
5	6	-3.7150	2.9108	18	-1.28	0.2181	0.7936

Tukey Grouping for trmt Least Squares Means (Alpha=0.05)

LS-means with the same letter are not significantly different.

trmt	Estimate	
3	46.7700	A
		A
2	44.0325	A
		A
6	43.2250	A
		A
4	40.6150	A
		A
5	39.5100	A
		A
1	38.2775	A

Note that we are using PROC GLIMMIX for all of our analyses. Also note that we are saving our residuals in a new dataset called **second**. we will be using these in a moment.

Let's review the output together.

Testing the Assumptions

Something that we learned when we were first taught ANOVAs was something about assumptions. One of those assumptions, was that the data going into the ANOVA had to have a normal distribution. I'm going to tell you now - to not worry about that. I remember being told this and also being told that ANOVA is a robust analysis for data that isn't too normal going in.

So - what assumptions should we be checking?

1. Residuals are random - no relationship to our treatment
2. Homogeneity of residuals across our treatment groups
3. Residuals are normally distributed
4. Residuals have a mean of 0

It's ALL about the residuals - we are testing the residuals to determine whether our model fits our data.

So how do we do this?

In SAS, we have a number of plots to review as well as the Shapiro-Wilk statistic to test normality of the residuals. Let's rerun our SAS code with the residual analysis below.

NOTE: In the SAS code you will see there is some coding to save your output as a PDF file. This added coding will be a great tool for you to learn and it is a way for me to show you the graphs that SAS creates when running the residual analysis. You will note that these are missing from this handout - you need to run them on your system to see.

```
ods pdf file="CRD_output.pdf";

Proc glimmix data=rcbd;
  class trmt;
  model Nitrogen = trmt;
  title "Proc GLIMMIX Results";
  lsmeans trmt / pdiff adjust=tukey ilink lines;
  output out=second predicted=pred residual=resid residual(noblup)=mresid stu
dent=studentresid student(noblup)=smresid;
Run;

/*  Linearity of fixed effects - both as a scatter and a boxplot */
Proc sgplot data=second;
  scatter y=smresid x=trmt;
  refline 0;
Run;

Proc sgplot data=second;
  vbox smresid / group=trmt datalabel;
Run;
```

```
/* Homogeneity of effects */
Proc sgscatter data=second;
  plot studentresid*(pred trmt block);
Run;

/* Q-Q plot and Shapiro-Wilk for normal distribution */
Proc univariate data=second normal plot;
  var studentresid;
Run;

ods pdf close;
```

Proc GLIMMIX Results

The GLIMMIX Procedure

Model Information

Data Set	WORK.RCBD
Response Variable	Nitrogen
Response Distribution	Gaussian
Link Function	Identity
Variance Function	Default
Variance Matrix	Diagonal
Estimation Technique	Restricted Maximum Likelihood
Degrees of Freedom Method	Residual

Class Level Information

Class	Levels	Values
trmt	6	1 2 3 4 5 6

Number of Observations Read	24
Number of Observations Used	24

Dimensions

Covariance Parameters	1
Columns in X	7
Columns in Z	0
Subjects (Blocks in V)	1
Max Obs per Subject	24

Optimization Information

Optimization Technique	None
Parameters	7

Lower Boundaries	1
Upper Boundaries	0
Fixed Effects	Not Profiled

Fit Statistics

-2 Res Log Likelihood	110.34
AIC (smaller is better)	124.34
AICC (smaller is better)	135.54
BIC (smaller is better)	130.57
CAIC (smaller is better)	137.57
HQIC (smaller is better)	125.20
Pearson Chi-Square	305.01
Pearson Chi-Square / DF	16.95

Type III Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
trmt	5	18	2.38	0.0802

trmt Least Squares Means

trmt	Estimate	Standard Error	DF	t Value	Pr > |t|	Mean	Standard Error Mean
1	38.2775	2.0582	18	18.60	<.0001	38.2775	2.0582
2	44.0325	2.0582	18	21.39	<.0001	44.0325	2.0582
3	46.7700	2.0582	18	22.72	<.0001	46.7700	2.0582
4	40.6150	2.0582	18	19.73	<.0001	40.6150	2.0582
5	39.5100	2.0582	18	19.20	<.0001	39.5100	2.0582
6	43.2250	2.0582	18	21.00	<.0001	43.2250	2.0582

