ANOVA

Multiple Regression with Qualitative Variables

Models

- Multiple Regression
  - Response: at least ordinal
  - Independent variables: at least ordinal
- ANOVA
  - Response: at least ordinal
  - Independent variables: nominal (qualitative)

Dummy Coding

- Multiple Regression:
  - With groups
  - Create a dummy variable(s) to code the groups
  - You can use 0's and 1's
  - Model:
    - $Y=b_o + b_1\ d1 + b_2\ d2 + e$

Comparing the Two Procedures

Proc Reg

```
data reg; input d1 d2 y @@; cards;
1 1 1 1 1 3
0 0 2 0 0 4
1 0 11 1 0 15
proc reg; model y= d1 d2;
test d1, d2; run;
```

Proc GLM

```
data anova; input group y @@;cards;
1 1 1 3
2 2 2 4
3 11 3 15
proc glm; class group; model y=group; run;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	148.00000	74.00000	18.50	0.0205
Error	3	12.00000	4.00000		
Corrected Total	5	160.00000			

Multiple Regression

Root MSE	2.00000	R-Square	0.9250
Dependent Mean	6.00000	Adj R-Sq	0.8750
Coeff Var	33.33333		

Multiple Regression

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.00000	1.41421	2.12	0.1240
d1	1	10.00000	2.00000	5.00	0.0154
d2	1	-11.00000	2.00000	-5.50	0.0118

Multiple Regression

Level of group	N	y Mean	y Std Dev
1	2	2.0000000	1.41421356
2	2	3.0000000	1.41421356
3	2	13.0000000	2.82842712

13-3=10 (g3 vs g2)

2-13=-11(g1 vs g3)

Test 1 Results for Dependent Variable y				
Source	DF	Mean Square	F Value	Pr > F
Numerator	2	74.00000	18.50	0.0205
Denominator	3	4.00000		

Multiple Regression

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	148.0000000	74.0000000	18.50	0.0205
Error	3	12.0000000	4.0000000		
Corrected Total	5	160.0000000			

ANOVA

R-Square	Coeff Var	Root MSE	y Mean
0.925000	33.33333	2.000000	6.000000

ANOVA

ANOVA Traditional Approach

- **A a procedure for testing the hypothesis that two or more population means are equal.**
- **For example:**

  H_0: $\mu_1 = \mu_2 = \mu_3 = \ldots \mu_k$

  H_1: At least one mean is different

ANOVA methods require the F-distribution

1. The F-distribution is not symmetric; it is skewed to the right.
2. The values of F can be 0 or positive, they cannot be negative.
3. There is a different F-distribution for each pair of degrees of freedom for the numerator and denominator.

One-Way ANOVA

Assumptions

1. The populations have normal distributions.
2. The populations have the same variance σ^2 (or standard deviation σ).
3. The samples are simple random samples.
4. The samples are independent of each other.

ANOVA Statistical Logic

Estimate the common value of σ^2 using

1. The variance between cells is an estimate of the common population variance σ^2 (the within variance) plus the variability among the sample means.

2. The variance within samples (also called variation due to error) is an estimate of the common population variance σ^2..

ANOVA Fundamental Concept

Test Statistic for One-Way ANOVA

ANOVA Fundamental Concept

Test Statistic for One-Way ANOVA

$$F = \frac{\text{variance between}}{\text{variance within}}$$

ANOVA Fundamental Concept

Test Statistic for One-Way ANOVA

$$F = \frac{\text{variance between samples}}{\text{variance within samples}}$$

A excessively large F test statistic is evidence against equal population means.

Calculations with Equal Sample Sizes

- Variance between samples = $ns_{\bar{x}}^2$

Calculations with Equal Sample Sizes

- Variance between samples = $ns_{\bar{x}}^2$

  where $s_{\bar{x}}^2$ = variance of samples means

Calculations with Equal Sample Sizes

- Variance between samples = $ns_{\bar{x}}^2$

  where $s_{\bar{x}}^2$ = variance of samples means

- Variance within samples = s_p^2

Critical Value of F

- Right-tailed test
- Degree of freedom with k samples of the same size n

  numerator df = k -1

  denominator df = k(n -1)

Sums of Squares Total

SS(total), or total sum of squares, is a measure of the total variation (around $\bar{\bar{x}}$) in all the sample data combined.

$$SS(total) = \Sigma\Sigma(x - \bar{\bar{x}})^2$$

Between Sums of Squares

SS(Between) is a measure of the variation between the samples.

$$SS(Bet) = \Sigma n_i(\bar{x}_i - \bar{\bar{x}})^2$$

Sums of Squares Error

SS(error) is a sum of squares representing the variability that is assumed to be common to all the populations being considered.

$$SS(error) = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2 \ldots n_k(x_k - 1)s_i^2$$

$$= \Sigma(n_i - 1)s_i^2$$

Sums of Squares

$$SS(total) = SS(Between) + SS(error)$$

Mean Squares (MS)

Sum of Squares SS(Between) and SS(error) divided by corresponding number of degrees of freedom.

MS (Between) is mean square for treatment, obtained as follows:

Mean Squares (MS)

Sum of Squares SS(Between) and SS(error) divided by corresponding number of degrees of freedom.

MS (Between) is mean square for treatment, obtained as follows:

$$MSB = \frac{SS\ (Between)}{k - 1}$$

Mean Squares (MS)

MS (error) is mean square for error, obtained as follows:

Mean Squares (MS)

MS (error) is mean square for error, obtained as follows:

$$MS\ (error) = \frac{SS\ (error)}{N - k}$$

SAS Setup: One way ANOVA

