STAT 401A - Statistical Methods for Research Workers

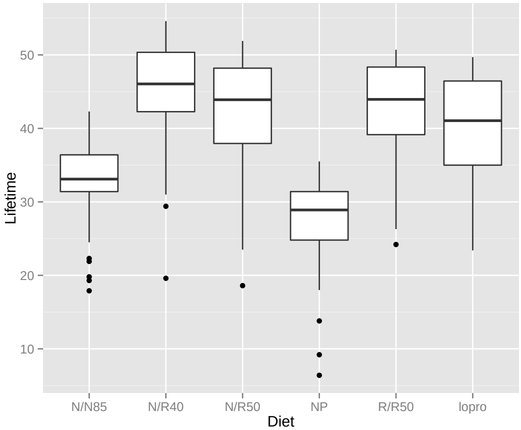
One-way ANOVA

Jarad Niemi (Dr. J)

Iowa State University

last updated: October 3, 2014

Lifetime (months) of mice on different diets



One-way ANOVA model/assumptions

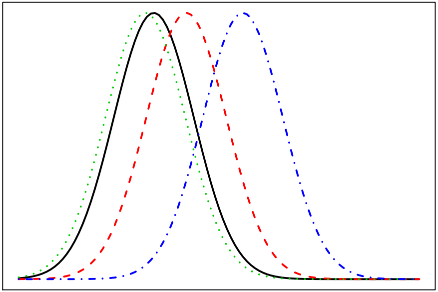
$$Y_{ij} \stackrel{ind}{\sim} N\left(\mu_j, \sigma^2\right)$$

for $j = 1, \ldots, J$ and $i = 1, \ldots, n_j$.
(n_j means there can be different # of observations in each group)

Assumptions:

- Normality
  - Not skewed
  - Not heavy-tailed
- Common variance for all groups
- Independence
  - No cluster effects
  - No serial effects
  - No spatial effects

ANOVA assumptions graphically



What if you want to compare two groups?

We may still be interested in comparing two groups.

Statistical hypothesis: Is there a difference in mean lifetimes between the mice in two groups, e.g. NP and N/N85?

Statistical question: What is the difference in mean lifetimes between the mice in two groups, e.g. NP and N/N85?

Two-group analysis

Begin with the two group (equal variance) model:

$$Y_{ij} \stackrel{ind}{\sim} N\left(\mu_j, \sigma^2\right)$$

with $j = 1, 2$ and $i = 1, \ldots, n_j$
To perform a hypothesis test or a CI for the difference in means, the relevant quantities are:

- $\overline{Y}_2 - \overline{Y}_1$
- $SE(\overline{Y}_2 - \overline{Y}_1) = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
- t distribution with $n_1 + n_2 - 2$ degrees of freedom

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n1 + n_2 - 2}$$

What if you have more than two groups?

Multi-group analysis

The multi-group (equal variance) model:

$$Y_{ij} \stackrel{ind}{\sim} N\left(\mu_j, \sigma^2\right)$$

but now $j = 1, \ldots, J$ and $i = 1, \ldots, n_j$
(n_j means there can be different # of observations in each group)
To perform a hypothesis test or a CI for the difference in means, the relevant quantities are:

- $\overline{Y}_2 - \overline{Y}_1$
- $SE(\overline{Y}_2 - \overline{Y}_1) = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
- t distribution with $n_1 + n_2 + \cdots + n_J - J$ degrees of freedom

where

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \cdots + (n_J-1)s_J^2}{n1 + n_2 + \cdots + n_J - J}$$

Hypothesis test for comparison of two means (in multi-group data)

If $Y_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma^2)$ for $j = 1, \ldots, J$ and we want to test the hypothesis

- $H_0 : \mu_1 = \mu_2$
- $H_1 : \mu_1 \neq \mu_2$

then we compute:

$$t = \frac{\overline{Y}_1 - \overline{Y}_2}{SE(\overline{Y}_1 - \overline{Y}_2)}$$

where

$$SE(\overline{Y}_1 - \overline{Y}_2) = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

and

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_J - 1)s_J^2}{n_1 + n_2 + \cdots + n_J - J}.$$

Then we compare t to a t distribution with $n_1 + n_2 + \cdots + n_J - J$ degrees of freedom.