Differences of trmt Least Squares Means
Adjustment for Multiple Comparisons: Tukey

trmt	_trmt	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P
1	2	-5.7550	2.9108	18	-1.98	0.0635	0.3917
1	3	-8.4925	2.9108	18	-2.92	0.0092	0.0828
1	4	-2.3375	2.9108	18	-0.80	0.4324	0.9633
1	5	-1.2325	2.9108	18	-0.42	0.6770	0.9980
1	6	-4.9475	2.9108	18	-1.70	0.1064	0.5489
2	3	-2.7375	2.9108	18	-0.94	0.3594	0.9304
2	4	3.4175	2.9108	18	1.17	0.2557	0.8432
2	5	4.5225	2.9108	18	1.55	0.1377	0.6366

2	6	0.8075	2.9108	18	0.28	0.7846	0.9997
3	4	6.1550	2.9108	18	2.11	0.0487	0.3234
3	5	7.2600	2.9108	18	2.49	0.0226	0.1776
3	6	3.5450	2.9108	18	1.22	0.2390	0.8227
4	5	1.1050	2.9108	18	0.38	0.7087	0.9988
4	6	-2.6100	2.9108	18	-0.90	0.3817	0.9424
5	6	-3.7150	2.9108	18	-1.28	0.2181	0.7936

Tukey Grouping for trmt Least Squares Means (Alpha=0.05)

LS-means with the same letter are not significantly different.

trmt	Estimate	
3	46.7700	A
		A
2	44.0325	A
		A
6	43.2250	A
		A
4	40.6150	A
		A
5	39.5100	A
		A
1	38.2775	A

Proc GLIMMIX Results

The UNIVARIATE Procedure
Variable: studentresid (Studentized Residual)

Moments

N	24	Sum Weights	24
Mean	0	Sum Observations	0
Std Deviation	1.02150784	Variance	1.04347826
Skewness	-0.0854656	Kurtosis	-0.0703732
Uncorrected SS	24	Corrected SS	24
Coeff Variation	.	Std Error Mean	0.20851441

Basic Statistical Measures

Location		Variability	
Mean	0.00000	Std Deviation	1.02151
Median	-0.04488	Variance	1.04348
Mode	.	Range	4.28057

Interquartile Range		1.47863		

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	0	Pr > |t|	1.0000
Sign	M	0	Pr >= |M|	1.0000
Signed Rank	S	-2	Pr >= |S|	0.9559

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.979823	Pr < W	0.8926
Kolmogorov-Smirnov	D	0.101629	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.035043	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.230648	Pr > A-Sq	>0.2500

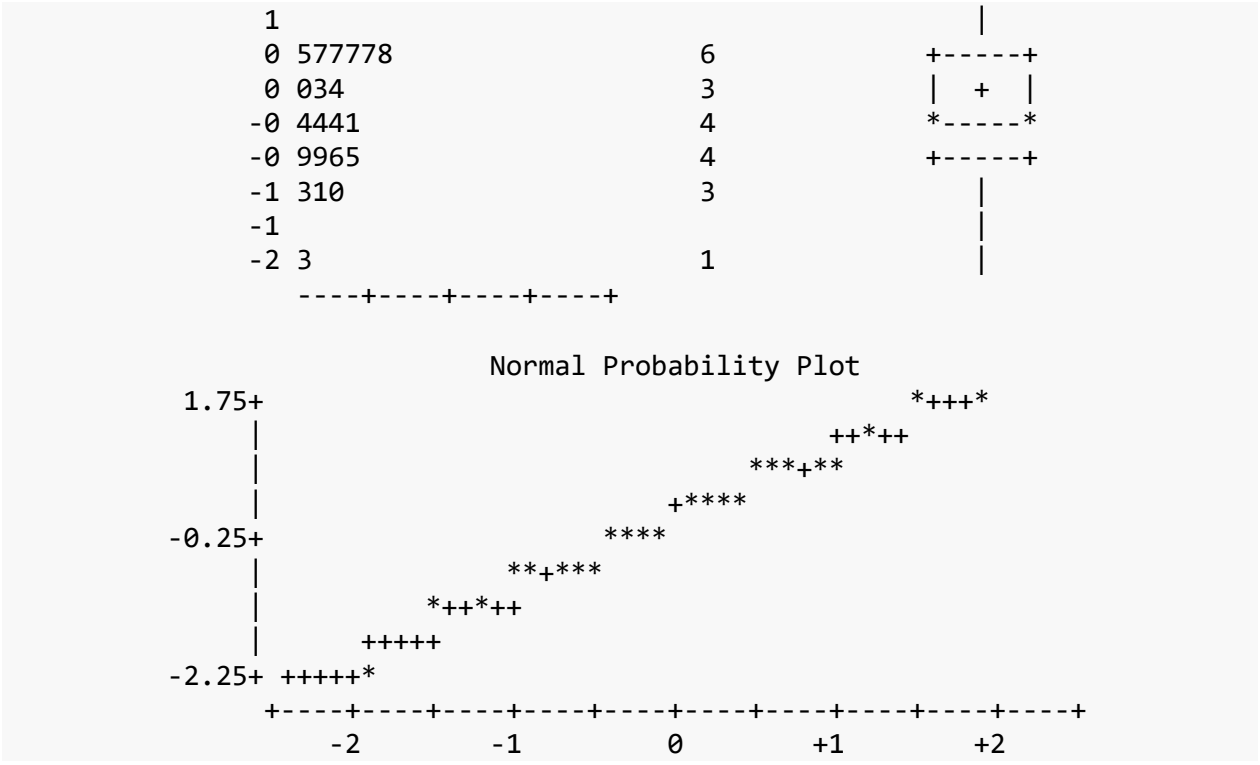
Quantiles (Definition 5)

