THE GENERAL ONE-WAY ANOVA

ANOVA Data AND ASSUMPTIONS

Assume we take independent RS’s of measurements from each of the "t" populations.

<table>
<thead>
<tr>
<th>Population (aka Sample)</th>
<th>Sample Values</th>
<th>Sample Variance</th>
<th>Sample Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Y_{11} Y_{12} ...</td>
<td>S_1^2</td>
<td>\bar{Y}_1</td>
</tr>
<tr>
<td>2</td>
<td>Y_{21} Y_{22} ...</td>
<td>S_2^2</td>
<td>\bar{Y}_2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>t</td>
<td>Y_{t1} Y_{t2} ...</td>
<td>S_t^2</td>
<td>\bar{Y}_t</td>
</tr>
</tbody>
</table>
MORE NOTATION

Notationally, \( Y_{ij} \) represents the \( j^{th} \) sample value from the \( i^{th} \) population.

\[
i = 1,2, \ldots, t \text{ (populations)} \\
j = 1, 2, \ldots, n_i \text{ (} n_i \text{ = number of observations from the } i^{th} \text{ population)}
\]

\[ n_T = n_1 + n_2 + \ldots + n_t \]

Assume \( Y_{ij} \sim \text{independent } N(\mu_i, \sigma^2), \quad i = 1,2, \ldots, t; \quad j = 1, 2, \ldots, n_i \)

QUESTIONS OF INTEREST

So our main or overall or initial hypothesis is

\[ H_0: \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_t \]

\[ H_A: \mu_i \neq \mu_j \text{ [at least two population means differ]} \]
**SOME QUICK TERMINOLOGY**

**RESPONSE:** Continuous measurement of interest

**FACTOR:** Categorical variable that "defines" the populations

**LEVELS:** The different values that the FACTOR can take on

Also known as **TREATMENTS**.

**DESIGNED EXPERIMENTS:**

**versus OBSERVATIONAL STUDIES:**

**COMPLETELY RANDOMIZED DESIGNS (CRD):**

**ONE-WAY ANOVA MODEL (aka One-Factor ANOVA)**

Model: $Y_{ij} = \mu_i + \varepsilon_{ij}$, random errors, $\varepsilon_{ij}$, are independent $N(0, \sigma^2)$

\[
\begin{align*}
\sigma & \\
\mu_i & \\
\end{align*}
\]

$i^{th}$- sample- average: $\bar{Y}_i = \frac{\sum_{j=1}^{n} Y_{ij}}{n_i}$

Overall- average: $\bar{Y}_* = \frac{\sum_{i=1}^{r} \sum_{j=1}^{n_i} Y_{ij}}{n_T}$
PARTITIONING THE TOTAL SUM OF SQUARES

Data on Crop Yield and Soil Type from the R Book.

<table>
<thead>
<tr>
<th>Soil Type</th>
<th>0</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>sand</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>clay</td>
<td>17</td>
<td>15</td>
<td>3</td>
<td>11</td>
<td>14</td>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td>loam</td>
<td>13</td>
<td>16</td>
<td>9</td>
<td>12</td>
<td>15</td>
<td>16</td>
<td>13</td>
</tr>
</tbody>
</table>

```
par(mfrow=c(2,2))
plot(Rep, Yield, pch=(1+14*(Rep>10)+9*(Rep>20)))
plot(Rep, Yield, pch=(1+14*(Rep>10)+9*(Rep>20)))
for (i in 1:30) lines(c(i,i), c(Yield[i], mean(Yield)))
abline(h=mean(Yield))
text(27,5, "SST")
plot(Rep, Yield, pch=(1+14*(Rep>10)+9*(Rep>20)))
for (i in 1:10) lines(c(i,i), c(Yield[i], mean(Yield[1:10])))
segments(1,mean(Yield[1:10]) , 10, mean(Yield[1:10]))
for (i in 11:20) lines(c(i,i), c(Yield[i], mean(Yield[11:20])))
segments(11,mean(Yield[11:20]) , 20, mean(Yield[11:20]))
for (i in 21:30) lines(c(i,i), c(Yield[i], mean(Yield[21:30])))
segments(21,mean(Yield[21:30]) , 30, mean(Yield[21:30]))
text(27,5, "SSE")
```
PARTITIONING THE TOTAL SUM OF SQUARES

Where's the average of these observations?

Define the TOTAL VARIATION or SUM OF SQUARES TOTAL as: \( \sum (y_i - \bar{y})^2 \)

Now suppose the Y's above actually came from four different groups/poplns, so that:

Can we find the average for each group?