```
data wheat;
  input id variety yield moist;
  datalines;
1       1         41       10
2       1         69       57

proc glm data=wheat; class variety;
model yield = variety;
means variety /hovtest;
run;
```

- The HOVTEST=BARTLETT option specifies Bartlett's test (Bartlett 1937), a modification of the normal-theory likelihood ratio test.
- The HOVTEST=BF option specifies Brown and Forsythe's variation of Levene's test (Brown and Forsythe 1974). Seems to be the best out of all of these, good power and good control of Type(I) error. See Olejnik and Algina, 1987.
- The HOVTEST=LEVENE option specifies Levene's test (Levene 1960), which is widely considered to be the standard homogeneity of variance test. You can use the TYPE= option in parentheses to specify whether to use the absolute residuals (TYPE=ABS) or the squared residuals (TYPE=SQUARE) in Levene's test. TYPE=SQUARE is the default.
- The HOVTEST=OBRIEN option specifies O'Brien's test (O'Brien 1979), which is basically a modification of HOVTEST=LEVENE(TYPE=SQUARE). You can use the W= option in parentheses to tune the variable to match the suspected kurtosis of the underlying distribution. By default, W=0.5, as suggested by O'Brien (1979, 1981).

Results

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	4089.06666	454.340741	4.78	0.0001
Error	50	4756.33333	95.126667		
Corrected Total	59	8845.40000			

R-Square	Coeff Var	Root MSE	yield Mean
0.462	17.14	9.753	56.90

Level of variety	N	yield Mean	yield Std Dev
1	6	59.500	10.559
2	6	47.000	5.4405
3	6	60.000	11.045
4	6	50.833	8.0849
5	6	64.500	6.7156
6	6	63.000	14.546
7	6	39.666	10.984
8	6	57.166	10.515
9	6	58.833	10.361
10	6	68.500	5.244

Levene's Test for Homogeneity of yield Variance
ANOVA of Squared Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
variety	9	116051	12894.6	1.50	0.1747
Error	50	430418	8608.4		

Post-hoc Multiple Comparisons

```
data wheat;
   input id variety yield moist;
   datalines;
1       1          41       10
2       1          69       57

proc glm data=wheat; class variety;
model yield = variety;
means variety /hovtest;
means variety / lsd waller tukey regwq;
run;
```

LSD Procedure

Alpha	0.05
Error Degrees of Freedom	50
Error Mean Square	95.12667
Critical Value of t	2.00856
Least Significant Difference	11.31

Waller-Ducan Test

Kratio	100
Error Degrees of Freedom	50
Error Mean Square	95.12667
F Value	4.78
Critical Value of t	2.02803
Minimum Significant Difference	11.42

Means with the same letter are not significantly different.

Waller Grouping			Mean	N	variety
	A		68.500	6	10
	A				
	A		64.500	6	5
	A				
	A		63.000	6	6
	A				
B	A		60.000	6	3
B	A				
B	A		59.500	6	1
B	A				
B	A		58.833	6	9
B	A				
B	A	C	57.167	6	8
B		C			
B	D	C	50.833	6	4
	D	C			
	D	C	47.000	6	2
	D				
	D		39.667	6	7

Tukey's Procedure

Alpha	0.05
Error Degrees of Freedom	50
Error Mean Square	95.12667
Critical Value of Studentized Range	4.68144
Minimum Significant Difference	18.64

Means with the same letter are not significantly different.

REGWQ Grouping			Mean	N	variety
	A		68.500	6	10
	A				
B	A		64.500	6	5
B	A				
B	A		63.000	6	6
B	A				
B	A		60.000	6	3
B	A				
B	A		59.500	6	1
B	A				
B	A		58.833	6	9
B	A				
B	A		57.167	6	8
B	A				
B	A	C	50.833	6	4
B		C			
B		C	47.000	6	2
		C			
		C	39.667	6	7

Planned and Post-hoc Comparisons

- Post-hoc comparisons should be conducted when there are no specific hypotheses about the means.
  - Regwq (Ryan's)
- Planned comparisons should be conducted when there are specific hypotheses about the means.
  - Contrast statement
  - Estimate statement

Estimate Statement