Level	Quantile
100% Max	1.9425253
99%	1.9425253
95%	1.6578086
90%	1.4684650
75% Q3	0.7387908
50% Median	-0.0448815
25% Q1	-0.7398427
10%	-1.0827650
5%	-1.3183926
1%	-2.3380431
0% Min	-2.3380431

Extreme Observations

------Lowest------		------Highest-----	
Value	Obs	Value	Obs
-2.338043	21	0.745453	6
-1.318393	9	0.825398	2
-1.082765	12	1.468465	14
-0.963549	13	1.657809	11
-0.924979	1	1.942525	22

Stem	Leaf	#	Boxplot
1	579	3	|



Let's review the results together.

Open and review the **CRD_output.pdf document**.

Once we are happy with our assumptions, then we can dig further into our results to answer which treatment is different from which. You will notice that these results are already available- **BUT** always look at the residuals first before getting too excited about the analysis results.

From the results we can see that all of our treatments are similar.

Mixed Model Analysis

A mixed model analysis is using a model that has fixed and random effects - mixed. How do we know if we have a mixed model? Always go back to your experimental design. Sorry - everything comes from that design. Earlier I said we were treating our data as a CRD - but I noted that this was not the case - that we were using it as an example.

Our experiment was conducted as an RCBD.

$Nitrogen_{ij} = \mu + block_i + trmt_j + e_{ij}$

trmt is a fixed effect - we are interested in differences between the treatments we set out in this trial. **block** however was our way to help explain variation we may see in the field - so it is not something I want to see if there are differences between, it is something that I want to acknowledge may exist, account for it, and then move one. So it is added to our model to explain some of the variation in our outcome variable and reduce our experimental error.

NOw we go back and translate our statistical model into SAS the same way we did earlier. THe only difference is that we need to add a RANDOM statement to list any effects that we are considering as random in our statistical model. Let's modify our code and rerun. Remember look at the residuals first!

Let's try it out:

```
ods pdf file="rcbd_output.pdf";

Proc glimmix data=rcbd;
  class block trmt;
  model Nitrogen = trmt;
  random block;
  title "Proc GLIMMIX Results";
  lsmeans trmt / pdiff adjust=tukey ilink lines;
  output out=second predicted=pred residual=resid residual(noblup)=mresid stu
dent=studentresid student(noblup)=smresid;
Run;

/*  Linearity of fixed effects - both as a scatter and a boxplot */
Proc sgplot;
  scatter y=smresid x=trmt;
  refline 0;
Run;

Proc sgplot;
  vbox smresid / group=trmt datalabel;
Run;

/* Homogeneity of effects */
Proc sgscatter;
  plot studentresid*(pred trmt block);
Run;
```

```
/* Q-Q plot and Shapiro-Wilk for normal distribution */
Proc univariate normal plot;
  var studentresid;
Run;

ods pdf close;
```

Proc GLIMMIX Results

The GLIMMIX Procedure

Model Information

Data Set	WORK.RCBD
Response Variable	Nitrogen
Response Distribution	Gaussian
Link Function	Identity
Variance Function	Default
Variance Matrix	Not blocked
Estimation Technique	Restricted Maximum Likelihood
Degrees of Freedom Method	Containment

Class Level Information

Class	Levels	Values
block	4	1 2 3 4
trmt	6	1 2 3 4 5 6

Number of Observations Read	24
Number of Observations Used	24

Dimensions

G-side Cov. Parameters	1
R-side Cov. Parameters	1
Columns in X	7
Columns in Z	4
Subjects (Blocks in V)	1
Max Obs per Subject	24

Optimization Information

Optimization Technique	Dual Quasi-Newton
Parameters in Optimization	1
Lower Boundaries	1
Upper Boundaries	0
Fixed Effects	Profiled

Residual Variance Starting From: Profiled Data

Iteration History

Iteration	Restarts	Evaluations	Objective Function	Change	Max Gradient
0	0	4	101.56577191	.	2.22E-16

Convergence criterion (ABSGCONV=0.00001) satisfied.

