

Statistics 514: Design of Experiments

Topic 6

Topic Overview

This topic will cover

- One-Factor Experiments
- Analysis of Variance
- Contrasts
- Random Effects

One Factor Experiments

Compare $a \geq 2$ group/treatment means.

One *factor* at a levels.

Examples: types of fertilizer, implant doses, heart drugs

Two Models

Cell mean model

$$Y_{i,j} = \mu_i + e_{i,j}, \quad j = 1, \dots, n_i, \quad i = 1, \dots, a$$

Treatment:	1	2	...	a
Mean:	μ_1	μ_2	...	μ_a

Treatment effect model

$$Y_{i,j} = \mu + \tau_i + e_{i,j}, \quad j = 1, \dots, n_i, \quad i = 1, \dots, a$$

μ is overall average

τ_i is “effect” of treatment i , $\sum_{i=1}^a \tau_i = 0$

$e_{i,j}$ taken independent $N(0, \sigma^2)$

Common Goals of Two Models

1. Estimate the μ_i or τ_i .
2. Test if they could be identical
3. Learn differences (and patterns therein) among μ_i

4. Pick “best” treatment
5. Estimate σ^2 (for next time)

Analysis of Variance

Balanced, $n_i = n$, case

$$\begin{aligned}\bar{y}_i &= \frac{1}{n} \sum_{j=1}^n y_{ij} && \text{(cell means)} \\ \bar{y}_{..} &= \frac{1}{an} \sum_{i=1}^a \sum_{j=1}^n y_{i,j} && \text{(grand mean)}\end{aligned}$$

Express $y_{i,j}$ as $\hat{\mu} + \hat{\tau}_i + \hat{\epsilon}_{i,j}$, where

$$\begin{aligned}\hat{\mu} &= \bar{y}_{..} \\ \hat{\tau}_i &= (\bar{y}_i - \bar{y}_{..}) \\ \hat{\epsilon}_{i,j} &= y_{i,j} - \bar{y}_i.\end{aligned}$$

Anova decomposition

$$\begin{aligned}SS_{TOT} &\equiv \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2 \\ &= \sum_{i=1}^a \sum_{j=1}^n (\bar{y}_i - \bar{y}_{..})^2 + \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_i)^2 \\ &= SS_{TRT} + SS_{ERR}\end{aligned}$$

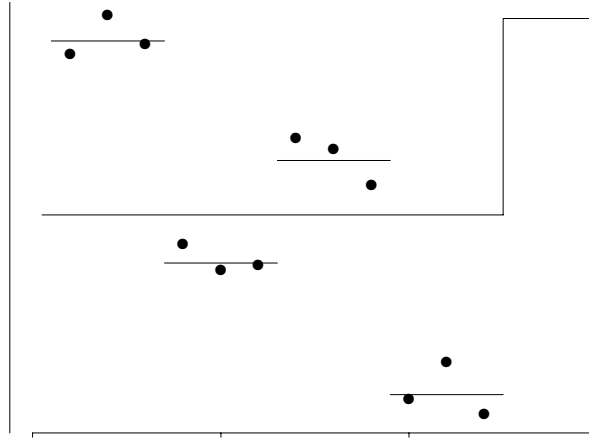
Proof

1. Substitute $y_{ij} - \bar{y}_{..} = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y}_{..})$.
2. Chug, chug, chug ...
3. Using $\sum_j (y_{ij} - \bar{y}_i) = 0$.

Physical Interpretation

- Unit masses
- Rotate with constant velocity around $\bar{y}_{..}$.
- Velocity $\propto |y_{ij} - \bar{y}_{..}|$
- Energy $\propto (y_{ij} - \bar{y}_{..})^2$

Variance as Energy



$SS_{TOT} \propto$ Total energy

$SS_{TRT} \propto$ kinetic energy: group means around grand mean

$SS_{ERR} \propto$ kinetic energy: points around group mean

Which energy is driving the crank?

Energy/variance and Pythagorean theorem analogies abound (use squares).

Hypotheses

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_a$$

$$\tau_1 = \tau_2 = \dots = \tau_a = 0$$

$$H_1 : \tau_i \neq 0 \text{ for at least one } i$$

Null Distributions

Under null hypothesis (and normality)

1. $SS_{TRT} \sim \sigma^2 \chi_{a-1}^2$

2. $SS_{ERR} \sim \sigma^2 \chi_{n(a-1)}^2$

3. $\bar{y}_{..} \sim N(\mu, \sigma^2/an)$

4. Above quantities are independent. (Cochran's Theorem)

5. $\bar{y}_{i.} \sim N(\mu_i = \mu, \sigma^2/n)$, independent

6. $SS_{TOT} \sim \sigma^2 \chi_{na-1}^2$

F statistic

$$F \equiv \frac{MS_{TRT}}{MS_{ERR}} \sim F_{a-1, n(a-1)},$$

where

$$MS_{TRT} \equiv SS_{TRT}/DF_{TRT} = SS_{TRT}/(a-1)$$

$$MS_{ERR} \equiv SS_{ERR}/DF_{ERR} = SS_{ERR}/(n(a-1))$$

More generally,

$$\begin{aligned} y_{i,j} - \bar{y}_i &= (\mu + \tau_i + e_{i,j}) - (\mu + \tau_i + \bar{e}_i) \\ &= e_{i,j} - \bar{e}_i. \text{ (Unaffected by } \tau_i) \\ \bar{y}_i - \bar{y}_.. &= (\mu + \tau_i + \bar{e}_i) - (\mu + \bar{\tau} + \bar{e}_..) \\ &= \tau_i + \bar{e}_i - \bar{e}_.. \end{aligned}$$