Then find the total variation of the Y's in that group around this average? \( \sum (y_{ij} - \bar{y}_j)^2 \)

Then combine them to obtain the TOTAL VARIATION WITHIN GROUPS: \( \sum \sum (y_{ij} - \bar{y}_j)^2 \)

Can we find the average of these four points and the variation of these four points around this average, but we'll weight each deviation by the number of obs in that group.

This is the TOTAL VARIATION BETWEEN GROUPS: \( \sum \eta_i (\bar{y}_i - \bar{y})^2 \)
PARTITIONING THE TOTAL SUM OF SQUARES

Defn: \( \text{SSTot} = \text{SST} = \text{TSS} = \sum_{i=1}^{t} \sum_{j=1}^{n} (y_{ij} - \bar{y})^2 . \)

Now define: Sum of Squares **between (among)** groups = \( \text{SSB} = \sum_{i=1}^{t} n_i (\bar{y}_i - \bar{y})^2 \) and

Sum of Squares **within** groups = \( \text{SSW} = \sum_{i=1}^{t} \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2 . \)

These sums of squares add so that \( \text{SST} = \text{SSB} + \text{SSW} . \)
TEST STATISTIC?

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

$H_a: \mu_i \neq \mu_j$ [at least two population means differ]

$$F_{obs} = \frac{S_B^2}{S_W^2}$$

where the **between (among)** group $S_B^2 = \frac{SSB}{t-1} = \frac{\sum_{i=1}^{t} n_i (\bar{y}_i - \bar{y})^2}{t-1}$ variability is

and the **within** group variability is $S_W^2 = \frac{SSW}{n_T-t} = \frac{\sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}{n_1 + n_2 + \ldots + n_t - t}$.

Reject $H_0$ if: $F_{obs} > F_{\alpha, t-1, n_T-t}$ OR $p-value = Pr (F_{t-1, n_T-t} > F_{obs}) < \alpha$.

ANOVA TABLE

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>$F_{obs}$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td>SSB</td>
<td>t-1</td>
<td>SSB/(t-1)</td>
<td>MSB/MSW</td>
<td>Pr ($F_{t-1, n_T-t} &gt; F_{obs}$)</td>
</tr>
<tr>
<td>Within</td>
<td>SSW</td>
<td>n_T-t</td>
<td>SSW/(n_T-t)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>TSS</td>
<td>n_T-1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example Bacteria growth in meat under different packaging conditions (revisited)

Using R

```r
> #
> # Bacteria Count and meat packaging Example
> #
>
> MeatData = as.data.frame(scan(what=list(Condition='', logCount=0), sep='', dec='.'))
1: plastic, 7.66
2: plastic, 6.98
3: plastic, 7.80
4: vacuum, 5.26
5: vacuum, 5.44
6: vacuum, 5.80
7: mixed, 7.41
8: mixed, 7.33
9: mixed, 7.04
10: CO2, 3.51
11: CO2, 2.91
12: CO2, 3.66
13:
Read 12 records
> #
> # Alternative Data Entry from a .csv file
> #
> # MeatData = read.table(file("C:\ .. \Bacteria Data.csv"), header=TRUE)
> #
>
> attach(MeatData)
> MeatData
   Condition logCount
1    plastic     7.66
2    plastic     6.98
3    plastic     7.80
4     vacuum     5.26
5     vacuum     5.44
6     vacuum     5.80
7      mixed     7.41
8      mixed     7.33
9      mixed     7.04
10      CO2     3.51
11      CO2     2.91
12      CO2     3.66
> by(logCount, Condition, mean)
INDICES: CO2
[1] 3.36
------------------------------------------------------------
INDICES: mixed
[1] 7.26
------------------------------------------------------------
INDICES: plastic
[1] 7.48
------------------------------------------------------------
INDICES: vacuum
[1] 5.5
> by(logCount, Condition, sd)
INDICES: CO2
[1] 0.3968627
------------------------------------------------------------------
INDICES: mixed
[1] 0.1946792
------------------------------------------------------------------
INDICES: plastic
[1] 0.4386342
------------------------------------------------------------------
INDICES: vacuum
[1] 0.2749545
> numSummary(logCount, groups=Condition, statistics=c("mean", "sd"))
Loading required package: abind
table
               mean         sd n
CO2         3.36 0.3968627 3
mixed       7.26 0.1946792 3
plastic     7.48 0.4386342 3
vacuum      5.50 0.2749545 3

> boxplot(logCount ~ Condition)
```r
> MeatData$iplas = as.numeric(Condition=="plastic")
> MeatData$imixed = as.numeric(Condition=="mixed")
> MeatData$ivac = as.numeric(Condition=="vacuum")
> MeatData

