

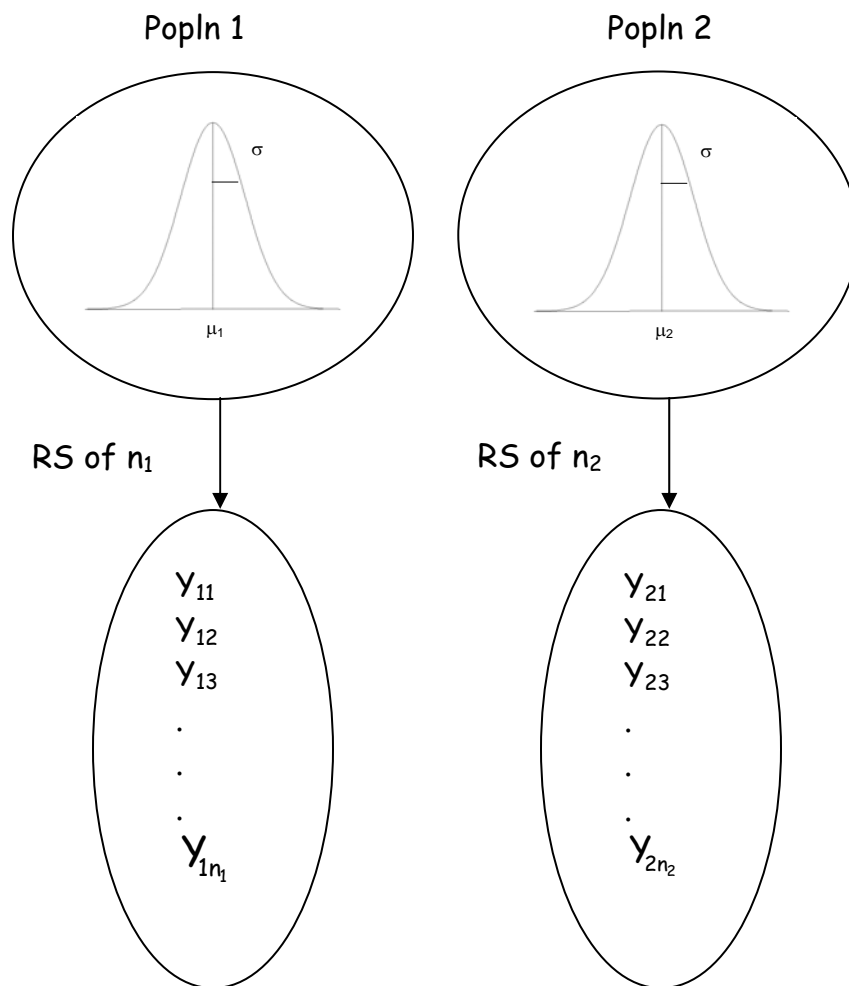
Week 6.1--IES 612-STA 4-573-STA 4-576.doc

IES 612/STA 4-573/STA 4-576 Winter 2009

ANOVA: models for comparing means of different treatments or populations

Recall the two-group pooled variance t-test

Assume we take **independent** RS's of measurements from each of the two populations.



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

$H_0: \mu_1 = \mu_2$ [two populations do NOT differ in mean response]

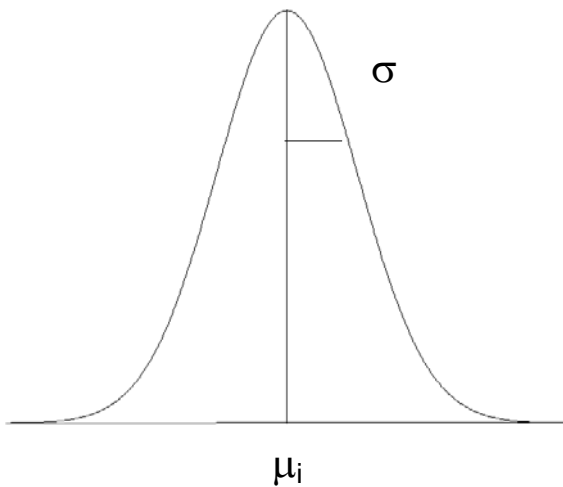
$H_a: \mu_1 \neq \mu_2$

Assumptions/Data?