Fit Statistics

-2 Res Log Likelihood	101.57
AIC (smaller is better)	105.57
AICC (smaller is better)	106.37
BIC (smaller is better)	104.34
CAIC (smaller is better)	106.34
HQIC (smaller is better)	102.87
Generalized Chi-Square	129.61
Gener. Chi-Square / DF	7.20

Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error
block	9.7446	8.9470
Residual	7.2006	2.6293

Type III Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
trmt	5	15	5.59	0.0042

trmt Least Squares Means

trmt	Estimate	Standard Error	DF	t Value	Pr > |t|	Mean	Standard Error Mean
1	38.2775	2.0582	15	18.60	<.0001	38.2775	2.0582
2	44.0325	2.0582	15	21.39	<.0001	44.0325	2.0582
3	46.7700	2.0582	15	22.72	<.0001	46.7700	2.0582
4	40.6150	2.0582	15	19.73	<.0001	40.6150	2.0582

5	39.5100	2.0582	15	19.20	<.0001	39.5100	2.0582
6	43.2250	2.0582	15	21.00	<.0001	43.2250	2.0582

Differences of trmt Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

trmt	_trmt	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P
1	2	-5.7550	1.8974	15	-3.03	0.0084	0.0742
1	3	-8.4925	1.8974	15	-4.48	0.0004	0.0049
1	4	-2.3375	1.8974	15	-1.23	0.2369	0.8150
1	5	-1.2325	1.8974	15	-0.65	0.5258	0.9849
1	6	-4.9475	1.8974	15	-2.61	0.0198	0.1553
2	3	-2.7375	1.8974	15	-1.44	0.1697	0.7025
2	4	3.4175	1.8974	15	1.80	0.0918	0.4934
2	5	4.5225	1.8974	15	2.38	0.0308	0.2226
2	6	0.8075	1.8974	15	0.43	0.6765	0.9978
3	4	6.1550	1.8974	15	3.24	0.0055	0.0505
3	5	7.2600	1.8974	15	3.83	0.0017	0.0168
3	6	3.5450	1.8974	15	1.87	0.0814	0.4559
4	5	1.1050	1.8974	15	0.58	0.5690	0.9907
4	6	-2.6100	1.8974	15	-1.38	0.1892	0.7402
5	6	-3.7150	1.8974	15	-1.96	0.0691	0.4079

Tukey-Kramer Grouping for trmt Least Squares Means (Alpha=0.05)

LS-means with the same letter are not significantly different.

trmt	Estimate		
3	46.7700		A
			A
2	44.0325	B	A
		B	A
6	43.2250	B	A
		B	A
4	40.6150	B	A
		B	
5	39.5100	B	
		B	
1	38.2775	B	

Proc GLIMMIX Results

The UNIVARIATE Procedure
Variable: studentresid (Studentized Residual)

Moments

N	24	Sum Weights	24
Mean	0	Sum Observations	0
Std Deviation	1.02150784	Variance	1.04347826
Skewness	0.12613884	Kurtosis	0.34838415
Uncorrected SS	24	Corrected SS	24
Coeff Variation	.	Std Error Mean	0.20851441

Basic Statistical Measures

Location		Variability	
Mean	0.00000	Std Deviation	1.02151
Median	-0.06155	Variance	1.04348
Mode	.	Range	4.51955
		Interquartile Range	1.33274

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	0	Pr > |t|	1.0000
Sign	M	0	Pr >= |M|	1.0000
Signed Rank	S	-3	Pr >= |S|	0.9338

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.988531	Pr < W	0.9916
Kolmogorov-Smirnov	D	0.089759	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.032838	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.191551	Pr > A-Sq	>0.2500

Quantiles (Definition 5)