<table>
<thead>
<tr>
<th>Condition</th>
<th>logCount</th>
<th>iplas</th>
<th>imixed</th>
<th>ivac</th>
</tr>
</thead>
<tbody>
<tr>
<td>plastic</td>
<td>7.66</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>plastic</td>
<td>6.98</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>plastic</td>
<td>7.80</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>vacuum</td>
<td>5.26</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>vacuum</td>
<td>5.44</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>vacuum</td>
<td>5.80</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>mixed</td>
<td>7.41</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>mixed</td>
<td>7.33</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>mixed</td>
<td>7.04</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>CO2</td>
<td>3.51</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CO2</td>
<td>2.91</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CO2</td>
<td>3.66</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

> MeatReg = lm(logCount ~ iplas + imixed + ivac, data=MeatData)
> par(mfrow=c(2,2))
> plot(MeatReg)

Hat values (leverages) are all = 0.3333333 and there are no factor predictors; no plot no. 5
> par(mfrow=c(1,1))
```

![Residuals vs Fitted](image1)

![Normal Q-Q](image2)

![Scale-Location](image3)
```r
> shapiro.test(MeatReg$residuals)

Shapiro-Wilk normality test

data:  MeatReg$residuals
W = 0.8942, p-value = 0.1336

> boxplot(MeatReg$residuals)

> summary(MeatReg)

Call:
  lm(formula = logCount ~ iplas + imixed + ivac, data = MeatData)

Residuals:
    Min     1Q Median     3Q    Max
-0.500 -0.225  0.110  0.210  0.320

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          3.36000     0.1965   17.10 1.39e-07 ***
iplas               4.12000     0.2779   14.82 4.22e-07 ***
imixed             3.90000     0.2779   14.03 6.45e-07 ***
ivac                2.14000     0.2779    7.70 5.74e-05 ***

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3404 on 8 degrees of freedom
Multiple R-squared: 0.9726,    Adjusted R-squared: 0.9623
F-statistic: 94.58 on 3 and 8 DF,  p-value: 1.376e-06
```
```r
> MeatANOVA = lm(logCount ~ Condition, data=MeatData)
> par(mfrow=c(2,2))
> plot(MeatANOVA)
> par(mfrow=c(1,1))

Residuals vs Fitted

Normal Q-Q

Scale-Location

Constant Leverage:
Residuals vs Factor Levels

> shapiro.test(MeatANOVA$residuals)

Shapiro-Wilk normality test

data:  MeatANOVA$residuals
W = 0.8942, p-value = 0.1336

> boxplot(MeatANOVA$residuals)
```
> summary(MeatANOVA)

Call:
  lm(formula = logCount ~ Condition, data = MeatData)

Residuals:
     Min      1Q  Median      3Q     Max
-0.500 -0.225  0.110  0.210  0.320

Coefficients:                  Estimate Std. Error t value Pr(>|t|)
(Intercept)            3.3600     0.1965   17.10 1.39e-07 ***
Condition[T.mixed]     3.9000     0.2779   14.03 6.45e-07 ***
Condition[T.plastic]   4.1200     0.2779   14.82 4.22e-07 ***
Condition[T.vacuum]    2.1400     0.2779    7.70 5.74e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3404 on 8 degrees of freedom
Multiple R-squared: 0.9726, Adjusted R-squared: 0.9623
F-statistic: 94.58 on 3 and 8 DF, p-value: 1.376e-06

> anova(MeatANOVA)

Analysis of Variance Table

Response: logCount
        Df Sum Sq Mean Sq F value Pr(>F)
Condition  3 32.873  10.958 94.584  1.376e-06 ***
Residuals  8  0.927   0.116
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```r
> library(multcomp)
> confint(MeatANOVA)

<table>
<thead>
<tr>
<th></th>
<th>2.5 %</th>
<th>97.5 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>2.906844</td>
<td>3.813156</td>
</tr>
<tr>
<td>Condition[T.mixed]</td>
<td>3.259141</td>
<td>4.540859</td>
</tr>
<tr>
<td>Condition[T.plastic]</td>
<td>3.479141</td>
<td>4.760859</td>
</tr>
<tr>
<td>Condition[T.vacuum]</td>
<td>1.499141</td>
<td>2.780859</td>
</tr>
</tbody>
</table>

> plot(print(confint(glht(MeatANOVA, linfct=mcp(Condition="Tukey")))))
```

Simultaneous Confidence Intervals for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts

Fit: lm(formula = logCount ~ Condition, data = MeatData)

Estimated Quantile = 3.2024

Linear Hypotheses:

<table>
<thead>
<tr>
<th>Linear Function</th>
<th>Estimate lwr</th>
<th>upr</th>
</tr>
</thead>
<tbody>
<tr>
<td>mixed - CO2 == 0</td>
<td>3.9000 3.0100</td>
<td>4.7900</td>
</tr>
<tr>
<td>plastic - CO2 == 0</td>
<td>4.1200 3.2300</td>
<td>5.0100</td>
</tr>
<tr>
<td>vacuum - CO2 == 0</td>
<td>2.1400 1.2500</td>
<td>3.0300</td>
</tr>
<tr>
<td>plastic - mixed == 0</td>
<td>0.2200 -0.6700</td>
<td>1.1100</td>
</tr>
<tr>
<td>vacuum - mixed == 0</td>
<td>-1.7600 -2.6500</td>
<td>-0.8700</td>
</tr>
<tr>
<td>vacuum - plastic == 0</td>
<td>-1.9800 -2.8700</td>
<td>-1.0900</td>
</tr>
</tbody>
</table>

95% family-wise confidence level
WHAT IF WE USED THE BACTERIAL COUNTS RATHER THAN THE logCOUNTS?

```r
> MeatData$Count = 10^logCount
> detach(MeatData)
> attach(MeatData)
> numSummary(Count, groups=Condition, statistics=c("mean", "sd"))

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>sd</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>CO2</td>
<td>2873.216</td>
<td>1905.102</td>
<td>3</td>
</tr>
<tr>
<td>mixed</td>
<td>19349453.561</td>
<td>7576411.476</td>
<td>3</td>
</tr>
<tr>
<td>plastic</td>
<td>39451493.090</td>
<td>27315819.586</td>
<td>3</td>
</tr>
<tr>
<td>vacuum</td>
<td>362783.434</td>
<td>236899.326</td>
<td>3</td>
</tr>
</tbody>
</table>

> boxplot(Count ~ Condition)
```

![Boxplot of bacterial counts by condition](image)
```r
> MeatANOVACount = lm(Count ~ Condition, data=MeatData)
> par(mfrow=c(2,2))
> plot(MeatANOVACount)
> par(mfrow=c(1,1))

> shapiro.test(MeatANOVACount$residuals)
Shapiro-Wilk normality test
data:  MeatANOVACount$residuals
W = 0.8102, p-value = 0.01227

> boxplot(MeatANOVACount$residuals)
```

---

C:\Users\baileraj\BAILERAJ\Classes\Web-CLASSES\ies-612\lectures\Week 6.2--IES 612-STA 4-573-STA 4-576-14feb09.doc2/15/2009 16
> summary(MeatANOVACount)

Call:
  lm(formula = Count ~ Condition, data = MeatData)

Residuals:
   Min     1Q Median     3Q    Max
-29901567 -110724   1030  3086957 23644241

Coefficients:               Estimate Std. Error  t value Pr(>|t|)
(Intercept)              2873    8183378 0.000351  0.99973
Condition[T.mixed]   19346580   11573044    1.672  0.13313
Condition[T.plastic] 39448620   11573044    3.409  0.00924 **
Condition[T.vacuum]    359910   11573044    0.031  0.97595
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14170000 on 8 degrees of freedom
Multiple R-squared: 0.6634,     Adjusted R-squared: 0.5371
F-statistic: 5.255 on 3 and 8 DF,  p-value: 0.02699

> anova(MeatANOVACount)

Analysis of Variance Table

Response: Count

  Df Sum Sq Mean Sq   F value Pr(>F)
Condition   3 3.1673e+15 1.0558e+15  5.2552   0.02699 *
Residuals   8 1.6072e+15 2.0090e+14
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Simultaneous Confidence Intervals for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts

Fit: lm(formula = Count ~ Condition, data = MeatData)

Estimated Quantile = 3.2009

Linear Hypotheses:

<table>
<thead>
<tr>
<th>Estimate</th>
<th>lwr</th>
<th>upr</th>
</tr>
</thead>
<tbody>
<tr>
<td>mixed - CO2 == 0</td>
<td>19346580.3451</td>
<td>-17697030.1975</td>
</tr>
<tr>
<td>plastic - CO2 == 0</td>
<td>39448619.8736</td>
<td>2405009.3311</td>
</tr>
<tr>
<td>vacuum - CO2 == 0</td>
<td>359910.2172</td>
<td>-36683700.3253</td>
</tr>
<tr>
<td>plastic - mixed == 0</td>
<td>20102039.5285</td>
<td>-16941571.0140</td>
</tr>
<tr>
<td>vacuum - mixed == 0</td>
<td>-18986670.1278</td>
<td>-56030280.6703</td>
</tr>
<tr>
<td>vacuum - plastic == 0</td>
<td>-39088709.6563</td>
<td>-76132320.1989</td>
</tr>
</tbody>
</table>