Population (aka Sample)	Sample Value			Sample variance	Sample Average
	1	2	...		
Popln 1 = $N(\mu_1, \sigma^2)$	Y_{11}	Y_{12}	...	s_1^2	\bar{Y}_1
Popln 2 = $N(\mu_2, \sigma^2)$	Y_{21}	Y_{22}	...	s_2^2	\bar{Y}_2

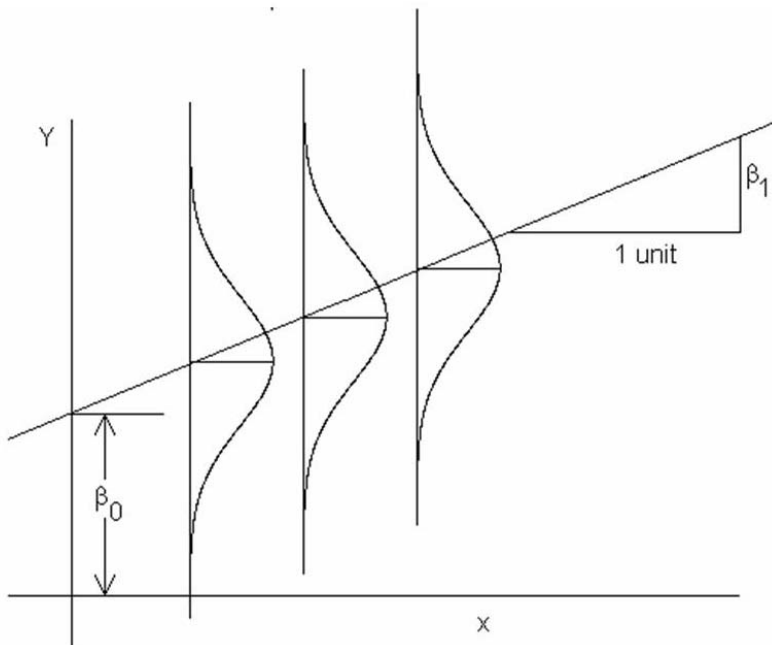
Assume $Y_{ij} \sim$ independent $N(\mu_i, \sigma^2)$ $i = 1, 2$ $j = 1, 2, \dots, n_i$

Another way of writing this is $Y_{ij} = \mu_i + \varepsilon_{ij}$ with $\varepsilon_{ij} \sim$ independent $N(0, \sigma^2)$.



In other words, the response of the " j^{th} " observation in the " i^{th} " population can be written in terms of the mean of the i^{th} population + how this observation differs from the mean. Does this look familiar? Aka SLR?

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$



Test Statistic?

$$t_{obs} = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad \text{where } s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{\sum_{i=1}^2 \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}{n_1+n_2-2}$$

The pooled variance looks like something from regression. What? $s^2 = \text{MSE} = \frac{\sum_{i=1}^n (y_i - \hat{y})^2}{n-2}$

Can we test the equality of two popln means $H_0: \mu_1 = \mu_2$ vs. $H_a: \mu_1 \neq \mu_2$ using Regression?

First recall tries to determine the LINEAR relationship between a CONTINUOUS RESPONSE variable (Y) and a CONTINUOUS INDEPENDENT variable (X). Hence regression data consists of NUMBERS!!!

Our Two-Sample Problem's Data looks like:

Data from population 1: $(y_{11}, y_{12}, \dots, y_{1n_1})$

Data from population 2: $(y_{21}, y_{22}, \dots, y_{2n_2})$

Let $X = 1$ (if popln 2) and $X=0$ (if popln 1) then our SLR is: $Y = \beta_0 + \beta_1 X + \varepsilon$

so for Popln 1: $Y = \beta_0 + \varepsilon$ and Popln 2: $Y = \beta_0 + \beta_1 + \varepsilon$

Recall that the " $\beta_0 + \beta_1 X$ " part of any regression model represented the _____!

Implying $\mu_1 = \beta_0$ and $\mu_2 = \beta_0 + \beta_1$ so $\beta_1 = \mu_2 - \mu_1$. Thus, $H_0: \mu_1 = \mu_2$ AND $H_0: \beta_1 = 0$ test the same hypothesis.

