

Comparison of a Population Means

Design of Experiments - Montgomery
Section 3-1 through 3-3

4

Analysis of Variance

- Interested in comparing
 - Several treatments
 - Several levels of one treatment
- Could do numerous two-sample t-tests but this approach does not test equality of all means at once
- ANOVA provides method of joint inference

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

μ - grand mean

τ_i - i th treatment effect

$\epsilon_{ij} \sim N(0, \sigma^2)$ - error component

$$H_0 : \tau_1 = \tau_2 = \dots = \tau_a = 0$$

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

4-1

Partitioning y_{ij}

• Notation

– $y_{i.} = \sum_{j=1}^{n_i} y_{ij} \rightarrow \bar{y}_{i.} = y_{i.}/n_i$ (treatment sample mean)

– $y_{..} = \sum \sum y_{ij} \rightarrow \bar{y}_{..} = y_{..}/N$ (grand sample mean)

• Decomposition of y_{ij} :

Rewrite y_{ij} as $\bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.})$

Can then express y_{ij} as $\hat{\mu} + \hat{\tau}_i + \hat{\epsilon}_{ij}$ where

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

Built in parameter estimate restrictions

$$\sum n_i \hat{\tau}_i = 0$$

$$\sum \hat{\epsilon}_{ij} = 0 \text{ for all } i$$

4-2

Partitioning the Sum of Squares

• Recall $y_{ij} - \bar{y}_{..} = (\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.})$

• Can show

$$\sum \sum (y_{ij} - \bar{y}_{..})^2 = \sum n_i (\bar{y}_{i.} - \bar{y}_{..})^2 + \sum \sum (y_{ij} - \bar{y}_{i.})^2$$

Total SS = Treatment SS + error SS

$$SS_T = SS_{T \text{ treatments}} + SS_E$$

• To test H_0 , look at $\sum n_i \hat{\tau}_i^2 = \sum n_i (\bar{y}_{i.} - \bar{y}_{..})^2$

Small if $|\hat{\tau}_i|'s \approx 0$

If large then reject H_0 , but how large is large?

Standardize to account for inherent variability

$$F_0 = \frac{\sum n_i (\bar{y}_{i.} - \bar{y}_{..})^2 / (a - 1)}{\sum \sum (y_{ij} - \bar{y}_{i.})^2 / (N - a)} = \frac{\text{"ave" square between trts}}{\text{"ave" square within trts}}$$

4-3

Why use F_0 ?

- From general sum of squares result, can show
 - $SS_E/(N - a)$ is unbiased estimate of σ^2
 - Under H_0 , $SS_{\text{Treatment}}/(a - 1)$ is also unbiased estimate
- Call these estimates **Mean Squares**
- Can show the expected value is

$$E(MS_E) = \sigma^2$$

$$E(MS_{\text{Treatment}}) = \sigma^2 + \sum n_i \tau_i^2 / (a - 1)$$
- Use ratio of mean squares as test statistic
- If statistic deviates from one, then reject H_0
- What is test statistic distribution?

4-4

Test Statistic Distribution

Assume $y_{ij} \sim N(\mu + \tau_i, \sigma^2)$ and independent

$$\bar{y}_i \sim N(\mu + \tau_i, \sigma^2/n_i)$$

$$\sum (y_{ij} - \bar{y}_i)^2 / \sigma^2 \sim \chi_{n_i-1}^2$$

$$\sum \sum (y_{ij} - \bar{y}_i)^2 / \sigma^2 \sim \chi_{N-a}^2$$

Under H_0 :

$$\bar{y}_i \sim N(\mu, \sigma^2/n_i)$$

$$\sum n_i (\bar{y}_i - \bar{y}_{..})^2 / \sigma^2 \sim \chi_{a-1}^2$$

If we can show independence, can use F-distribution

Cochran's Thm

Since SS_T with $N - 1$ degrees of freedom is partitioned into $SS_{\text{Treatment}}$ and SS_E and $(a - 1) + (N - a) = N - 1$, then the two sum of squares are independent

Use F with $a - 1$ and $N - a$ degrees of freedom

4-5

Analysis of Variance Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Between	$SS_{\text{Treatment}}$	$a - 1$	$MS_{\text{Treatment}}$	F_0
Within	SS_E	$N - a$	MS_E	
Total	SS_T	$N - 1$		