Diet effect on mice lifetime

Table: Summary statistics for mice lifetime (months) on different diets

	Diet	n	mean	sd
1	N/N85	57	32.7	5.1
2	N/R40	60	45.1	6.7
3	N/R50	71	42.3	7.8
4	NP	49	27.4	6.1
5	R/R50	56	42.9	6.7
6	lopro	56	39.7	7.0

Test for difference in mean lifetime between NP and N/N85, i.e.

$$H_0 : \mu_4 = \mu_1 \text{ vs } H_1 : \mu_4 \neq \mu_1.$$

Showing work

$$
\begin{aligned}
\overline{Y}_1 - \overline{Y}_4 &= 32.7 - 27.4 = 5.3 \\
df &= 57 + 60 + 71 + 49 + 56 + 56 - 6 = 343 \\
s_p^2 &= \frac{(57-1)5.1^2+(60-1)6.7^2+(71-1)7.8^2+(49-1)6.1^2+(56-1)6.7^2+(56-1)7.0^2}{57+60+71+49+56+56-6} \\
&= \frac{15314}{343} = 44.6 \\
s_p &= \sqrt{s_p^2} = \sqrt{44.6} = 6.7 \\
SE(\overline{Y}_1 - \overline{Y}_4) &= s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_4}} = 6.7\sqrt{\frac{1}{57} + \frac{1}{49}} = 1.3 \\
t &= \frac{\overline{Y}_1 - \overline{Y}_4}{SE(\overline{Y}_1 - \overline{Y}_4)} = \frac{5.3}{1.2} = 4.1 \\
p &= 2P(t_{343} < -|t|) = 2P(t_{343} < -4.1) = 0.000052
\end{aligned}
$$

So we reject the null hypothesis that there is no difference between mean lifetime of mice on the NP and N/N85 diets.

Confidence interval for the difference of two means (in multi-group data)

If $Y_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma^2)$ for $j = 1, \ldots, J$, a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

$$\overline{Y}_1 - \overline{Y}_2 \pm t_{n_1+n_2+\cdots+n_J-J}(1-\alpha/2) SE(\overline{Y}_1 - \overline{Y}_2)$$

where the t critical value, $t_{n_1+n_2+\cdots+n_J-J}(1-\alpha/2)$, needs to be calculated using a statistical software.

A 95% confidence interval for the difference in mean lifetime for N/N85 minus NP ($\mu_1 - \mu_4$) is

$$5.3 \pm 1.96 \times 1.3 = (2.8, 7.8).$$

The statistical conclusion would be

> *In this study, mice on the N/N85 diet lived an average of 5.3 months longer than mice on the NP diet (95% CI (2.8,7.8)).*

```
DATA mice;
  INFILE 'case0501.csv' DSD FIRSTOBS=2;
  INPUT lifetime diet $;

PROC GLM DATA=mice;
  CLASS diet;
  MODEL lifetime = diet;
  LSMEANS diet / ADJUST=T CL;
  RUN;
```

The GLM Procedure
Least Squares Means

diet	lifetime LSMEAN	LSMEAN Number
N/N85	32.6912281	1
N/R40	45.1166667	2
N/R50	42.2971831	3
NP	27.4020408	4
R/R50	42.8857143	5
lopro	39.6857143	6

Least Squares Means for effect diet
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: lifetime

i/j	1	2	3	4	5	6
1		<.0001	<.0001	<.0001	<.0001	<.0001
2	<.0001		0.0166	<.0001	0.0731	<.0001
3	<.0001	0.0166		<.0001	0.6223	0.0293
4	<.0001	<.0001	<.0001		<.0001	<.0001
5	<.0001	0.0731	0.6223	<.0001		0.0117

lifetime

diet	LSMEAN	95% Confidence Limits	
N/N85	32.691228	30.951394	34.431062
N/R40	45.116667	43.420886	46.812447
N/R50	42.297183	40.738291	43.856075
NP	27.402041	25.525547	29.278535
R/R50	42.885714	41.130415	44.641014
lopro	39.685714	37.930415	41.441014

Least Squares Means for Effect diet

i	j	Difference Between Means	95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-12.425439	-14.854984	-9.995893
1	3	-9.605955	-11.942013	-7.269897
1	4	5.289187	2.730232	7.848142
1	5	-10.194486	-12.665943	-7.723030
1	6	-6.994486	-9.465943	-4.523030
2	3	2.819484	0.516048	5.122919
2	4	17.714626	15.185417	20.243835
2	5	2.230952	-0.209692	4.671597
2	6	5.430952	2.990308	7.871597
3	4	14.895142	12.455599	17.334686
3	5	-0.588531	-2.936130	1.759068
3	6	2.611469	0.263870	4.959068
4	5	-15.483673	-18.053169	-12.914178
4	6	-12.283673	-14.853169	-9.714178
5	6	3.200000	0.717632	5.682368

One-way ANOVA F-test

Are any of the means different?