```
data wheat;
   input id variety yield moist;
   datalines;
1       1         41       10
2       1         69       57

proc glm data=wheat; class variety;
model yield = variety;
means variety /hovtest;
means variety / lsd waller tukey regwq;
estimate 'one vs three' variety 1 0 -1 0 0 0 0 0 0 0;
run;
```

Estimate

Parameter	Estimate	Standard Error	t Value	Pr > |t|
one vs three	-0.50000000	5.63106463	-0.09	0.9296

Bonferroni Inequality

- The Bonferrroni inequality provides a way to control for the overall probability of Type One Error:

$$\alpha_s \leq P(I) \leq \sum \alpha_i$$

ANOVA Model

$$y_{ij} = \mu + \alpha_j + e_{ij}$$

Or

$$y_{ij} = \mu_{ij} + e_{ij}$$

Two-Way Analysis of Variance

- **Involves two nominal factors**
- **Partitions data into subcategories called cells**

Factors

- Factors can be:
  - Fixed: inference valid only to levels present in the design.
  - Random: inference valid to the population of levels.
- Factors can also be
  - Crossed
  - nested

Crossed Design

	A1	A2
B1		
B2		

We can test for interaction.

B is Nested in the A Factor

A1		A2	
B1	B2	B3	B4

We cannot test for Interaction

Model for a Two-way ANOVA

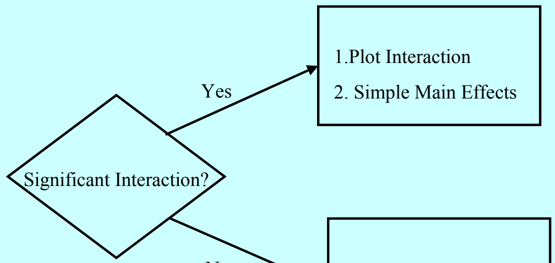
$$y_{ijk} = \mu + \alpha_j + \beta_k + \alpha\beta_{jk} + e_{ijk}$$

Assumptions

1. For each cell, the sample values come from a population with a distribution that is approximately normal.
2. The populations have the same variance.
3. The samples are simple random samples.

Definition

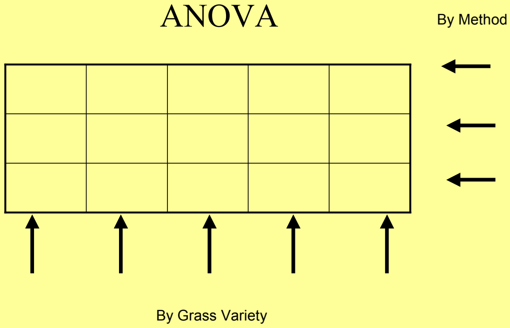
There is an interaction between two factors if the effect of one of the factors changes for different categories of the other factor.



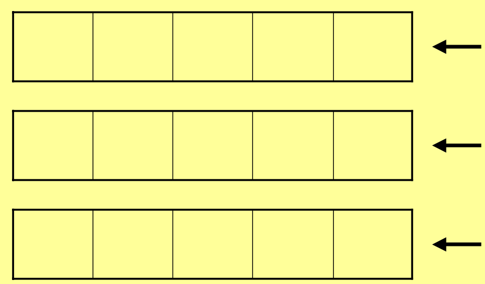
Simple Main Effects

- When the two-way interaction is significant we generally do not interpret the main effects.
- Instead of interpreting the main, we divide the two-way ANOVA into one-way ANOVA'S– these one-way anova's are called simple main effects.

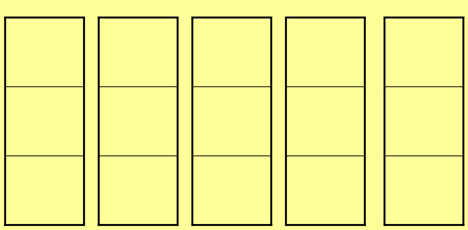
Consider the Grass by Method ANOVA



Simple Main Effects by Method



Simple Main Effects by Variety



SAS Setup: Two-way ANOVA d

(Data from Little, Stroup, Fruend, 2002)