If balanced:

$$SS_T = \sum \sum y_{ij}^2 - y_{..}^2 / N$$

$$SS_{\text{Treatment}} = \frac{1}{n} \sum y_i^2 - y_{..}^2 / N$$

$$SS_E = SS_T - SS_{\text{Treatment}}$$

If unbalanced:

$$SS_T = \sum \sum y_{ij}^2 - y_{..}^2 / N$$

$$SS_{\text{Treatment}} = \sum \frac{y_i^2}{n_i} - y_{..}^2 / N$$

$$SS_E = SS_T - SS_{\text{Treatment}}$$

If $F_0 > F_{\alpha, a-1, N-a}$ then reject H_0

4-6

Similarity With t-test

- Consider the square of the t-test statistic

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

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

4-7

Using SAS (lambs.sas)

Example

Twelve lambs are randomly assigned to three different diets. The weight gain (in 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	

$$\sum \sum y_{ij} = 156 \text{ and } \sum \sum y_{ij}^2 = 2274$$

$$y_1 = 33, y_2 = 75, \text{ and } y_3 = 48$$

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

$$SS_T = 2274 - 156^2/12 = 246$$

$$SS_{\text{Treatment}} = (33^2/3 + 75^2/5 + 48^2/4) - 156^2/12 = 36$$

$$SS_E = 246 - 36 = 210$$

$$F_0 = (36/2)/(210/9) = 0.77$$

P-value > 0.20 (DNR)

```
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;

proc glm;
class diet;
model wtgain=diet;
output out=diag r=res p=pred;

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

proc sort; by pred;
symbol1 v=circle i=sm50;
proc gplot; plot res*pred / haxis=axis1;
run;
```

4-8

4-9

Log File (.log)

```
393 option nocenter ps=65 ls=80;
395 data lambs;
396 input diet wtgain;
397 cards;
```

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

NOTE: DATA statement used:
real time 0.10 seconds

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

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

NOTE: PROCEDURE GPLOTT used:
real time 2.57 seconds

```
418 proc glm;
419 class diet;
420 model wtgain=diet;
421 output out=diag r=res p=pred;
422 run;
```

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

NOTE: The data set WORK.DIAG has 12 observations and 4 variables.

NOTE: PROCEDURE GLM used:
real time 0.44 seconds

4-10

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			

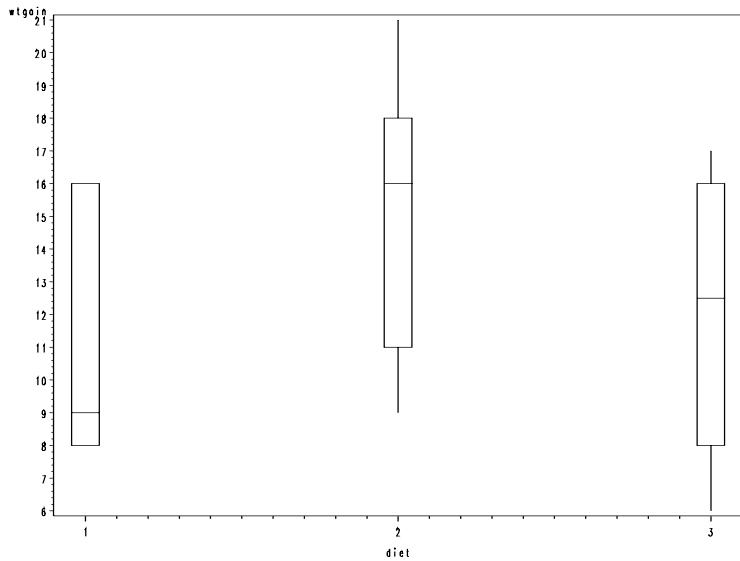
R-Square	Coeff Var	Root MSE	wtgain Mean
0.146341	37.15738	4.830459	13.00000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
diet	2	36.0000000	18.0000000	0.77	0.4907

Source	DF	Type III SS	Mean Square	F Value	Pr > F
diet	2	36.0000000	18.0000000	0.77	0.4907