Hypotheses in English:

H_0: all the means are the same

H_1: at least one of the means is different

Statistical hypotheses:

$$\begin{array}{llll} H_0: & \mu_j = \mu \text{ for all } j & \qquad & Y_{ij} \stackrel{iid}{\sim} N(\mu, \sigma^2) \\ H_1: & \mu_j \neq \mu_{j'} \text{ for some } j \text{ and } j' & \qquad & Y_{ij} \stackrel{ind}{\sim} N\left(\mu_j, \sigma^2\right) \end{array}$$

An ANOVA table organizes the relevant quantities for this test and computes the pvalue.

ANOVA table

A start of an ANOVA table:

Source of variation	Sum of squares	d.f.	Mean square
Factor A (Between groups)	$SSA = \sum_{j=1}^{J} n_j \left(\overline{Y}_j - \overline{Y}\right)^2$	$J-1$	$\frac{SSA}{J-1}$
Error (Within groups)	$SSE = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \left(Y_{ij} - \overline{Y}_j\right)^2$	$n-J$	$\frac{SSE}{n-J}$ $(= s_p^2)$
Total	$SST = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \left(Y_{ij} - \overline{Y}\right)^2$	$n-1$	

where

- J is the number of groups,
- n_j is the number of observations in group j,
- $n = \sum_{j=1}^{J} n_j$ (total observations),
- $\overline{Y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ij}$ (average in group j),
- and $\overline{Y} = \frac{1}{n} \sum_{j=1}^{J} \sum_{i=1}^{n_j} Y_{ij}$ (overall average).

ANOVA table

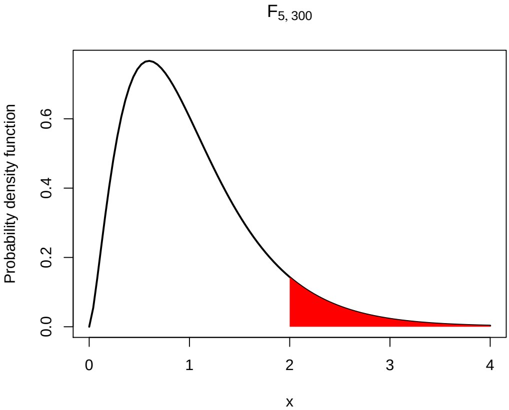
An easier to remember ANOVA table:

Source of variation	Sum of squares	df	Mean square	F-statistic	p-value
Factor A (between groups)	SSA	$J-1$	MSA $=$ SSA$/J-1$	MSA/MSE	(see below)
Error (within groups)	SSE	$n-J$	MSE $=$ SSE$/n-J$		
Total	SST=SSA+SSE	$n-1$			

Under H_0,

- the quantity MSA/MSE has an F-distribution with $J-1$ numerator and $n-J$ denominator degrees of freedom,
- larger values of MSA/MSE indicate evidence against H_0, and
- the p-value is determined by $P(F_{J-1,n-J} > MSA/MSE)$.

F-distribution



One-way ANOVA F-test (by hand)

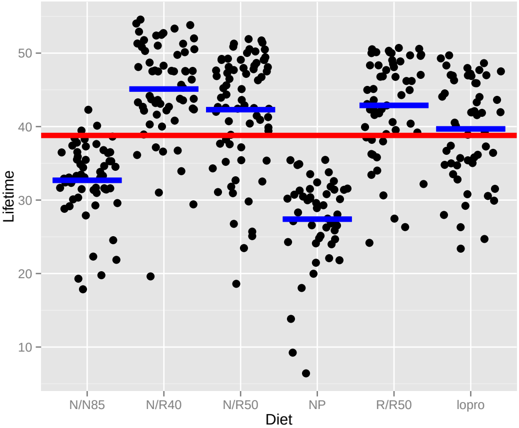
Table: Summary statistics for mice lifetime (months) on different diets

	Diet	n	mean	sd
1	N/N85	57	32.7	5.1
2	N/R40	60	45.1	6.7
3	N/R50	71	42.3	7.8
4	NP	49	27.4	6.1
5	R/R50	56	42.9	6.7
6	lopro	56	39.7	7.0
7	Total	349	38.8	