95% family-wise confidence level
title "One-way ANOVA/ CRD example + contrasts + multiple comparisons"
title2 "Bacteria in meat data"
data meat;
  input condition $ logcount @@;
  ivac = (condition="vacuum");
  imix = (condition="mixed");
  iCO2 = (condition="CO2");
cards;
  plastic 7.66 plastic 6.98 plastic 7.80
  vacuum 5.26 vacuum 5.44 vacuum 5.80
  mixed 7.41 mixed 7.33 mixed 7.04
  CO2 3.51 CO2 2.91 CO2 3.66;
run;
proc print data=meat;
run;
proc sort out=smeat; by condition;
proc univariate plot; by condition;
  title3 summary statistics and boxplot;
  var logcount;
run;
proc reg data=meat;
  title3 Regression with indicators;
  model logcount = ivac imix iCO2;
run;
proc glm data=meat order=data;
  title3 One-way anova + contrast + model adequacy;
  class condition;
  model logcount=condition;
  output out=new p=yhat r=resid;
  contrast 'plastic vs. rest' condition 3 -1 -1 -1;
  estimate 'plastic vs. rest' condition 3 -1 -1 -1;
  contrast 'CO2 vs. plastic' condition -1 0 0 1;
  estimate 'CO2 vs. plastic' condition -1 0 0 1;
  contrast 'CO2 vs. vacuum' condition 0 -1 0 1;
  estimate 'CO2 vs. vacuum' condition 0 -1 0 1;
  contrast 'CO2 vs. mixed' condition 0 0 -1 1;
  estimate 'CO2 vs. mixed' condition 0 0 -1 1;
  lsmeans condition / stderr pdiff;
  means condition / lsd clm;
  means condition / bon scheffe tukey;
  means condition / bon tukey cldiff;
run;
proc plot data=new;
  plot logcount*condition yhat*condition='p' /overlay;
  plot resid*condition resid*yhat / vref=0;
run;
proc univariate plot;
  var resid;
run;
* construct the normal scores - Z[(i-.375)/(n+.25)];
* note not multiplied by sqrt(mse);

```
proc rank data=new normal=blom out=rnew;
  var resid;
  ranks nscore;
run;
```

* generate plot analogous to univariate's normal prob. plot;

```
proc plot;
  plot resid*nscore;
run;
```

data moremeat; set meat;
  count = exp(logcount);
  title3 raw count data analyzed;
run;

proc glm data=moremeat;
  class condition;
  model count=condition;
  output out=mnew p=yhat r=resid;
  lsmeans condition / stderr pdiff;
  * means condition / clm bon scheffe lsd tukey snk;
run;

proc plot data=mnew;
  plot count*condition yhat*condition='p' /overlay;
  plot resid*condition resid*yhat / vref=0;
run;

proc univariate data=mnew plot;
  var resid;
run;

proc rank data=mnew normal=blom out=rnew;
  var resid;
  ranks nscore;
run;

proc plot;
  plot resid*nscore;
run;
```
proc print data=meat;
run;
Obs   condition logcount   ivac   imix   iCO2
 1    plastic   7.66    0       0       0
 2    plastic   6.98    0       0       0
 3    plastic   7.80    0       0       0
 4    vacuum   5.26    1       0       0
 5    vacuum   5.44    1       0       0
 6    vacuum   5.80    1       0       0
 7    mixed    7.41    0       1       0
 8    mixed    7.33    0       1       0
 9    mixed    7.04    0       1       0
10   CO2       3.51    0       0       1
11   CO2       2.91    0       0       1
12   CO2       3.66    0       0       1

proc sort out=smeat; by condition;
run;
proc univariate plot; by condition;
  title3 summary statistics and boxplot;
  var logcount;
run;

The UNIVARIATE Procedure
Variable: logcount

Schematic Plots

8 +
 |    *-----*
 |    +-----+    | + |
 |    *+-+-*    |   |
7 +    +-----+    +-----+
 |
6 +    +-----+    +-----+
 |      |   |
5 +    +-----+    *+-+-*
 |      |   |
4 +    +-----+    *-----*
 |      | + | 
3 +    +-----+    |   |
2 +    +-----+    +-----+    +-----+    +-----+    +-----+
condition   CO2   mixed   plastic   vacuum
```latex
\begin{verbatim}
proc reg data=meat;
title3 Regression with indicators;
model logcount = ivac imix iCO2;
run;
\end{verbatim}
```

The REG Procedure
Model: MODEL1
Dependent Variable: logcount

Analysis of Variance

\begin{tabular}{|c|c|c|c|c|c|}
\hline
Source & DF & Sum of Squares & Mean Square & F Value & Pr > F \\
\hline
Model & 3 & 32.87280 & 10.95760 & 94.58 & <.0001 \\
Error & 8 & 0.92680 & 0.11585 & & \\
Corrected Total & 11 & 33.79960 & & & \\
\hline
\end{tabular}

Root MSE 0.34037 R-Square 0.9726 Dependent Mean 5.90000 Adj R-Sq 0.9623 Coeff Var 5.76894

Parameter Estimates

\begin{tabular}{|c|c|c|c|c|c|}
\hline
Variable & Parameter & DF & Estimate & Standard Error & t Value & Pr > |t| \\
\hline
Intercept & 1 & 7.48000 & 0.19651 & 38.06 & <.0001 \\
ivac & 1 & -1.98000 & 0.27791 & -7.12 & <.0001 \\
imix & 1 & -0.22000 & 0.27791 & -0.79 & 0.4514 \\
iCO2 & 1 & -4.12000 & 0.27791 & -14.83 & <.0001 \\
\hline
\end{tabular}