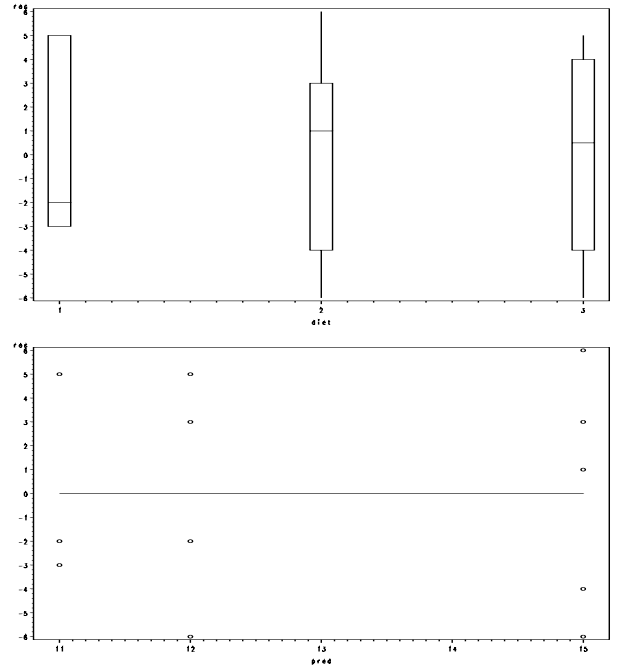
4-11

Plots



4-12

Plots



4-13

Handout Example

```
options ls=80 ps=60 nocenter;
options device=win target=winprtm rotate=landscape ftext=swiss
      hsize=8.0in vszie=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;
```

```
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;
  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;
```

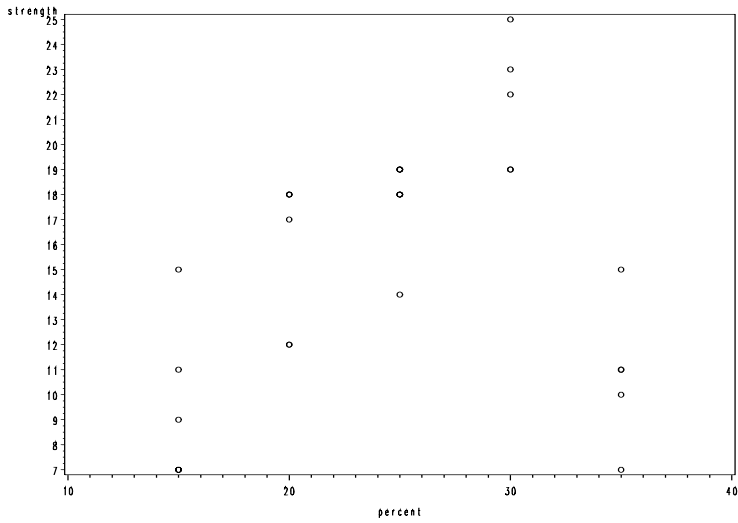
4-14

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

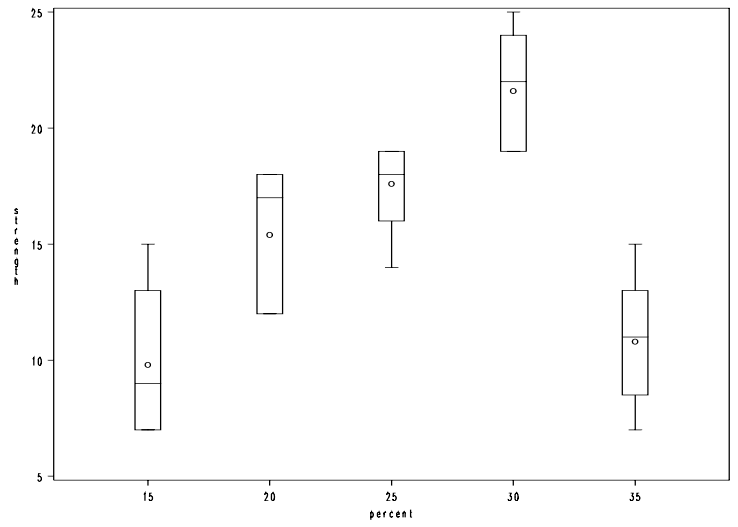
4-15

Plot of Strength vs Percent Blend



4-16

Plot of Strength vs Percent Blend



4-17

The GLM Procedure

Dependent Variable: strength

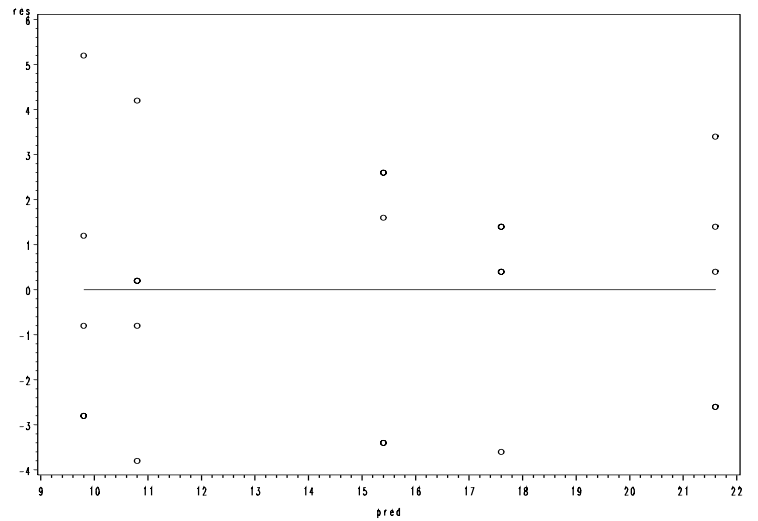
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	475.7600000	118.9400000	14.76	<.0001
Error	20	161.2000000	8.0600000		
Corrected Total	24	636.9600000			