Example: Bacteria Count on meat found using two different packing conditions

Packing Condition	LogBacteria Count	X
Vacuum	5.26	
Vacuum	5.44	
Vacuum	5.8	
Mixed	7.41	
Mixed	7.33	
Mixed	7.04	

How about a quick of the assumptions of the Two-Sample T-Test?

What about the variance being the same in both populations?

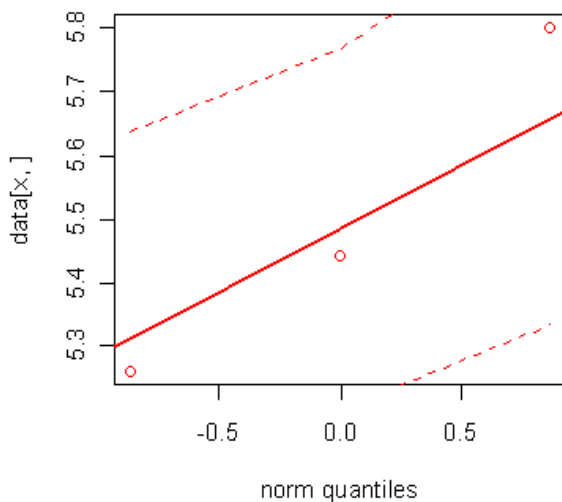
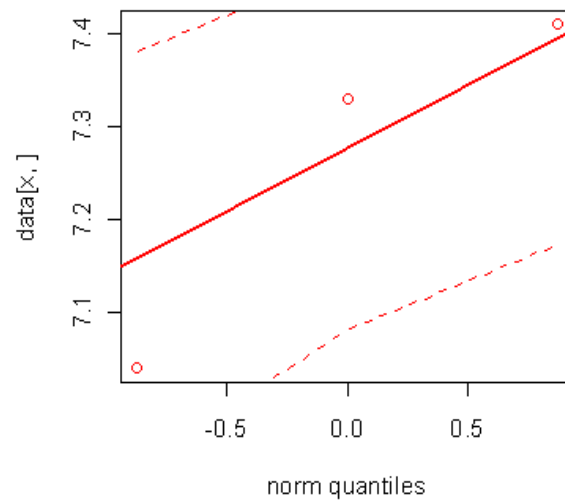
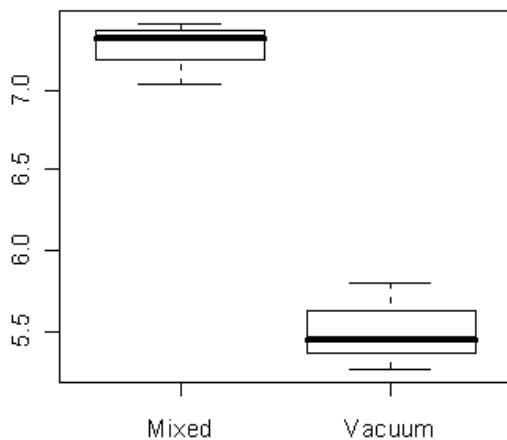
What about the populations both Normal?

We assume the data set has already been entered or imported so that we have this data available:

	Condition	LogBacteria
1	Vacuum	5.26
2	Vacuum	5.44
3	Vacuum	5.80
4	Mixed	7.41
5	Mixed	7.33
6	Mixed	7.04

While RComander will give us the Boxplots, to obtain separate NPP's we need to use commands.

```
> par(mfrow=c(2,2))
> boxplot(LogBacteria ~ Condition)
> by(LogBacteria, Condition, qq.plot)
```



Conclusions?

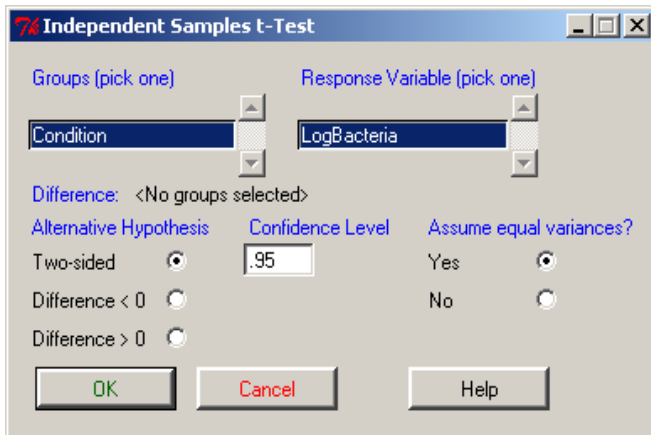
EQUIVALENCE OF 2-SAMPLE T-TEST, REGRESSION, AND ONE-WAY ANOVA

Using R

Two-Sample T-Test Analysis

With *MeatBacteria* the active data set, using RCommander:

Statistics > Means > Independent samples t-test ...



```
> t.test(LogBacteria~Condition, alternative='two.sided', conf.level=.95,  
var.equal=TRUE, data=MeatBacteria)
```

Two Sample t-test

```
data: LogBacteria by Condition  
t = 9.0485, df = 4, p-value = 0.0008266  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 1.219959 2.300041  
sample estimates:  
mean in group Mixed mean in group Vacuum  
          7.26          5.50
```

Regression Analysis

Create the X variable using Recode from RCommander, with *MeatBacteria* the active data set.

Data > Manage Variables in Active data set > Recode variables ...

The image shows two windows from RCommander. The left window is titled "Recode Variables" and has the following fields: "Variables to recode (pick one or more)" with "Condition" and "LogBacteria" selected; "New variable name or prefix for multiple recodes:" with "X" entered; "Make (each) new variable a factor" with an unchecked checkbox; and "Enter recode directives" with the text: "Vacuum" = 0 and "Mixed" = 1. The right window is titled "MeatBacteria" and displays a data table with 6 rows and 4 columns: "Condition", "LogBacteria", and "X".

	Condition	LogBacteria	X
1	Vacuum	5.26	0
2	Vacuum	5.44	0
3	Vacuum	5.80	0
4	Mixed	7.41	1
5	Mixed	7.33	1
6	Mixed	7.04	1

Statistics > Fit models > Linear regression ...

```
> BacteriaSLR <- lm(LogBacteria~X, data=MeatBacteria)
> summary(BacteriaSLR)

Call:
lm(formula = LogBacteria ~ X, data = MeatBacteria)

Residuals:
    1     2     3     4     5     6 
-0.24 -0.06  0.30  0.15  0.07 -0.22 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.5000     0.1375  39.989 2.34e-06 ***
X             1.7600     0.1945   9.048 0.000827 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2382 on 4 degrees of freedom
Multiple R-Squared:  0.9534,    Adjusted R-squared:  0.9418 
F-statistic: 81.87 on 1 and 4 DF,  p-value: 0.0008266
```

Recall that to test two popln means equal we test whether $\beta_1 = 0$.

What do we conclude?

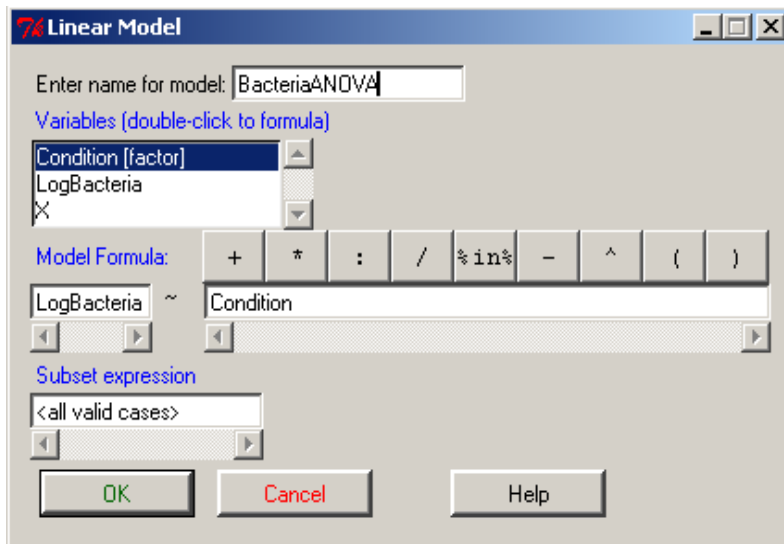
How do these results compare to the Two-Sample T-Test results?