Non-null Expectations

$$\begin{aligned} E(SS_{TRT}) &= \sum_i \sum_j E((\tau_i + \bar{e}_i - \bar{e}_i)^2) \\ &= n \sum_i \tau_i^2 + n \sum_i E((\bar{e}_i - \bar{e}_i)^2) \\ &= n \sum_i \tau_i^2 + (a-1)\sigma^2 \\ E(MS_{TRT}) &= \frac{n}{a-1} \sum_i \tau_i^2 + \sigma^2 \\ &= n \sigma_{TRT}^2 + \sigma^2 \end{aligned}$$

Expected Mean Squares

1. $E(MS_{TRT}) = n \left(\frac{1}{a-1} \sum_{i=1}^a \tau_i^2 \right) + \sigma^2$
2. But $E(MS_{ERR}) = \sigma^2$.
3. So $\tau_i \neq 0$ increases $F = MS_{TRT}/MS_{ERR}$ (otherwise, ratio is close to 1).
4. This is why reject null for **large** F .
5. Note role of n . (How does n affect the power?)

Distributions

1. $\frac{SS_{TRT}}{\sigma^2} \sim \chi_{a-1}^2 \left(\frac{n}{\sigma^2} \sum_{i=1}^a \tau_i^2 \right)$ (noncentral)
2. So $F \sim F'_{a-1, N-a} \left(\frac{n}{\sigma^2} \sum_{i=1}^a \tau_i^2 \right)$.

Anova table

Source	D. F.	Sum Sq.	Mean Sq.	F	p
Treatment	$a - 1$	SS_{TRT}	MS_{TRT}	F	p
Error	$N - a$	SS_{ERR}	MS_{ERR}		
Total	$N - 1$	SS_{TOT}			

Inference

1. Inference not too sensitive to normality assumption.
2. Balance ($n_i = n$) mitigates $\sigma_i^2 \neq \sigma_j^2$
3. Randomization mitigates correlations, trends
4. F test approximates randomization/permutation inference

Similarity with t -test

- Consider the square of the t -test statistic

$$\begin{aligned}
 t_0^2 &= \left(\frac{\bar{y}_1. - \bar{y}_2.}{S_p \sqrt{2/n}} \right)^2 \\
 &= \left(\frac{(\bar{y}_1. - \bar{y}_{..}) - (\bar{y}_2. - \bar{y}_{..})}{S_p \sqrt{2/n}} \right)^2 \\
 &= \left(\frac{2(\bar{y}_1. - \bar{y}_{..})}{S_p \sqrt{2/n}} \right)^2 \\
 &= \left(\frac{4(\bar{y}_1. - \bar{y}_{..})^2}{S_p^2(2/n)} \right) \\
 &= \frac{2((\bar{y}_1. - \bar{y}_{..})^2 + (\bar{y}_2. - \bar{y}_{..})^2)}{S_p^2(2/n)} \\
 &= \frac{n((\bar{y}_1. - \bar{y}_{..})^2 + (\bar{y}_2. - \bar{y}_{..})^2)}{S_p^2} \\
 &= \frac{MS(\text{Between})}{MS(\text{Within})} = \frac{MS_{Treatment}}{MS_E}
 \end{aligned}$$

- When $a = 2$, $t_0^2 = F_0$; gives identical results as t -test $H_A : \neq$.

Unbalanced Case

1. SS_{TRT} and SS_{ERR} defined as before.
2. Still take $N = \sum_{i=1}^a n_i - 1$ and $\sum_i \tau_i = 0$.

3. Now $DF_{ERR} = \sum_{i=1}^a n_i - 1 = N - a$

4. $SS_{TRT} \sim \sigma^2 \chi_{a-1}^2 (\sum_i n_i \tau_i^2 / \sigma^2)$

Example of Unbalanced Design

Twelve lambs are randomly assigned to three different diets. The weight gain (over two weeks) is recorded. Is there a difference among the diets?

Diet 1	Diet 2	Diet 3
8	9	15
16	16	10
9	21	17
	11	6
	18	

$$\bar{y}_{..} = 156/12 \text{ and } \bar{y}_{..}^2 = 2274/12$$

$$\bar{y}_{1.} = 33/3, \bar{y}_{2.} = 75/5, \bar{y}_{3.} = 48/4$$

$$n_1 = 3, n_2 = 5, n_3 = 4, N_{Trt} = 3 \text{ and } N = 12$$

$$SS_{Tot} = N\bar{y}_{..}^2 - N(\bar{y}_{..})^2 = 2274 - 156^2/12 = 246$$

$$SS_{Trt} = (n_1(\bar{y}_{1.})^2 + n_2(\bar{y}_{2.})^2 + n_3(\bar{y}_{3.})^2) - N(\bar{y}_{..})^2 = (33^2/3 + 75^2/5 + 48^2/4) - 156^2/12 = 36$$

$$SS_E = SS_{Tot} - SS_{Trt} = 246 - 36 = 210 \quad F_0 = \frac{MS_{Trt}}{MS_E} = \frac{\left(\frac{SS_{Trt}}{N_{Trt}-1}\right)}{\left(\frac{SS_E}{N-N_{Trt}}\right)} = (36/2)/(210/9) = 0.77$$

$p\text{-value} \geq 0.20$ (DNR)

Using SAS (lambs.sas)

```
option nocenter ps=65 ls=80;
```

```
data lambs;
  input diet wtgain;
  cards;
  1 8
  1 16
  1 9
  2 9
  2 16
  2 21
  2 11
  2 18
  3 15
  3 10
  3 17
  3 6
  ;
```

```
symbol1 bwidth=5 i=box;
axis1 offset=(5);
proc gplot;
  plot wtgain*diet / frame haxis=axis1;
run;
```

```
proc glm;
  class diet;
  model wtgain=diet;
  output r=res p=pred;
run;
```

```
proc gplot;
  plot res*diet /frame haxis=axis1;
run;
```

```
proc sort; by pred;
symbol1 v=circle i=sm50;
```

```
proc gplot;
  plot res*pred / haxis=axis1;
run;
```

Log File (.log)

```
1  option nocenter ps=65 ls=80;
2
3  data lambs;
4  input diet wtgain;
5  cards;
```

NOTE: The data set WORK.LAMBS has 12 observations and 2 variables.