```

```latex
\begin{verbatim}
proc glm data=meat order=data;
title3 One-way anova + contrast + model adequacy;
class condition;
model logcount=condition;
output out=new p=yhat r=resid;
run;
\end{verbatim}
```

The GLM Procedure

Class Level Information

\begin{tabular}{|c|c|}
\hline
Class & Levels \\
\hline
condition & 4 plastic vacuum mixed CO2 \\
\hline
\end{tabular}

Number of observations 1

The GLM Procedure
Dependent Variable: logcount

Analysis of Variance

\begin{tabular}{|c|c|c|c|c|c|}
\hline
Source & DF & Sum of Squares & Mean Square & F Value & Pr > F \\
\hline
Model & 3 & 32.87280000 & 10.95760000 & 94.58 & <.0001 \\
Error & 8 & 0.92680000 & 0.11585000 & & \\
Corrected Total & 11 & 33.79960000 & & & \\
\hline
\end{tabular}

R-Square 0.972580 Coeff Var 5.768940 Root MSE 0.340367 logcount Mean 5.900000

\begin{tabular}{|c|c|c|c|c|c|}
\hline
Source & DF & Type I SS & Mean Square & F Value & Pr > F \\
\hline
condition & 3 & 32.87280000 & 10.95760000 & 94.58 & <.0001 \\
\hline
\end{tabular}