```
• proc glm data=factorial;
• class method variety;
• model yield= method variety method*variety;
• run;
• proc means data=factorial noprint;
•  by method variety;
•  output out=factmean mean=yldmean;
•  run;
•  proc plot data=factmean;
•  plot yldmean*variety=method;run;
```

Interaction

Overall Model Results

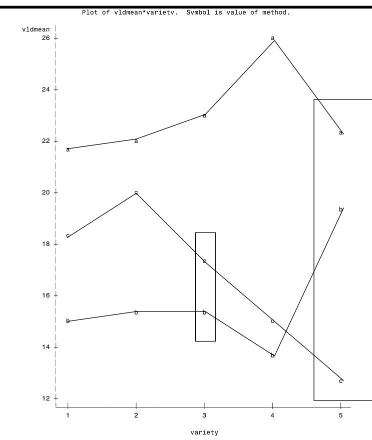
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	1339.024889	95.644635	4.87	<.0001
Error	75	1473.766667	19.650222		
Corrected Total	89	2812.791556			

Two-way ANOVA Results

Source	DF	Type I SS	Mean Square	F Value	Pr > F
method	2	953.1562222	476.5781111	24.25	<.0001
variety	4	11.3804444	2.8451111	0.14	0.9648
method*variety	8	374.4882222	46.8110278	2.38	0.0241

Effect Size

R-Square	Coeff Var	Root MSE	yield Mean
0.476048	24.04225	4.432857	18.43778



Using SAS's by Statement

- We can use the SAS's by statement to obtain simple main effects

  proc sort data=   ; by rows;
  proc glm; class column;
  model y=column;
  means column / regwq;
  by rows; **run**;

  Simple Main Effects

  The problem with using the 'by' statement to obtain simple main effects is that the error term is recomputed at each level.

Using the "by" statement to obtain simple main effect by variety

- **proc sort** data=factorial; by variety;
- **proc glm**; class method;
- model yield=method;
- means method / regwq;
- by variety; **run**;

Simple Main Effects

Not a very powerful approach

Homogeneity Test on the Two-Way ANOVA

```
proc glm data=factorial;
class trt;
model yield=trt;
means trt /hovtest=bf regwq;
run;
proc glm data=factorial;
class method variety;
model yield= method variety method*variety;
run;
```

One-way setup with all of the means

Treating the Two-way as a One-way

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	1339.0248	95.644635	4.87	.0001
Error	75	1473.7666	19.650222		
Corrected Total	89	2812.7915			

Brown and Forsythe's Test for Homogeneity of yield Variance ANOVA of Absolute Deviations from Group Medians					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
trt	14	92.4716	6.6051	0.88	0.5862
Error	75	565.0	7.5339		

Variety 1

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	134.00111	67.0005556	3.11	0.074
Error	15	323.25000	21.5500000		
Corrected Total	17	457.25111			

Means with the same letter are not significantly different.			
REGWQ Grouping	Mean	N	method
A	21.767	6	a
A			
A	18.417	6	c
A			
A	15.083	6	b

Variety 1

Simple Main Effects: Using the Slice Statement within Lsmeans

```
proc glm data=factorial;
class method variety;
model yield=method|variety;
means method variety / regwq;
/* lsmeans method*variety/slice=variety cl
  pdiff adjust=tukey; */
Lsmeans method*variety/slice=variety;
run;
```

method*variety Effect Sliced by variety for yield					
variety	DF	Sum of Squares	Mean Square	F Value	Pr > F
1	2	134.0011	67.000556	3.41	0.038
2	2	138.9033	69.451667	3.53	0.034
3	2	192.7033	96.351667	4.90	0.010
4	2	562.293333	281.146667	14.31	<.0001
5	2	299.743333	149.871667	7.63	0.0010

Table of Means

	V1	V2	V3	V4	v5
A1	21.76	21.85	23.13	25.96	22.33
A2	15.08	15.23	15.45	13.50	19.21
A3	18.41	19.91	17.31	14.83	12.55

Executing the Planned Comparisons: Two-way ANOVA

```
proc glm data=factorial;
class method variety;
model yield= method variety method*variety;
estimate 'method a vs methods b & c' method 2 -1 -1;
/* The ordering of the table is determine by the
   class statement */
estimate 'v1 vs v2 within a1' variety 1 -1 0 0 0
   method*variety 1 -1 0 0 0 0 0 0 0 0
0 0 0 0 0;
estimate 'a1 vs a3 within v1' method 1 0 -1
   method*variety 1 0 0 0 0 0 0 0 0 0
-1 0 0 0 0;
run;
```

Planned Comparisons

Parameter	Estimate	Standard Error	t Value	Pr > |t|
method a vs methods b & c	13.71666	1.982433	6.92	<.0001
v1 vs v2 within a1	-0.0833333	2.559311	-0.03	0.9741
a1 vs a3 within v1	3.350000	2.559311	1.31	0.1945

Special Case: One Observation Per Cell and No Interaction

- **When you have only one observation per cell the interaction effect can not be calculated.**
- **If it seems reasonable to assume (based on knowledge about the circumstances) that there is no interaction between the two factors, make the assumption and then proceed with a two-way anova with no interaction.**