NOTE: DATA statement used:

real time	0.03 seconds
cpu time	0.03 seconds

```
18 ;
19
20 symbol1 bwidth=5 i=box;
21 axis1 offset=(5);
22 proc gplot;
23   plot wtgain*diet / frame haxis=axis1;
24 run;
```

25

NOTE: There were 12 observations read from the data set WORK.LAMBS.

NOTE: PROCEDURE Gplot used:

real time	0.20 seconds
-----------	--------------

cpu time 0.01 seconds

Output File (.lst)

The GLM Procedure

Class Level Information

Class	Levels	Values
diet	3	1 2 3

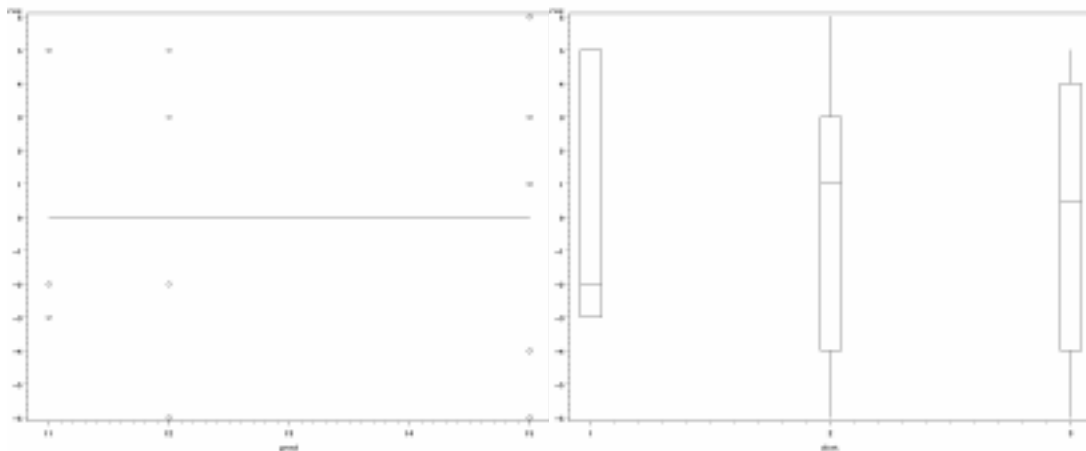
Number of observations 12

The GLM Procedure

Dependent Variable: wtgain

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	36.0000000	18.0000000	0.77	0.4907
Error	9	210.0000000	23.3333333		
Corrected Total	11	246.0000000			

R-Square	Coeff Var	Root MSE	wtgain Mean
0.146341	37.15738	4.830459	13.00000



Another SAS Example (tensile.sas)

```
options ls=75 ps=60 nocenter;  
options colors=(none) device=win target=winprtm rotate=landscape ftext=swiss  
      hsize=8.0in vsize=6.0in htext=1.5 htitle=1.5 hpos=60 vpos=60  
      horigin=0.5in vorigin=0.5in;
```

```
data one;
```

```

infile 'c:\saswork\data\tensile.dat';
input percent strength time;

title1 'Chapter 3 Example';
proc print data=one; run;

symbol1 v=circle i=none;
title1 'Plot of Strength vs Percent Blend';
proc gplot data=one; plot strength*percent/frame; run;

proc boxplot;
plot strength*percent/boxstyle=skeletal pctldef=4;

proc glm;
class percent; model strength=percent;
output out=oneres p=pred r=res; run;

proc sort; by pred;
symbol1 v=circle i=sm50; title1 'Residual Plot';
proc gplot; plot res*pred/frame; run;

proc univariate data=oneres pctldef=4;
var res; qqplot res / normal (L=1 mu=est sigma=est);
histogram res / normal; run;

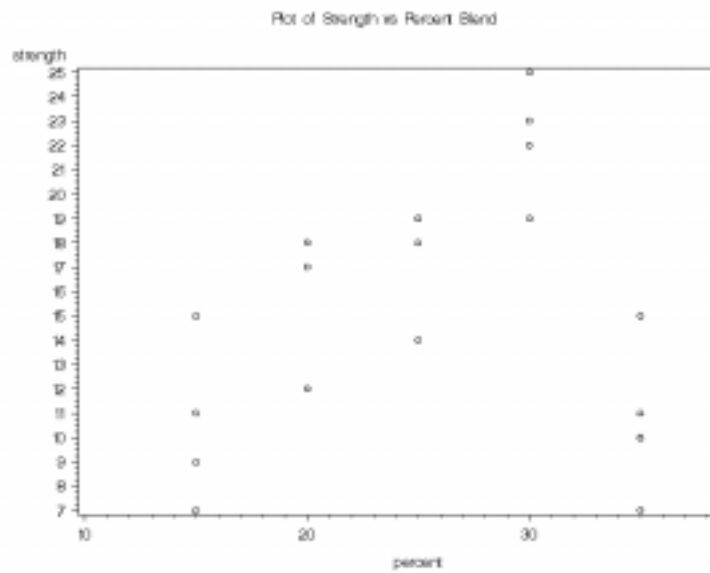
symbol1 v=circle i=none;
title1 'Plot of residuals vs time';
proc gplot; plot res*time / vref=0 vaxis=-6 to 6 by 1;
run;

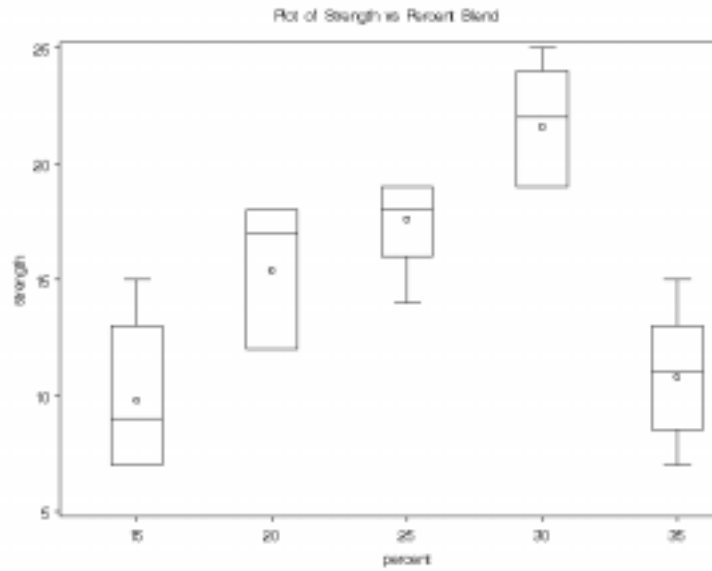
```