\begin{tabular}{|c|c|c|c|c|c|}
\hline
Source & DF & Type III SS & Mean Square & F Value & Pr > F \\
\hline
condition & 3 & 32.87280000 & 10.95760000 & 94.58 & <.0001 \\
\hline
\end{tabular}
```
contrast 'plastic vs. rest' condition 3 -1 -1 -1;
estimate 'plastic vs. rest' condition 3 -1 -1 -1;
contrast 'CO2 vs. plastic' condition -1 0 0 1;
estimate 'CO2 vs. plastic' condition -1 0 0 1;
contrast 'CO2 vs. vacuum' condition 0 -1 0 1;
estimate 'CO2 vs. vacuum' condition 0 -1 0 1;
contrast 'CO2 vs. mixed' condition 0 0 -1 1;
estimate 'CO2 vs. mixed' condition 0 0 -1 1;

Contrast                    DF     Contrast SS     Mean Square    F Value    Pr > F
plastic vs. rest             1      9.98560000      9.98560000      86.19    <.0001
CO2 vs. plastic              1     25.46160000     25.46160000     219.78    <.0001
CO2 vs. vacuum               1      6.86940000      6.86940000      59.30    <.0001
CO2 vs. mixed                1     22.81500000     22.81500000     196.94    <.0001
Dependent Variable: logcount

Parameter                  Estimate           Error    t Value    Pr > |t|
plastic vs. rest          6.32000000      0.68073490       9.28      <.0001
CO2 vs. plastic          -4.12000000      0.27790886     -14.83      <.0001
CO2 vs. vacuum           -2.14000000      0.27790886      -7.70      <.0001
CO2 vs. mixed            -3.90000000      0.27790886     -14.03      <.0001

lsmeans condition / stderr pdiff;

The GLM Procedure
Least Squares Means

| condition | logcount   | Standard Error | Pr > |t|  | Number |
|-----------|------------|----------------|------|---|--------|
| plastic   | 7.48000000 | 0.19651124     | <.0001 | 1 |
| vacuum    | 5.50000000 | 0.19651124     | <.0001 | 2 |
| mixed     | 7.26000000 | 0.19651124     | <.0001 | 3 |
| CO2       | 3.36000000 | 0.19651124     | <.0001 | 4 |

Least Squares Means for effect condition
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: logcount

<table>
<thead>
<tr>
<th>i/j</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>&lt;.0001</td>
<td>0.4514</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>&lt;.0001</td>
<td>0.0002</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.4514</td>
<td>0.0002</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td></td>
</tr>
</tbody>
</table>

NOTE: To ensure overall protection level, only probabilities associated with pre-planned comparisons should be used
t Confidence Intervals for logcount

<table>
<thead>
<tr>
<th>condition</th>
<th>N</th>
<th>Mean</th>
<th>Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>plastic</td>
<td>3</td>
<td>7.4800</td>
<td>7.0268</td>
</tr>
<tr>
<td>mixed</td>
<td>3</td>
<td>7.2600</td>
<td>6.8068</td>
</tr>
<tr>
<td>vacuum</td>
<td>3</td>
<td>5.5000</td>
<td>5.0468</td>
</tr>
<tr>
<td>C02</td>
<td>3</td>
<td>3.3600</td>
<td>2.9068</td>
</tr>
</tbody>
</table>

Tukey's Studentized Range (HSD) Test for logcount

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

<table>
<thead>
<tr>
<th>condition</th>
<th>N</th>
<th>Mean</th>
<th>Minimum Significant Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3</td>
<td>7.4800</td>
<td>0.89</td>
</tr>
<tr>
<td>A</td>
<td>3</td>
<td>7.2600</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>3</td>
<td>5.5000</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>3.3600</td>
<td></td>
</tr>
</tbody>
</table>

Bonferroni (Dunn) t Tests for logcount

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

<table>
<thead>
<tr>
<th>condition</th>
<th>N</th>
<th>Mean</th>
<th>Minimum Significant Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3</td>
<td>7.4800</td>
<td>0.9668</td>
</tr>
<tr>
<td>A</td>
<td>3</td>
<td>7.2600</td>
<td></td>
</tr>
</tbody>
</table>
**Scheffe's Test for logcount**

NOTE: This test controls the Type I experimentwise error rate.

<table>
<thead>
<tr>
<th>Mean</th>
<th>N</th>
<th>condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>7.4800</td>
<td>plastic</td>
</tr>
<tr>
<td>A</td>
<td>7.2600</td>
<td>mixed</td>
</tr>
<tr>
<td>B</td>
<td>5.5000</td>
<td>vacuum</td>
</tr>
<tr>
<td>C</td>
<td>3.3600</td>
<td>CO2</td>
</tr>
</tbody>
</table>

Means with the same letter are not significantly different.

**Tukey's Studentized Range (HSD) Test for logcount**

NOTE: This test controls the Type I experimentwise error rate.

<table>
<thead>
<tr>
<th>Difference</th>
<th>Between Means</th>
<th>Simultaneous 95% Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>plastic - mixed</td>
<td>0.2200</td>
<td>-0.6700 to 1.1100</td>
</tr>
<tr>
<td>plastic - vacuum</td>
<td>1.9800</td>
<td>1.0900 to 2.8700</td>
</tr>
<tr>
<td>plastic - CO2</td>
<td>4.1200</td>
<td>3.2300 to 5.0100</td>
</tr>
<tr>
<td>mixed - plastic</td>
<td>-0.2200</td>
<td>-1.1100 to 0.6700</td>
</tr>
<tr>
<td>mixed - vacuum</td>
<td>1.7600</td>
<td>0.8700 to 2.6500</td>
</tr>
<tr>
<td>mixed - CO2</td>
<td>3.9000</td>
<td>3.0100 to 4.7900</td>
</tr>
<tr>
<td>vacuum - plastic</td>
<td>-1.9800</td>
<td>-2.8700 to -1.0900</td>
</tr>
<tr>
<td>vacuum - mixed</td>
<td>-1.7600</td>
<td>-2.6500 to -0.8700</td>
</tr>
<tr>
<td>vacuum - CO2</td>
<td>2.1400</td>
<td>1.2500 to 3.0300</td>
</tr>
<tr>
<td>CO2 - plastic</td>
<td>-4.1200</td>
<td>-5.0100 to -3.2300</td>
</tr>
<tr>
<td>CO2 - mixed</td>
<td>-3.9000</td>
<td>-4.7900 to -3.0100</td>
</tr>
<tr>
<td>CO2 - vacuum</td>
<td>-2.1400</td>
<td>-3.0300 to -1.2500</td>
</tr>
</tbody>
</table>

**Bonferroni (Dunn) t Tests for logcount**

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than Tukey's for all pairwise comparisons.

<table>
<thead>
<tr>
<th>Mean</th>
<th>N</th>
<th>condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>7.4800</td>
<td>plastic</td>
</tr>
<tr>
<td>A</td>
<td>7.2600</td>
<td>mixed</td>
</tr>
<tr>
<td>B</td>
<td>5.5000</td>
<td>vacuum</td>
</tr>
<tr>
<td>C</td>
<td>3.3600</td>
<td>CO2</td>
</tr>
</tbody>
</table>
Comparisons significant at the 0.05 level are indicated by ***.

<table>
<thead>
<tr>
<th>Difference</th>
<th>Between Means</th>
<th>Simultaneous 95% Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>plastic - mixed</td>
<td>0.2200</td>
<td>-0.7468 to 1.1868</td>
</tr>
<tr>
<td>plastic - vacuum</td>
<td>1.9800</td>
<td>1.0132 to 2.9468 ***</td>
</tr>
<tr>
<td>plastic - CO2</td>
<td>4.1200</td>
<td>3.1532 to 5.0868 ***</td>
</tr>
<tr>
<td>mixed - plastic</td>
<td>-0.2200</td>
<td>-1.1868 to 0.7468</td>
</tr>
<tr>
<td>mixed - vacuum</td>
<td>1.7600</td>
<td>0.7932 to 2.7268 ***</td>
</tr>
<tr>
<td>mixed - CO2</td>
<td>3.9000</td>
<td>2.9332 to 4.8668 ***</td>
</tr>
<tr>
<td>vacuum - plastic</td>
<td>-1.9800</td>
<td>-2.9468 to -1.0132 ***</td>
</tr>
<tr>
<td>vacuum - mixed</td>
<td>-1.7600</td>
<td>-2.7268 to -0.7932 ***</td>
</tr>
<tr>
<td>vacuum - CO2</td>
<td>2.1400</td>
<td>1.1732 to 3.1068 ***</td>
</tr>
<tr>
<td>C02 - plastic</td>
<td>-4.1200</td>
<td>-5.0868 to -3.1532 ***</td>
</tr>
<tr>
<td>C02 - mixed</td>
<td>-3.9000</td>
<td>-4.8668 to -2.9332 ***</td>
</tr>
<tr>
<td>C02 - vacuum</td>
<td>-2.1400</td>
<td>-3.1068 to -1.1732 ***</td>
</tr>
</tbody>
</table>