Level	Quantile
100% Max	2.3233687
99%	2.3233687
95%	1.6452580
90%	1.1151230
75% Q3	0.7761524
50% Median	-0.0615464
25% Q1	-0.5565866
10%	-1.3184409
5%	-1.4740124

```
                  1%           -2.1961817
                  0% Min       -2.1961817

                    Extreme Observations

      ------Lowest------        ------Highest-----

          Value      Obs            Value      Obs

      -2.196182       21         0.857210       14
      -1.474012       12         0.981659       17
      -1.318441       19         1.115123        4
      -1.056147        3         1.645258       22
      -0.668478        8         2.323369       11

    Stem Leaf                     #             Boxplot
       2 3                        1                |
       1 6                        1                |
       1 01                       2                |
       0 889                      3             +-----+
       0 12224                    5             |  +  |
      -0 43332                    5             *-----*
      -0 765                      3             +-----+
      -1 31                       2                |
      -1 5                        1                |
      -2 2                        1                |
         ----+----+----+----+

                      Normal Probability Plot
    2.25+                                                *+++++
        |                                             *++++
        |                                        ++*++
    0.75+                                   +**+**
        |                              +*****
        |                         *****
   -0.75+                   *+**+
        |           *++*+*
        |      +++++
   -2.25+ +++++*
        +----+----+----+----+----+----+----+----+----+----+
            -2        -1         0        +1        +2
```

Remember - check your residuals first!!

Open and review the **rcbd_output.pdf document**.

Take a note of the means comparisons results - are they the same as when we ran the data using a CRD? Why or why not?

Generalized Linear Mixed Model (GLMM)

The last type of analysis I'd like to work through - GLMM. Think of this term as the umbrella term for all types of ANOVAs. Our computing power and statistical methodologies have caught up and now allow us to analyse most of the different types of data we collect. As an example - our field data, we've been working with the nitrogen variable. Have you noticed that these are continuous measures - data types we are comfortable with because we can take the mean and talk about variation around that mean. But, we don't only collect continuous measures. For example, we may have a variable that counts an object - weeds in our example.

Until recently, we would force these types of data to make them normal and then run our ANOVA. But no more! We can now let our program now, what distribution we believe our data has, test the residuals to make sure our model fits the data, and we're off to the races. Sounds easy? Well… as you can imagine, finding that right distribution can be challenging. There are other aspects of our model that we can modify now, that we couldn't before, but let's wade into this new forum with changing the distribution.

Let's start with the RCBD mixed model and we will be studying the **weed** variable.

Our statistical model is the same - so try it out and see what happens with the results. Remember to check the residuals before getting too excited about the results!

```
ods pdf file="glmm1_output.pdf";

Proc glimmix data=rcbd;
  class block trmt;
  model weed = trmt;
  random block;
  title "Proc GLIMMIX Results";
  lsmeans trmt / pdiff adjust=tukey ilink lines;
  output out=second predicted=pred residual=resid residual(noblup)=mresid stu
dent=studentresid student(noblup)=smresid;
Run;

/*  Linearity of fixed effects - both as a scatter and a boxplot */
Proc sgplot data=second;
  scatter y=smresid x=trmt;
  refline 0;
Run;

Proc sgplot data=second;
  vbox smresid / group=trmt datalabel;
Run;

/* Homogeneity of effects */
Proc sgscatter data=second;
  plot studentresid*(pred trmt block);
Run;
```

```
/* Q-Q plot and Shapiro-Wilk for normal distribution */
Proc univariate data=second normal plot;
  var studentresid;
Run;

ods pdf close;
```

Proc GLIMMIX Results

The GLIMMIX Procedure

Model Information

Data Set	WORK.RCBD
Response Variable	Weed
Response Distribution	Gaussian
Link Function	Identity
Variance Function	Default
Variance Matrix	Not blocked
Estimation Technique	Restricted Maximum Likelihood
Degrees of Freedom Method	Containment

Class Level Information

Class	Levels	Values
block	4	1 2 3 4
trmt	6	1 2 3 4 5 6

Number of Observations Read	24
Number of Observations Used	24

Dimensions

G-side Cov. Parameters	1
R-side Cov. Parameters	1
Columns in X	7
Columns in Z	4
Subjects (Blocks in V)	1
Max Obs per Subject	24

Optimization Information

Optimization Technique	Dual Quasi-Newton
Parameters in Optimization	1
Lower Boundaries	1
Upper Boundaries	0
Fixed Effects	Profiled

Residual Variance Starting From: Profiled Data

Iteration History

Iteration	Restarts	Evaluations	Objective Function	Change	Max Gradient
0	0	4	169.67391681	.	0

Convergence criterion (ABSGCONV=0.00001) satisfied.

Estimated G matrix is not positive definite.