Chapter 3 Example

Obs	percent	strength	time
1	15	7	15
2	15	7	19
3	15	15	25
4	15	11	12
5	15	9	6
6	20	12	8
7	20	17	14
8	20	12	1
9	20	18	11
10	20	18	3
11	25	14	18
12	25	18	13
13	25	18	20
14	25	19	7
15	25	19	9
16	30	19	22

17	30	25	5
18	30	22	2
19	30	19	24
20	30	23	10
21	35	7	17
22	35	10	21
23	35	11	4
24	35	15	16
25	35	11	23





The GLM Procedure

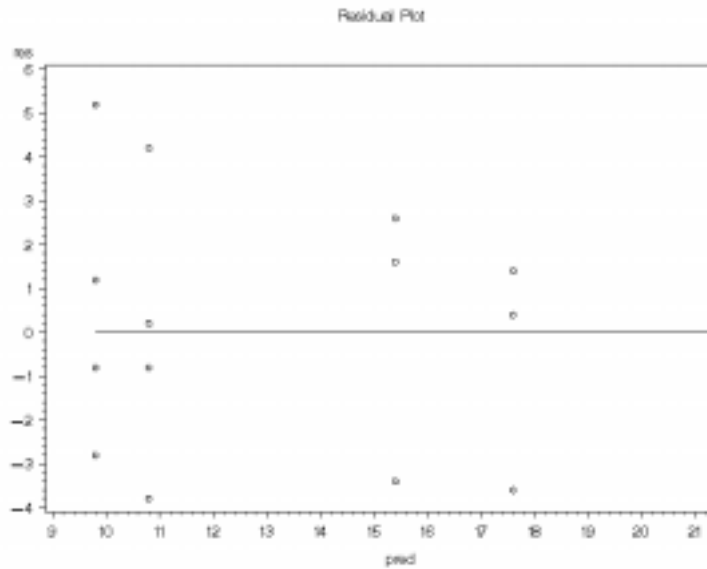
Dependent Variable: strength

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	475.7600000	118.9400000	14.76	<.0001
Error	20	161.2000000	8.0600000		
Corrected Total	24	636.9600000			

R-Square Coeff Var Root MSE strength Mean
 0.746923 18.87642 2.839014 15.04000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
percent	4	475.7600000	118.9400000	14.76	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
percent	4	475.7600000	118.9400000	14.76	<.0001



The UNIVARIATE Procedure
Variable: res

Moments

N	25	Sum Weights	25
Mean	0	Sum Observations	0
Std Deviation	2.59165327	Variance	6.71666667
Skewness	0.11239681	Kurtosis	-0.8683604
Uncorrected SS	161.2	Corrected SS	161.2
Coeff Variation	.	Std Error Mean	0.51833065

Basic Statistical Measures

Location		Variability	
Mean	0.00000	Std Deviation	2.59165
Median	0.40000	Variance	6.71667
Mode	-3.40000	Range	9.00000
		Interquartile Range	4.20000

NOTE: The mode displayed is the smallest of 7 modes with a count of 2.

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value-----
Student's t	t	0	Pr > t 1.0000
Sign	M	2.5	Pr >= M 0.4244
Signed Rank	S	0.5	Pr >= S 0.9896