One-Way ANOVA Analysis

Recall that ANOVA tests equality of Means so we will test $\mu_{\text{Vacuum}} = \mu_{\text{Mixed}}$.

Using RCommander with *MeatBacteria* the active data set.

Statistics > Fit models > Linear model ...



```
> BacteriaANOVA <- lm(LogBacteria ~ Condition , data=MeatBacteria)
> summary(BacteriaANOVA)
```

Call:

```
lm(formula = LogBacteria ~ Condition, data = MeatBacteria)
```

Residuals:

```
    1    2    3    4    5    6
-0.24 -0.06  0.30  0.15  0.07 -0.22
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	7.2600	0.1375	52.785	7.71e-07	***
Condition[T.Vacuum]	-1.7600	0.1945	-9.048	0.000827	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2382 on 4 degrees of freedom

Multiple R-squared: 0.9534, Adjusted R-squared: 0.9418

F-statistic: 81.87 on 1 and 4 DF, p-value: 0.0008266

What do we conclude? Which test result do we use?

How do these results compare to the Two-Sample T-Test AND Regression results?

Two-Sample T-Test Analysis

```

Two Sample t-test

data:  LogBacteria by Condition
t = 9.0485, df = 4, p-value = 0.0008266
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.219959 2.300041
sample estimates:
mean in group Mixed mean in group Vacuum
           7.26                5.50
    
```

Regression Analysis

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.5000     0.1375   39.989 2.34e-06 ***
X              1.7600     0.1945    9.048 0.000827 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2382 on 4 degrees of freedom
Multiple R-Squared:  0.9534,    Adjusted R-squared:  0.9418
F-statistic: 81.87 on 1 and 4 DF,  p-value: 0.0008266
    
```

One-Way ANOVA Analysis

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.2600     0.1375   52.785 7.71e-07 ***
Condition[T.Vacuum] -1.7600     0.1945   -9.048 0.000827 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2382 on 4 degrees of freedom
Multiple R-Squared:  0.9534,    Adjusted R-squared:  0.9418
F-statistic: 81.87 on 1 and 4 DF,  p-value: 0.0008266
    
```

Summary of Results

Analysis	Test statistic	P-value	Comment
T-test	$t_{obs}=9.0485$	0.0008266	
Regression	$t_{obs}=9.048$ $F_{obs}=81.87$	0.000827	Test of " $\beta_1=0$ " in model
ANOVA	$F_{obs}=81.87$	0.0008266	Test of " $\mu_1 = \mu_2$ "

What's the Moral you should take away from these results?

```

options ls=110 formdlim="-" nocenter nodate;
data meat;
  input condition $ logcount @@;
  ivac = (condition="vacuum");
  imix = (condition="mixed");
  datalines;
vacuum 5.26  vacuum 5.44  vacuum 5.80
mixed 7.41  mixed 7.33  mixed 7.04
;

title "Log(bacteria count) for different packaging conditions";

proc boxplot;
title2 "Boxplots of log(count)";
  plot logcount*condition;

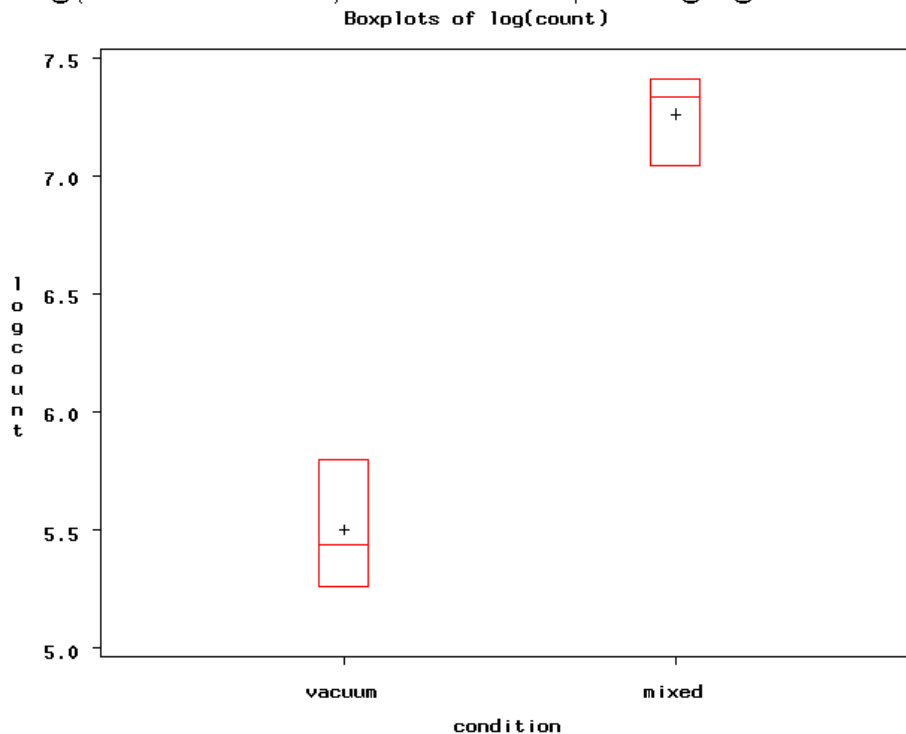
proc ttest;
title2 "T-test comparing mix to vacuum conditions";
  class condition;
  var logcount;