Fit Statistics

-2 Res Log Likelihood	169.67
AIC (smaller is better)	171.67
AICC (smaller is better)	171.92
BIC (smaller is better)	171.06
CAIC (smaller is better)	172.06
HQIC (smaller is better)	170.33
Generalized Chi-Square	8239.75
Gener. Chi-Square / DF	457.76

Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error
block	0	.
Residual	457.76	152.59

Type III Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
trmt	5	15	8.98	0.0004

trmt Least Squares Means

trmt	Estimate	Standard Error	DF	t Value	Pr > |t|	Mean	Standard Error Mean
1	84.0000	10.6977	15	7.85	<.0001	84.0000	10.6977
2	89.5000	10.6977	15	8.37	<.0001	89.5000	10.6977

3	26.0000	10.6977	15	2.43	0.0281	26.0000	10.6977
4	28.0000	10.6977	15	2.62	0.0194	28.0000	10.6977
5	36.0000	10.6977	15	3.37	0.0043	36.0000	10.6977
6	14.7500	10.6977	15	1.38	0.1882	14.7500	10.6977

Differences of trmt Least Squares Means
Adjustment for Multiple Comparisons: Tukey

trmt	_trmt	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P
1	2	-5.5000	15.1288	15	-0.36	0.7213	0.9990
1	3	58.0000	15.1288	15	3.83	0.0016	0.0166
1	4	56.0000	15.1288	15	3.70	0.0021	0.0214
1	5	48.0000	15.1288	15	3.17	0.0063	0.0575
1	6	69.2500	15.1288	15	4.58	0.0004	0.0040
2	3	63.5000	15.1288	15	4.20	0.0008	0.0083
2	4	61.5000	15.1288	15	4.07	0.0010	0.0107
2	5	53.5000	15.1288	15	3.54	0.0030	0.0292
2	6	74.7500	15.1288	15	4.94	0.0002	0.0020
3	4	-2.0000	15.1288	15	-0.13	0.8966	1.0000
3	5	-10.0000	15.1288	15	-0.66	0.5186	0.9837
3	6	11.2500	15.1288	15	0.74	0.4686	0.9729
4	5	-8.0000	15.1288	15	-0.53	0.6047	0.9940
4	6	13.2500	15.1288	15	0.88	0.3949	0.9467
5	6	21.2500	15.1288	15	1.40	0.1805	0.7241

Tukey Grouping for trmt Least Squares Means (Alpha=0.05)

LS-means with the same letter are not significantly different.

trmt	Estimate		
2	89.5000		A
			A
1	84.0000	B	A
		B	
5	36.0000	B	C
			C
4	28.0000		C
			C
3	26.0000		C
			C
6	14.7500		C

Proc GLIMMIX Results

The UNIVARIATE Procedure
Variable: studentresid (Studentized Residual)

Moments

N	24	Sum Weights	24
Mean	0	Sum Observations	0
Std Deviation	1.02150784	Variance	1.04347826
Skewness	0.58893287	Kurtosis	0.35621279
Uncorrected SS	24	Corrected SS	24
Coeff Variation	.	Std Error Mean	0.20851441

Basic Statistical Measures

Location		Variability	
Mean	0.00000	Std Deviation	1.02151
Median	-0.17540	Variance	1.04348
Mode	-0.97145	Range	4.04771
		Interquartile Range	1.01193

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	0	Pr > |t|	1.0000
Sign	M	-3	Pr >= |M|	0.3075
Signed Rank	S	-17	Pr >= |S|	0.6372

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.946023	Pr < W	0.2218
Kolmogorov-Smirnov	D	0.156583	Pr > D	0.1299
Cramer-von Mises	W-Sq	0.101609	Pr > W-Sq	0.1011
Anderson-Darling	A-Sq	0.570407	Pr > A-Sq	0.1290

Quantiles (Definition 5)

Level	Quantile
100% Max	2.212751
99%	2.212751
95%	2.104812
90%	1.511147
75% Q3	0.425010
50% Median	-0.175401
25% Q1	-0.586919

```
                    10%           -0.971451
                    5%            -1.619086
                    1%            -1.834964
                    0% Min        -1.834964

                     Extreme Observations

          ------Lowest------        ------Highest-----

              Value      Obs             Value      Obs

          -1.834964       19          0.661127       23
          -1.619086       18          1.349238       17
          -0.971451       15          1.511147        9
          -0.971451       13          2.104812       20
          -0.809543       11          2.212751       16

        Stem Leaf                      #             Boxplot
           2 12                        2                0
           1 5                         1                |
           1 3                         1                |
           0 57                        2                |
           0 233                       3             +--+--+
          -0 43322221                  8             *-----*
          -0 865                       3             +-----+
          -1 00                        2                |
          -1 86                        2                |
             ----+----+----+----+

                         Normal Probability Plot
       2.25+                                          *   *+++++
           |                                       *  +++++
           |                                    *++++
           |                                ++*+*
       0.25+                           ++++***
           |                       ***** ***
           |              * **+**+
           |           +++++
      -1.75+      *+++*
           +----+----+----+----+----+----+----+----+----+----+
               -2        -1         0        +1        +2
```

What do you think?