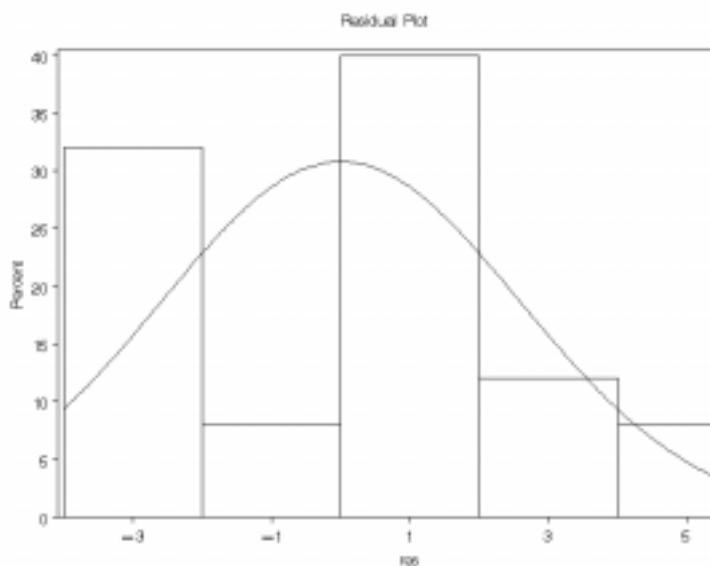
Fitted Distribution for res

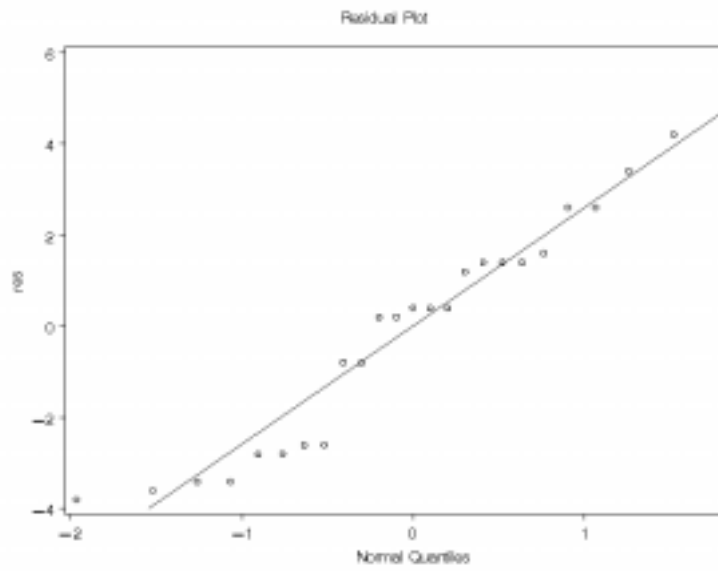
Parameters for Normal Distribution

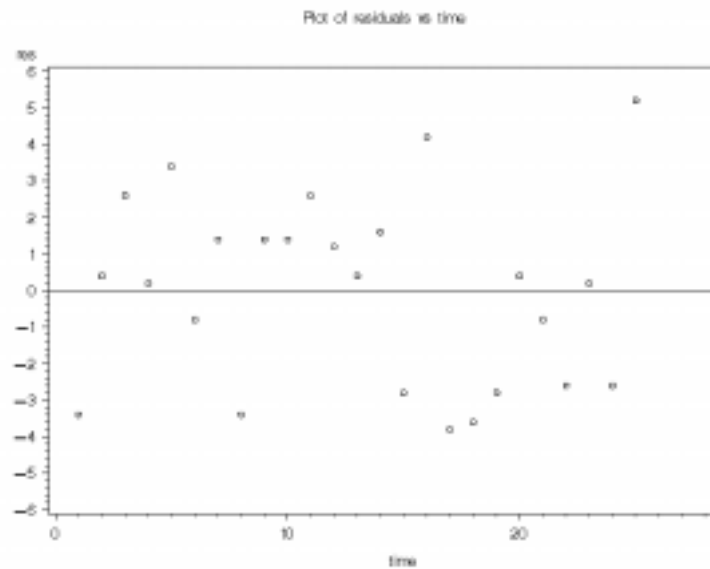
Parameter	Symbol	Estimate
Mean	Mu	0
Std Dev	Sigma	2.591653

Goodness-of-Fit Tests for Normal Distribution

Test	---Statistic---	-----p Value-----
Kolmogorov-Smirnov	D 0.16212279	Pr > D 0.088
Cramer-von Mises	W-Sq 0.08045523	Pr > W-Sq 0.203
Anderson-Darling	A-Sq 0.51857191	Pr > A-Sq 0.177







Regression and ANOVA

$\hat{\mu}_i$, $\hat{\tau}_i$ are least squares (LS) estimates.

ANOVA Representation with 3 levels

$$y_{i,j} = \mu + \tau_i + e_{i,j} \begin{cases} i = 1, 2, 3 \\ j = 1, \dots, n \end{cases}$$

Equivalent regression model:

$$y_{i,j} = \beta_0 + \beta_1 x_{1,j} + \beta_2 x_{2,j} + \epsilon_{i,j} \begin{cases} i = 1, 2, 3 \\ j = 1, \dots, n \end{cases}$$

$$x_{1,j} = I(\text{obs } j \text{ from trt 1}) = \begin{cases} 1 & \text{obj } j \text{ from trt 1} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{2,j} = I(\text{obs } j \text{ from trt 2})$$

Relationship to ANOVA parameters

$$\begin{aligned}\beta_0 &= \mu_3 \\ \beta_1 &= \mu_1 - \mu_3 \\ \beta_2 &= \mu_2 - \mu_3\end{aligned}$$

If a treatments $\Rightarrow a - 1$ regressor variables

$$y_{ij} = \beta_0 + \beta_1 x_{1,j} + \beta_2 x_{2,j} + \dots + \beta_{a-1} x_{a-1,j} + \epsilon_{i,j} \begin{cases} i = 1, \dots, a \\ j = 1, \dots, n \end{cases}$$
$$x_{i,j} = I(\text{obs } i \text{ from trt } j)$$

$$\begin{aligned}\beta_0 &= \mu_a \\ \beta_i &= \mu_i - \mu_a, \quad i = 1, 2, \dots, a - 1\end{aligned}$$

Hypothesis Tests

$$\begin{aligned}H_0 : \quad \beta_0 &= \mu_a \\ \beta_i &= 0, \quad i = 1, \dots, a - 1\end{aligned}$$

Regression test procedure $\iff F$ -test in one-way ANOVA.