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

Residual Plot



4-18

4-19

The UNIVARIATE Procedure
Variable: res

Moments			
N	25	Sum Weights	25
Mean	0	Sum Observations	0
Std Deviation	2.59165327	Variance	6.7166667
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	4.00000

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

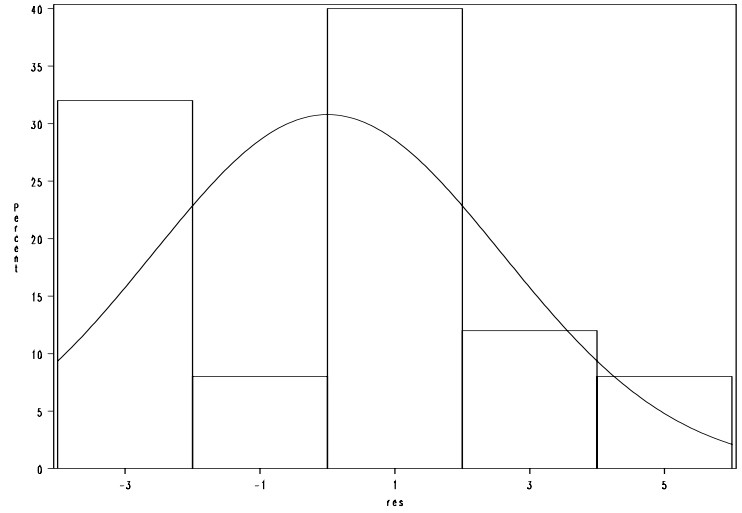
Fitted Distribution for res

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

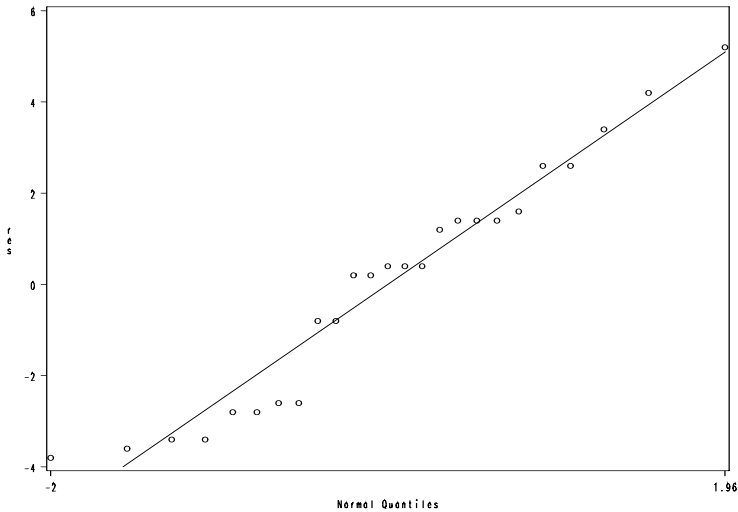
4-20

Residual Plot



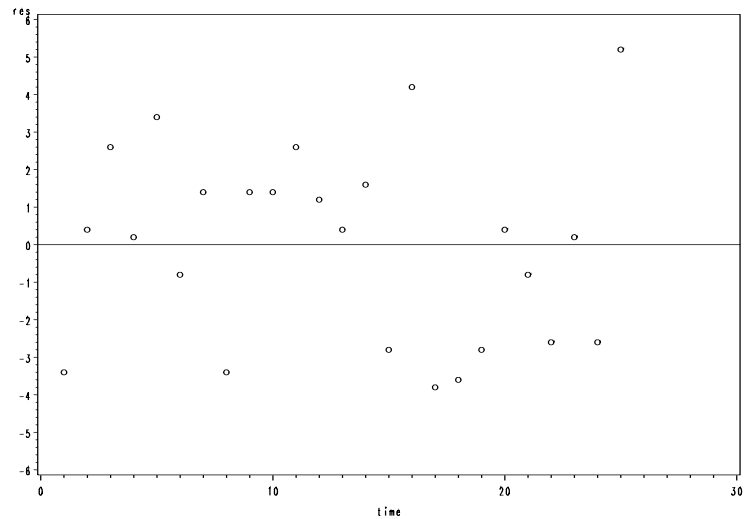
4-21

Residual Plot



4-22

Plot of residuals vs time



4-23