Open and review the **glmm1_output.pdf document**.

GLMM - Poisson Distribution

Based on the residuals we saw above, there are a few issues. So, let's go back and think about our data. We have number of weeds counted in a plot. It's a COUNT. We are counting how many weeds we see. Yes - someone can make the argument that the COUNT can come from a normal distribution - but - what range do we have in our data? It is a very narrow range and our residuals didn't look the greatest so let's try a different distribution.

Traditionally for COUNT data you would use a Poisson Distribution.

So, our statistical model has not changed, only the type of data we are analyzing has changed. with PROC GLIMMIX we can add an option at the end of our model statement to let SAS know that we are now working with count data or data that comes from a POISSON distribution.

Let's try it out - but remember check those residuals!

```
ods pdf file="glmm2_poisson_output.pdf";
Proc glimmix data=rcbd method=laplace;
  class block trmt;
  model weed = trmt /dist=poisson link=log;
  random block;
  title "Proc GLIMMIX Results";
  lsmeans trmt / pdiff adjust=tukey ilink lines;
  output out=second predicted=pred residual=resid residual(noblup)=mresid stu
dent=studentresid student(noblup)=smresid;
Run;


/*  Linearity of fixed effects - both as a scatter and a boxplot */
Proc sgplot data=second;
  scatter y=studentresid x=trmt;
  refline 0;
Run;

Proc sgplot data=second;
  vbox studentresid / group=trmt datalabel;
Run;

/* Homogeneity of effects */
Proc sgscatter data=second;
  plot studentresid*(pred trmt block);
Run;

/* Q-Q plot and Shapiro-Wilk for normal distribution */
Proc univariate data=second normal plot;
  var studentresid;
Run;
```

```
ods pdf close;
```

Proc GLIMMIX Results

The GLIMMIX Procedure

Model Information

Data Set	WORK.RCBD
Response Variable	Weed
Response Distribution	Poisson
Link Function	Log
Variance Function	Default
Variance Matrix	Not blocked
Estimation Technique	Maximum Likelihood
Likelihood Approximation	Laplace
Degrees of Freedom Method	Containment

Class Level Information

Class	Levels	Values
block	4	1 2 3 4
trmt	6	1 2 3 4 5 6

Number of Observations Read	24
Number of Observations Used	24

Dimensions

G-side Cov. Parameters	1
Columns in X	7
Columns in Z	4
Subjects (Blocks in V)	1
Max Obs per Subject	24

Optimization Information

Optimization Technique	Dual Quasi-Newton
Parameters in Optimization	7
Lower Boundaries	1
Upper Boundaries	0
Fixed Effects	Not Profiled
Starting From	GLM estimates

Iteration History

		Objective		Max

Iteration	Restarts	Evaluations	Function	Change	Gradient
0	0	4	393.95171425	.	16.31912
1	0	3	393.92307395	0.02864029	4.878291
2	0	3	393.90913699	0.01393696	4.78523
3	0	4	393.90836371	0.00077328	4.652532
4	0	3	393.908229	0.00013471	4.528016
5	0	3	393.90762557	0.00060343	3.785906
6	0	3	393.90604612	0.00157945	1.584237
7	0	3	393.90579638	0.00024975	1.11238
8	0	2	393.905129	0.00066738	0.954611
9	0	3	393.90501786	0.00011114	0.054695
10	0	3	393.90501699	0.00000087	0.002244

Convergence criterion (GCONV=1E-8) satisfied.

Fit Statistics

-2 Log Likelihood	393.91
AIC (smaller is better)	407.91
AICC (smaller is better)	414.91
BIC (smaller is better)	403.61
CAIC (smaller is better)	410.61
HQIC (smaller is better)	398.48

Fit Statistics for Conditional Distribution

-2 log L(Weed | r. effects)	382.80
Pearson Chi-Square	236.50
Pearson Chi-Square / DF	9.85

Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error
block	0.02120	0.01760

Type III Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
trmt	5	15	76.42	<.0001

trmt Least Squares Means

trmt	Estimate	Standard Error	DF	t Value	Pr > |t|	Mean	Standard Error Mean
1	4.4203	0.09109	15	48.53	<.0001	83.1173	7.5709
2	4.4837	0.09008	15	49.78	<.0001	88.5595	7.9772
3	3.2475	0.1222	15	26.57	<.0001	25.7268	3.1441
4	3.3216	0.1194	15	27.83	<.0001	27.7057	3.3072
5	3.5730	0.1107	15	32.26	<.0001	35.6217	3.9450
6	2.6807	0.1492	15	17.96	<.0001	14.5949	2.1780