Contrasts

- Linear combination of coeff's that sum to zero
- $\Gamma = \sum c_i \mu_i$ with $\sum c_i = 0$
 - To compare Trt 1 with Trt 2: $c_i = \{1, -1, 0, \dots, 0\}$
 - To compare Trt 1 vs $0.5(\text{Trt2} + \text{Trt3})$: $c_i = \{1, -0.5, -0.5, 0, \dots, 0\}$
- Will estimate Γ using $C = \sum c_i \bar{y}_i$.
- Notice estimate uses trt **means**, not totals.
- Under H_0 : $t_0 = C / \sqrt{\text{Var}(C)} \sim t_{N-a}$.
- Also $t_{N-a}^2 = F_{1, N-a}$, so could present as F -test.

Contrasts often presented in terms of Sum of Squares.

$$SS_C = (\sum c_i \bar{y}_i)^2 / \sum (c_i^2 / n_i)$$

If divide by MS_E , simply t_0^2
Can then compare to $F_{1, N-a}$.

Orthogonal Contrasts

- Suppose you have two contrasts $\{c_i\}$ and $\{d_i\}$.
- **Orthogonal** if $\sum c_i d_i / n_i = 0$ (using trt means)
- Can divide up SS_{Trt} into $a - 1$ orthogonal contrasts.
- By Cochran's Theorem \rightarrow comparisons independent

Orthogonal Polynomial Model

- Treatments are quantitative
- General polynomial model to fit trend

$$f(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3$$

Regression can be used to get estimates for a_1 , a_2 , and a_3 .

- Suppose want to test whether that trend follows in data.
- Orthogonal polynomial (as connected to orthogonal contrasts)

$$f(t) = \beta_0 + \beta_1 P_1(t) + \beta_2 P_2(t) + \beta_3 P_3(t)$$

where $P_1(t)$, $P_2(t)$, and $P_3(t)$ are pre-specified polynomials of order 1, 2, and 3, respectively. $P_1(t)$ is linear, $P_2(t)$ is quadratic, and $P_3(t)$ is cubic.

coefficients	t_1	t_2	\dots	t_a
Γ_1	$P_1(t_1)$	$P_1(t_2)$	\dots	$P_1(t_a)$
Γ_2	$P_2(t_1)$	$P_2(t_2)$	\dots	$P_2(t_a)$
Γ_3	$P_3(t_1)$	$P_3(t_2)$	\dots	$P_3(t_a)$

Algorithm

Function points: t_1, \dots, t_a

1. List functions in order: x, x^2, x^3, \dots, x^k .
2. Compute $\{(y_{1,0}, \dots, y_{a,0}), \dots, (y_{1,k}, \dots, y_{a,k})\}$, where $y_{i,j} = t_i^j$.
3. Center them (to make them contrasts)

$$y'_{i,j} = y_{i,j} - \frac{1}{a} \sum_i y_{i,j}$$

4. Orthogonalize them to get $y''_{i,j}$
 - Possible algorithms: Gram-Schmidt, etc.
 - Do in order from linear to last.
 - Want orthogonality/independence. (Ex: quartic significance should be independent of quadratic significant.)

Example

- Treatment levels t_k : 15, 20, 25, 30, 35
- Orthogonal polynomials ($x = (t - 25)/5$)

$$\begin{aligned}P_1(t) &= x \\P_2(t) &= x^2 - 2 \\P_3(t) &= 5/6(x^3 - 17x/5)\end{aligned}$$

t	15	20	25	30	35	
$P_1(t)$	-2	-1	0	1	2	linear
$P_2(t)$	2	-1	-2	-1	2	quadratic
$P_3(t)$	-1	2	0	-2	1	cubic

(for equally spaced t , more examples at Table IX in Montgomery)

If significant, can use $P_i(t)$ to predict $f(t)$ when t is not a treatment in the experiment. (Since functions are ordered, all functions up to highest significant order are included in predictive model.)

Determining Orthogonal Polynomial Coefficients Using SAS

- Often the levels of the treatment are not equally spaced
- Can use `proc iml` to determine coefficients

```
proc iml;
levels = {1 2 5 10 20}; /* Consider these 5 levels */
print levels;
coef = ORPOL(levels, 3); /* Gives coefs up through cubic */
coef = t(coef); /* Puts coefs in rows instead of cols */
coef = coef[2:4,]; /* Eliminates first row of coef matrix */
print coef; /* 1st row linear, 2nd quadratic, 3rd cubic */
run
```

```
-----
                                LEVELS
                                1      2      5      10      20
                                COEF
-0.424967 -0.360578 -0.167411 0.1545335 0.798423
0.4348974 0.2072899 -0.325207 -0.711616 0.3946361
-0.433125 0.1365799 0.7252914 -0.510844 0.0820972
```

Punchline

Testing

$$\begin{aligned}H_0 : & \quad \mu_1 = \mu_2 = \dots = \mu_a \\H_1 : & \quad \mu_i \neq \mu_j \text{ for some } (i, j)\end{aligned}$$

Equivalent to testing

$$\begin{aligned}H_0 : & \quad c_{1,1}\mu_1 + c_{2,1}\mu_2 + \dots + c_{a,1}\mu_a = 0 \\& \quad c_{1,2}\mu_1 + c_{2,2}\mu_2 + \dots + c_{a,2}\mu_a = 0 \\& \quad \vdots \\& \quad c_{1,a-1}\mu_1 + c_{2,a-1}\mu_2 + \dots + c_{a,a-1}\mu_a = 0 \\H_1 : & \quad c_{1,k}\mu_1 + c_{2,k}\mu_2 + \dots + c_{a,k}\mu_k \neq 0 \text{ for some } k\end{aligned}$$

for any set of orthogonal contrasts $\{c_{1,1}, \dots, c_{a,1}\}, \dots, \{c_{1,a-1}, \dots, c_{a,a-1}\}$.