```plaintext
options ls=70;
proc plot data=new;
plot logcount*condition yhat*condition='p' /overlay;
plot resid*condition resid*yhat / vref=0;
run;
```
Plot of resid*hat. Legend: A = 1 obs, B = 2 obs, etc.

```
proc univariate plot;
  var resid;
* construct the normal scores - Z[(i-.375)/(n+.25)];
* note not multiplied by sqrt(mse);
proc rank data=new normal=blom out=rnew;
  var resid;
  ranks nscore;
* generate plot analogous to univariate's normal prob. plot;
proc plot;
  plot resid*nscore;
```
data moremeat; set meat;
count = exp(logcount);
title3 raw count data analyzed;

proc glm data=moremeat;
  class condition;
  model count=condition;
  output out=mnew p=yhat r=resid;
  lsmeans condition / stderr pdiff;
  * means condition / clm bon scheffe lsd tukey snk;
run;

The GLM Procedure
Dependent Variable: count

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>7282652.348</td>
<td>2427550.783</td>
<td>16.56</td>
<td>0.0009</td>
</tr>
<tr>
<td>Error</td>
<td>8</td>
<td>1172820.616</td>
<td>146602.577</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>11</td>
<td>8455472.964</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-Square: 0.861294
Coeff Var: 42.54159
Root MSE: 382.8872
count Mean: 900.0303

C:\Users\baileraj\BAILERAJ\Classes\Web-CLASSES\ies-612\lectures\Week 6.2--IES 612-STA 4-573-STA 4-576-14feb09.doc2/15/200928
Source                      DF     Type III SS     Mean Square    F Value    Pr > F
condition                      3     7282652.348     2427550.783      16.56    0.0009

proc plot data=mnew;
  plot count*condition yhat*condition='p' /overlay;
  plot resid*condition resid*yhat / vref=0;
run;

Plot of resid*condition. Legend: A = 1 obs, B = 2 obs, etc.
Plot of resid*yhat. Legend: A = 1 obs, B = 2 obs, etc.
proc rank data=mnew normal=blom out=rnew;
  var resid;
  ranks nscore;
run;
proc plot;
  plot resid*nscore;
run;