So

$$
\begin{aligned}
SSA =& \quad 57 \times (32.7 - 38.8)^2 + 60 \times (45.1 - 38.8)^2 + 71 \times (42.3 - 38.8)^2 + 49 \times (27.4 - 38.8)^2 \\
& +56 \times (42.9 - 38.8)^2 + 56 \times (39.7 - 38.8)^2 = 12734 \\
SST =& \quad (35.5 - 38.8)^2 + (35.4 - 38.8)^2 + (34.9 - 38.8)^2 + \cdots + (19.6 - 38.8)^2 + (47.6 - 38.8)^2 = 28031 \\
SSE =& \quad SST - SSA = 28031 - 12734 = 15297 \\
J - 1 =& \quad 5 \\
n - J =& \quad 349 - 6 = 343 \\
n - 1 =& \quad 348 \\
MSA =& \quad SSA/J - 1 = 12734/5 = 2547 \\
MSE =& \quad SSE/n - J = 15297/343 = 44.6 = s_p^2 \\
F =& \quad MSA/MSE = 2547/44.6 = 57.1 \\
p =& \quad P(F_{5,343} > 57.1) < 0.0001
\end{aligned}
$$

As a picture



SAS code and output for one-way ANOVA

```
DATA mice;
  INFILE 'case0501.csv' DSD FIRSTOBS=2;
  INPUT lifetime diet $;

PROC GLM DATA=mice;
  CLASS diet;
  MODEL lifetime = diet;
  RUN;

                                        The GLM Procedure

Dependent Variable: lifetime

                                                Sum of
       Source                      DF          Squares     Mean Square    F Value    Pr > F
       Model                        5      12733.94181      2546.78836      57.10    <.0001
       Error                      343      15297.41532        44.59888
       Corrected Total            348      28031.35713
```