Differences of trmt Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

trmt	_trmt	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P
1	2	-0.06342	0.07596	15	-0.83	0.4168	0.9560
1	3	1.1727	0.1122	15	10.45	<.0001	<.0001
1	4	1.0986	0.1091	15	10.07	<.0001	<.0001
1	5	0.8473	0.09960	15	8.51	<.0001	<.0001
1	6	1.7396	0.1412	15	12.32	<.0001	<.0001
2	3	1.2361	0.1114	15	11.10	<.0001	<.0001
2	4	1.1620	0.1083	15	10.73	<.0001	<.0001
2	5	0.9107	0.09868	15	9.23	<.0001	<.0001
2	6	1.8030	0.1405	15	12.83	<.0001	<.0001
3	4	-0.07410	0.1362	15	-0.54	0.5943	0.9932
3	5	-0.3254	0.1287	15	-2.53	0.0231	0.1767
3	6	0.5669	0.1630	15	3.48	0.0034	0.0326
4	5	-0.2513	0.1260	15	-1.99	0.0646	0.3889
4	6	0.6410	0.1609	15	3.98	0.0012	0.0124
5	6	0.8923	0.1546	15	5.77	<.0001	0.0004

Tukey-Kramer Grouping for trmt Least Squares Means (Alpha=0.05)

LS-means with the same letter are not significantly different.

trmt	Estimate	
2	4.4837	A
		A
1	4.4203	A
5	3.5730	B
		B
4	3.3216	B
		B
3	3.2475	B

6	2.6807	C

Proc GLIMMIX Results

The UNIVARIATE Procedure
Variable: studentresid (Studentized Residual)

Moments

N	24	Sum Weights	24
Mean	-0.0200584	Sum Observations	-0.4814025
Std Deviation	3.97877608	Variance	15.8306591
Skewness	0.39291046	Kurtosis	-0.7715362
Uncorrected SS	364.114815	Corrected SS	364.105159
Coeff Variation	-19835.92	Std Error Mean	0.81216427

Basic Statistical Measures

Location		Variability	
Mean	-0.02006	Std Deviation	3.97878
Median	-0.88166	Variance	15.83066
Mode	.	Range	14.41473
		Interquartile Range	5.82008

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	-0.0247	Pr > |t|	0.9805
Sign	M	-1	Pr >= |M|	0.8388
Signed Rank	S	-9	Pr >= |S|	0.8032

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.96443	Pr < W	0.5336
Kolmogorov-Smirnov	D	0.146491	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.054972	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.316331	Pr > A-Sq	>0.2500

Quantiles (Definition 5)

Level	Quantile

```
                100% Max       8.068747
                99%            8.068747
                95%            6.595293
                90%            5.744689
                75% Q3         2.771780
                50% Median    -0.881662
                25% Q1        -3.048299
                10%           -4.683652
                5%            -5.399651
                1%            -6.345982
                0% Min        -6.345982

                 Extreme Observations

   ------Lowest-----        -----Highest-----

      Value      Obs           Value      Obs

   -6.34598       19         4.15645       17
   -5.39965       18         4.54562       23
   -4.68365       13         5.74469        9
   -4.13442       24         6.59529       20
   -3.48669       15         8.06875       16

Stem Leaf                     #             Boxplot
   8 1                        1                |
   6 6                        1                |
   4 257                      3                |
   2 441                      3             +-----+
   0 234                      3             |     |
  -0 8652                     4             *--+--*
  -2 52964                    5             +-----+
  -4 471                      3                |
  -6 3                        1                |
     ----+----+----+----+

                Normal Probability Plot
 9+                                                  * +++++
  |                                               *+++++
  |                                         **+*+++
  |                                     ***+++
 1+                                 +***+
  |                           +***+*
  |                      +*+****
  |           *++*+*
-7+     +*+++
   +----+----+----+----+----+----+----+----+----+----+
        -2        -1         0        +1        +2
```

Open and review the **glmm2_poisson_output.pdf document**.

What do you think? Better? NOtice that it can be extremely subjective. It will be up to you to decide whether you feel you can defend your choice or not.

Your turn!

If you happen to have some of your data with you - try this out. You should know what your exeprimental design is - from there create your statistical model - then trying running the analysis to partition your variance.

If you do not have your data with you, let me know and I can give you a practice RCBD dataset.

Workshop Review

ANOVA - History and updates

Fixed Effects ANOVA

Mixed Model ANOVA

GLMM