Note that

$$H_0 : \quad \mu_1 = \mu_2 = \dots = \mu_a$$

is obviously the same as

$$\begin{aligned}H_0 : & \quad \mu_1 - \mu_2 = 0 \\& \quad \mu_2 - \mu_3 = 0 \\& \quad \vdots \\& \quad \mu_{a-1} - \mu_a = 0\end{aligned}$$

Poser

What if we're interested in just a few non-orthogonal contrasts?

Random Effects vs Fixed Effects (Section 13-1)

- Consider factor with numerous possible levels
- Want to draw inference **on population of levels** of which samples drawn are tiny fraction
- Not concerned with any specific levels
- Example of difference
 - (**Fixed**) Compare reading ability of 10 second grade classes in NY

Select $a = 10$ specific classes of interest. Randomly choose n students from each classroom. Want to compare τ_i (class-specific effects).

– (**Random**) Study the variability **among all** second grade class in NY

Randomly choose $a = 10$ classes from large number of classes. Randomly choose n students from each classroom. Want to assess σ_τ^2 (class-to-class variability).

- Inference broader in random effects case (applies to cases not seen before)
- Levels chosen randomly \rightarrow inference on population

Another, More Precise Definition: Future Prediction

Only interested in whether that factor has an effect; not worried about (in other words, can't) predict the effect for a treatment level we haven't seen before.

- If we will see *just one* (or more) factor levels that we haven't seen before, we want to quantify if the factor level is going to have an effect.
- **Example (continued)**: Want to know how much of an effect class will have on experimental variability in trying to predict reading ability on the *11th* classroom.
- Criticism: inference based analyses are hard to interpret usefully. (Large p -value: predicting average reading ability in the 11th classroom can be extrapolated using grand average?)
- Common trick question: what the variance of an observation from a model with a fixed factor.

Model

- Same model as in the fixed case

$$y_{i,j} = \mu + \tau_i + e_{i,j} \begin{cases} i = 1, 2, \dots, a \\ j = 1, 2, \dots, n_i \end{cases}$$

μ – grand mean

τ_i – i th treatment effect

$e_{i,j} \sim N(0, \sigma^2)$

But view number of treatment levels as infinite.

- Instead of $\sum \tau_i = 0$, assume

$\tau_i \sim N(0, \sigma_\tau^2)$

$\{\tau_i\}$ and $\{e_{i,j}\}$ independent

Inference

- $\text{Var}(y_{i,j}) = \sigma_\tau^2 + \sigma^2$
 σ^2 and σ_τ^2 are called *variance components*.
- The hypotheses:

$$H_0 : \sigma_\tau^2 = 0 \text{ vs. } H_1 : \sigma_\tau^2 > 0$$

- Identical ANOVA table:

Source	SS	DF	MS	F_0
Between	SS_{Trt}	$a - 1$	MS_{Trt}	$F_0 = \frac{MS_{Trt}}{MS_E}$
Within	SS_E	$N - a$	MS_E	
Total	$SS_{Tot} = SS_{Trt} + SS_E$	$N - 1$		

- $E(MS_E) = \sigma^2$
- $E(MS_{Trt}) = \sigma^2 + n\sigma_\tau^2$

- Under H_0 , $F_0 \sim F_{a-1, N-a}$
- Same test as before (between vs within variability)
- Conclusions, however, pertain to entire population. Trade-off: lose predictive power.

Model Estimates

- Usually interested in estimating variances
- Use mean squares (known as ANOVA method)

$$\hat{\sigma}^2 = MS_E$$

$$\hat{\sigma}_\tau^2 = (MS_{Trt} - MS_E)/n$$

If unbalanced, replace n with

$$n_0 = \frac{1}{a-1} \left(\sum_{i=1}^a n_i - \frac{\sum_{i=1}^a n_i^2}{\sum_{i=1}^a n_i} \right)$$

- Estimate of σ_τ^2 can be negative.
 - Support H_0 ? Use zero as estimate (lose unbiasedness)? (Can negative estimate of σ_τ^2 yield significant F -test?)
 - Validity of model? Nonlinear?
 - Other approaches (MLE, Bayesian with nonnegative prior, MIVQUE: minimum variance quadratic unbiased estimator, etc.)

Confidence Intervals

σ^2

Same as fixed case

$$\frac{(N-a)MS_E}{\sigma^2} \sim \chi_{N-a}^2$$
$$\frac{(N-a)MS_E}{\chi_{\alpha/2, N-a}^2} \leq \sigma^2 \leq \frac{(N-a)MS_E}{\chi_{1-\alpha/2, N-a}^2}$$

σ_τ^2

Linear combination of χ^2

$$\frac{(a-1)MS_{Trt}}{\sigma^2 + n\sigma_\tau^2} \sim \chi_{a-1}^2$$
$$\hat{\sigma}_\tau^2 = (MS_{Trt} - MS_E)/n$$

so

$$\hat{\sigma}_\tau^2 \sim \frac{\sigma^2 + n\sigma_\tau^2}{n(a-1)} \chi_{a-1}^2 - \frac{\sigma^2}{n(N-a)} \chi_{N-a}^2$$

No closed form expressions for this distribution.
Approximations available (Section 13-7)

Proportion of σ_τ^2 in $\text{Var}(y_{i,j})$

Common estimate if goal is to reduce variance

- We have that

$$\frac{MS_{Trt}/(\sigma^2 + n\sigma_\tau^2)}{MS_E/\sigma^2} \sim F_{a-1, N-a}$$

- For σ_τ^2/σ^2 :

$$L \leq \frac{\sigma_\tau^2}{\sigma^2} \leq U,$$

where

$$L = \frac{1}{n} \left(\frac{MS_{Trt}}{MS_E F_{\alpha/2, a-1, N-a}} - 1 \right)$$
$$U = \frac{1}{n} \left(\frac{MS_{Trt}}{MS_E F_{1-\alpha/2, a-1, N-a}} - 1 \right)$$