R code and output for one-way ANOVA

```
m = lm(Lifetime~Diet, case0501)
anova(m)


Analysis of Variance Table

Response: Lifetime
           Df Sum Sq Mean Sq F value Pr(>F)
Diet        5  12734    2547    57.1 <2e-16 ***
Residuals 343  15297      45
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

General F-tests

The one-way ANOVA F-test is an example of a general hypothesis testing framework that uses F-tests. This framework can be used to test

- composite alternative hypothesesor, equivalently,
- a full vs a reduced model.

The general idea is to balance the amount of variability remaining when moving from the reduced model to the full model measured using the sums of squared errors (SSEs) relative to the amount of complexity, i.e. parameters, added to the model.

Simple vs Composite Hypotheses

Suppose

$$Y_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma^2)$$

for $j = 1, \ldots, 3$ then a simple hypothesis is

- $H_0 : \mu_1 = \mu_2$
- $H_1 : \mu_1 \neq \mu_2$

and a composite hypothesis is

- $H_0 : \mu_1 = \mu_2 = \mu_3$
- $H_1 : \mu_j \neq \mu_{j'}$ for some j and j'

since there are four possibilities under H_1

- $\mu_1 = \mu_2 \neq \mu_3$
- $\mu_2 = \mu_3 \neq \mu_1$
- $\mu_3 = \mu_1 \neq \mu_2$
- none of μ_1, μ_2, μ_3 are equal

Testing Composite hypotheses

If $Y_{ij} \overset{ind}{\sim} N(\mu_j, \sigma^2)$ for $j = 1, \ldots, J$ and we want to test the composite hypothesis

- $H_0 : \mu_j = \mu$ for all j
- $H_1 : \mu_j \neq \mu_{j'}$ for some j and j'

think about this as two models:

- $H_0 : Y_{ij} \overset{ind}{\sim} N(\mu, \sigma^2)$ (reduced)
- $H_1 : Y_{ij} \overset{ind}{\sim} N(\mu_j, \sigma^2)$ (full)

We can use an F-test to calculate a p-value for tests of this type.

Nested models: full vs reduced

Definition

Two models are nested if the reduced model is a special case of the full model.

For example, consider the full model

$$Y_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma^2).$$

One special case of this model occurs when $\mu_j = \mu$ and thus

$$Y_{ij} \stackrel{ind}{\sim} N(\mu, \sigma^2).$$

is a reduced model and these two models are nested.

Calculating the sum of squared residuals (errors)

Model	*Full*	*Reduced*
Assumption	$H_1 : Y_{ij} \overset{ind}{\sim} N\left(\mu_j, \sigma^2\right)$	$H_0 : Y_{ij} \overset{iid}{\sim} N(\mu, \sigma^2)$
Mean	$\hat{\mu}_j = \overline{Y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ij}$	$\hat{\mu} = \overline{Y} = \frac{1}{n} \sum_{j=1}^{J} \sum_{i=1}^{n_j} Y_{ij}$
Residual	$r_{ij} = Y_{ij} - \hat{\mu}_j = Y_{ij} - \overline{Y}_j$	$r_{ij} = Y_{ij} - \hat{\mu} = Y_{ij} - \overline{Y}$
SSE	$\sum_{j=1}^{J} \sum_{i=1}^{n_j} r_{ij}^2$	$\sum_{j=1}^{J} \sum_{i=1}^{n_j} r_{ij}^2$

General F-tests

Do the following

1. Calculate

   Extra sum of squares =
   Residual sum of squares (reduced) - Residual sum of squares (full)

2. Calculate

   Extra degrees of freedom =
   # of mean parameters (full) - # of mean parameters (reduced)

3. Calculate

$$\text{F-statistic} = \frac{\text{Extra sum of squares / Extra degrees of freedom}}{\hat{\sigma}^2_{full}}$$

4. Compare this to an F-distribution with
   - numerator degrees of freedom = Extra degrees of freedom
   - denominator degrees of freedom = n - # of mean parameters (full)

Example

Recall the mice data set.
Consider the hypothesis that all diets have a common mean lifetime except NP.

Let

$$Y_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma^2)$$

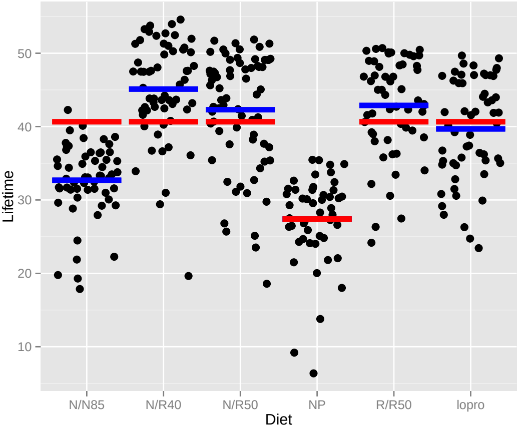
with $j = 1$ being the NP group then the hypotheses are

- $H_0 : \mu_j = \mu$ for $j \neq 1$
- $H_1 : \mu_j \neq \mu_{j'}$ for some $j, j' = 2, \ldots, 6$

As models:

- $H_0 : Y_{i1} \sim N(\mu_1, \sigma^2)$ and $Y_{ij} \sim N(\mu, \sigma^2)$ for $j \neq 1$
- $H_1 : Y_{ij} \sim N(\mu_j, \sigma^2)$

As a picture



```
DATA mice;
  INFILE 'case0501.csv' DSD FIRSTOBS=2;
  INPUT lifetime diet $;
  IF diet='NP' THEN NP=1; ELSE NP=0;

PROC PRINT DATA=mice; RUN;

PROC GLM DATA=mice;
  CLASS diet;
  MODEL lifetime = diet;
  TITLE 'Full Model';
  RUN;

PROC GLM DATA=mice;
  CLASS NP;
  MODEL lifetime = NP;
  TITLE 'Reduced Model';
  RUN;
```

```
                            Full Model

                        The GLM Procedure

Dependent Variable: lifetime

                                        Sum of
       Source                DF        Squares     Mean Square    F Value    Pr > F
       Model                  5    12733.94181      2546.78836      57.10    <.0001
       Error                343    15297.41532        44.59888
       Corrected Total      348    28031.35713


                           Reduced Model

                        The GLM Procedure

Dependent Variable: lifetime

                                        Sum of
       Source                DF        Squares     Mean Square    F Value    Pr > F
       Model                  1     7401.77817      7401.77817     124.50    <.0001
       Error                347    20629.57896        59.45124
       Corrected Total      348    28031.35713
```

General F-test calculations

$$\begin{aligned} ESS &= 20629.57896 - 15297.41532 = 5332.164 \\ Edf &= 5 - 1 = 4 \\ F &= (ESS/Edf)/\hat{\sigma}^2_{full} = (5332.164/4)/44.59888 = 29.88956 \end{aligned}$$

Finally, we calculate the pvalue (using statistical software):

$$P(F_{4,343} > F) < 0.0001$$

Since this is very small, we reject the null hypothesis that the reduced model is adequate. So there is evidence that the mean is not the same for all the non-NP groups.

Making SAS do the calculations