proc reg;
title2 "Regression with indicator variable for mix condition";
  model logcount = imix;

proc glm;
title2 "One-way anova model";
  class condition;
  model logcount = condition;
run;

```

Log(bacteria count) for different packaging conditions



Log(bacteria count) for different packaging conditions

1

T-test comparing mix to vacuum conditions

The TTEST Procedure

		Statistics							
Variable	condition	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err
logcount	mixed	3	6.7764	7.26	7.7436	0.1014	0.1947	1.2235	0.1124
logcount	vacuum	3	4.817	5.5	6.183	0.1432	0.275	1.728	0.1587
logcount	Diff (1-2)		1.22	1.76	2.3	0.1427	0.2382	0.6845	0.1945

T-Tests

Variable	Method	Variances	DF	t Value	Pr > t
logcount	Pooled	Equal	4	9.05	0.0008
logcount	Satterthwaite	Unequal	3.6	9.05	0.0013

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F
logcount	Folded F	2	2	1.99	0.6678

Log(bacteria count) for different packaging conditions

2

Regression with indicator variable for mix condition

The REG Procedure

Model: MODEL1

Dependent Variable: logcount

Number of Observations Read 6
 Number of Observations Used 6

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4.64640	4.64640	81.87	0.0008
Error	4	0.22700	0.05675		
Corrected Total	5	4.87340			

Root MSE 0.23822 R-Square 0.9534
 Dependent Mean 6.38000 Adj R-Sq 0.9418
 Coeff Var 3.73390

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	5.50000	0.13754	39.99	<.0001
imix	1	1.76000	0.19451	9.05	0.0008

Log(bacteria count) for different packaging conditions One-way anova model	3				
The GLM Procedure					
Class Level Information					
Class	Levels Values				
condition	2 mixed vacuum				
Number of Observations Read	6				
Number of Observations Used	6				

Log(bacteria count) for different packaging conditions One-way anova model	4				
The GLM Procedure					
Dependent Variable: logcount					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4.64640000	4.64640000	81.87	0.0008
Error	4	0.22700000	0.05675000		
Corrected Total	5	4.87340000			
R-Square	Coeff Var	Root MSE	logcount Mean		
0.953421	3.733896	0.238223	6.380000		
Source	DF	Type I SS	Mean Square	F Value	Pr > F
condition	1	4.64640000	4.64640000	81.87	0.0008
Source	DF	Type III SS	Mean Square	F Value	Pr > F
condition	1	4.64640000	4.64640000	81.87	0.0008

Summary of Results and some Comments

	Test statistic	P-value	Comment
T-test	$t_{obs}=9.05$	0.0008	Test of " $\mu_1 = \mu_2$ " - note unequal variance t-test has same value test statistic [b/c sample sizes are the same]; however, slight modification in degrees of freedom
Regression	$t_{obs}=9.05$ $F_{obs}=81.87$	0.0008	Test of " $\beta_1=0$ " in model $logcount = \beta_0 + \beta_1 I[condition="mix"] + \varepsilon$
One-way ANOVA	$F_{obs}=81.87$	0.0008	Test of " $\mu_1 = \mu_2$ "