- Intraclass correlation:

$$\text{Corr}(y_{i,j}, y_{k,\ell}) = \begin{cases} 0 & i \neq k \\ \sigma_\tau^2 / (\sigma^2 + \sigma_\tau^2) & \text{for } i = k \text{ and } j \neq \ell \\ 1 & i = k \text{ and } j = \ell \end{cases}$$

Consider

$$\frac{L}{L+1} \leq \frac{\sigma_\tau^2}{\sigma^2 + \sigma_\tau^2} \leq \frac{U}{U+1}$$

Becomes

$$\frac{F_0 - F_{\alpha/2, a-1, N-a}}{F_0 + (n-1)F_{\alpha/2, a-1, N-a}} \leq \frac{\sigma_\tau^2}{\sigma^2 + \sigma_\tau^2} \leq \frac{F_0 - F_{1-\alpha/2, a-1, N-a}}{F_0 + (n-1)F_{1-\alpha/2, a-1, N-a}}$$

Example

A supplier delivers several hundred batches of raw material to a company each year. The company is interested in a high yield from each batch of raw material (percent usable). Therefore, to investigate the consistency of this supplier, an experiment is done where five batches were selected at random and three yield determinations were made on each batch.

Batch					Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
1	2	3	4	5					
74	68	75	72	79	Between	147.74	4	36.93	20.5
76	71	77	74	81	Within	18.00	10	1.80	
75	72	77	73	79	Total	165.73	14		

Highly significant result ($F_{0.05, 4, 10} = 3.47$)

$$\hat{\sigma}_\tau^2 = (36.93 - 1.80)/3 = 11.71$$

86.7% (= 11.71/(11.71 + 1.80)) is attributable to batch differences.

Conclusion: Time to improve consistency of the batches.

Confidence Intervals

- 95% CI for σ^2

$$\begin{aligned} \frac{SS_E}{\chi_{0.025, 10}^2} \leq \sigma^2 \leq \frac{SS_E}{\chi_{0.975, 10}^2} &= (18.00/20.48, 18.00/3.25) \\ &= (0.879, 5.538) \end{aligned}$$

- 95% CI for Intraclass Correlation

$$\left(\frac{20.52-4.47}{20.52+(3-1)4.47}, \frac{20.52-1/8.84}{20.52+(3-1)(1/8.84)} \right) \\ (0.545, 0.984)$$

using property that

$$F_{1-\alpha/2, \nu_1, \nu_2} = 1/F_{\alpha/2, \nu_2, \nu_1}$$

Using SAS

```
options nocenter ps = 35 ls = 72;
```

```
data example;
  input batch percent;
  cards;
  1 74
  1 76
  1 75
  2 68
  2 71
  2 72
  3 75
  3 77
  3 77
  4 72
  4 74
  4 73
  5 79
  5 81
  5 79
;

proc glm;
  class batch;
  model percent = batch;
  random batch;
  output out = diag r = res p = pred;
  proc plot;
  plot res*pred;
```

```
proc varcomp method = type1;
  class batch;
  model percent = batch;
```

```
proc mixed cl;
  class batch;
  model percent = ;
  random batch;
run;
```

The GLM Procedure

Dependent Variable: percent

Source	DF	Sum of Squares	Mean Square	F Value
Model	4	147.7333333	36.9333333	20.52

Error	10	18.0000000	1.8000000
Corrected Total	14	165.7333333	

Source	Pr > F
Model	<.0001

Variance Components Estimation Procedure

Type 1 Estimates

Variance Component	Estimate
Var(batch)	11.71111
Var(Error)	1.80000

The Mixed Procedure

Covariance Parameter Estimates

Cov Parm	Estimate	Alpha	Lower	Upper
batch	11.7111	0.05	4.0450	114.21
Residual	1.8000	0.05	0.8788	5.5436

Fit Statistics

-2 Res Log Likelihood	62.8
AIC (smaller is better)	66.8
AICC (smaller is better)	67.8
BIC (smaller is better)	66.0

Negative σ^2 Estimate Example

```
options nocenter ps=39 ls=64;
```

```
data new;
input class subj score @@;
cards;
1 1 74.62 1 2 73.90 1 3 72.27 1 4 71.60 1 5 73.80
1 6 77.42 1 7 72.16 1 8 76.69 1 9 75.84 1 10 70.35
2 1 72.55 2 2 71.44 2 3 72.67 2 4 72.59 2 5 71.25
2 6 68.99 2 7 69.61 2 8 77.44 2 9 73.99 2 10 73.90
3 1 76.66 3 2 74.76 3 3 70.47 3 4 75.38 3 5 68.32
3 6 76.69 3 7 73.34 3 8 68.24 3 9 69.33 3 10 78.22
;

proc glm;
class class;
```

```

model score = class;
random class / test;

proc varcomp method = type1;
class class;
model score = class;

proc varcomp method = reml;
class class;
model score = class;

proc mixed cl;
class class;
model score = ;
random class;
run;

```

The GLM Procedure

Dependent Variable: score

Source	DF	Sum of Squares	Mean Square
Model	2	10.1115467	5.0557733
Error	27	227.3489500	8.4203315
Corrected Total	29	237.4604967	

Source	F Value	Pr > F
Model	0.60	0.5557
Error		
Corrected Total		

R-Square	Coeff Var	Root MSE	score Mean
0.042582	3.966909	2.901781	73.14967

Source	Type III Expected Mean Square
class	Var(Error) + 10 Var(class)

Variance Components Estimation Procedure

Type 1 Estimates

Variance Component	Estimate
Var(class)	-0.33646
Var(Error)	8.42033

REML Estimates

Variance Component	Estimate
Var(class)	0
Var(Error)	8.18829