```
DATA mice;
  INFILE 'case0501.csv' DSD FIRSTOBS=2;
  INPUT lifetime diet $;
  IF diet='NP' THEN NP=1; ELSE NP=0;

PROC GLM DATA=mice;
  CLASS diet NP;
  MODEL lifetime = NP diet(NP);
  RUN;
```

```
                                        The GLM Procedure

Dependent Variable: lifetime

                                                 Sum of
      Source                        DF          Squares     Mean Square    F Value    Pr > F
      Model                          5      12733.94181      2546.78836      57.10    <.0001
      Error                        343      15297.41532        44.59888
      Corrected Total              348      28031.35713


      Source                        DF      Type III SS     Mean Square    F Value    Pr > F
      NP                             1      7256.758693     7256.758693     162.71    <.0001
      diet(NP)                       4      5332.163640     1333.040910      29.89    <.0001
```

Making R do the calculations

```
case0501$NP = factor(case0501$Diet == "NP")

modR = lm(Lifetime~NP,   case0501)
modF = lm(Lifetime~Diet, case0501)
anova(modR,modF)


Analysis of Variance Table

Model 1: Lifetime ~ NP
Model 2: Lifetime ~ Diet
  Res.Df   RSS Df Sum of Sq    F Pr(>F)
1    347 20630
2    343 15297  4      5332 29.9 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Are there differences in means amongst low calorie diets?

Let Y_{ij} be the lifetime in months for mouse i in group j where the groups are N/N85 (j=1), N/R40 (j=2), N/R50 (j=3), NP (j=4), R/R50 (j=5), and lopro (j=6). Assume

$$Y_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma^2)$$

and test the hypotheses

H_0: $\mu_2 = \mu_3 = \mu_5 = \mu_6$

H_1: at least one of $\mu_2, \mu_3, \mu_5, \mu_6$ is different from the rest

Implicitly, we are allowing μ_1 and μ_4 to be different from the others.

Making SAS do the calculations

```
DATA mice;
  INFILE 'case0501.csv' DSD FIRSTOBS=2;
  INPUT lifetime diet $;
  IF diet='N/N85' THEN local=1; ELSE local=2; /* NP is 2 here */
  IF diet='NP' THEN local=0;                  /* NP is now 0  */

/* I needed to run this PROC PRINT to set the data up appropriately
PROC PRINT DATA=mice; RUN;
*/

PROC GLM DATA=mice;
  CLASS diet local;
  MODEL lifetime = local diet(local);
  RUN;
```

```
                                The GLM Procedure

Dependent Variable: lifetime

                                            Sum of
      Source                     DF        Squares     Mean Square    F Value    Pr > F
      Model                       5    12733.94181      2546.78836      57.10    <.0001
      Error                     343    15297.41532        44.59888
      Corrected Total           348    28031.35713

      Source                     DF      Type I SS     Mean Square    F Value    Pr > F
      local                       2    11868.52098      5934.26049     133.06    <.0001
      diet(local)                 3      865.42083       288.47361       6.47    0.0003
```

Making R do the calculations

```
case0501$local = ifelse(case0501$Diet=='N/N85', 1, 2) # NP is 2 here
case0501$local[case0501$Diet=='NP'] = 0               # now NP is 1
case0501$local = factor(case0501$local)
mod1 = lm(Lifetime~1,     case0501)
modR = lm(Lifetime~local, case0501)
modF = lm(Lifetime~Diet,  case0501)
anova(mod1, modR, modF)


Analysis of Variance Table

Model 1: Lifetime ~ 1
Model 2: Lifetime ~ local
Model 3: Lifetime ~ Diet
  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1    348 28031
2    346 16163  2     11869 133.06 < 2e-16 ***
3    343 15297  3       865   6.47 0.00029 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


anova(modF) # To get the pooled estimate of the variance for the full model


Analysis of Variance Table

Response: Lifetime
           Df Sum Sq Mean Sq F value Pr(>F)
Diet        5  12734    2547    57.1 <2e-16 ***
Residuals 343  15297      45
---
```

Summary

- Use t-tests for simple hypothesis tests and CIs
- Use F-tests for composite hypothesis tests
  - One-way ANOVA F-test
  - General F-tests

Think about F-tests as comparing models.