General F-tests

STAT 401 - Statistical Methods for Research Workers

Jarad Niemi

Iowa State University

20 September 2013

Simple vs Composite Hypotheses

Suppose

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

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

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

and a composite hypothesis is

- $H_0 : \mu_1 = \mu_2 = \mu_3$
- $H_1 : \mu_i \neq \mu_j$ for some $i \neq j$

since there are four possibilities under H_1

- $\mu_1 = \mu_2 \neq \mu_3$
- $\mu_2 = \mu_3 \neq \mu_1$
- $\mu_3 = \mu_1 \neq \mu_2$
- $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_1$

Equality of two means

If $Y_{ij} \stackrel{ind}{\sim} N(\mu_i, \sigma^2)$ for $i = 1, \ldots, \mathrm{I}$ and we want to test the simple hypothesis

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

then we use the same t-test and confidence interval formulas from the two-sample t-test:

$$t = \frac{\overline{Y}_1 - \overline{Y}_2}{SE(\overline{Y}_1 - \overline{Y}_2)} \qquad CI : \overline{Y}_1 - \overline{Y}_2 \pm t_{df}(1-\alpha/2) SE(\overline{Y}_1 - \overline{Y}_2)$$

where

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

with two modifications:

$$\begin{aligned} df &= n_1 + n_2 + \cdots + n_{\mathrm{I}} - \mathrm{I} \\ s_p^2 &= \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \cdots + (n_{\mathrm{I}}-1)s_{\mathrm{I}}^2}{n_1 + n_2 + \cdots + n_{\mathrm{I}} - \mathrm{I}} = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \cdots + (n_{\mathrm{I}}-1)s_{\mathrm{I}}^2}{df} \end{aligned}$$

```
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

                     lifetime      LSMEAN
        diet           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

Testing Composite hypotheses

If $Y_{ij} \stackrel{ind}{\sim} N(\mu_i, \sigma^2)$ for $i = 1, \ldots, \mathrm{I}$ and we want to test the composite hypothesis

- $H_0 : \mu_i = \mu_j$
- $H_1 : \mu_i \neq \mu_j$ for some $i \neq j$

think about this as two models:

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

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

F-tests

Do the following

1. Calculate

$$\text{Extra sum of squares} = \text{Residual sum of squares (reduced) - Residual sum of squares (full)}$$

2. Calculate

$$\text{Extra degrees of freedom} = \text{\# of mean parameters (full) - \# of mean parameters (reduced)}$$

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 except NP have a common mean and this mean is different from the NP.

Let

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

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

- $H_0 : \mu_i = \mu$ for $i \neq 1$
- $H_1 : \mu_i \neq \mu_j$ for $i \neq j$ and $i, j = 2, \ldots, 6$

As models:

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

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

PROC PRINT DATA=mice; RUN;

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

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

Full Model

The GLM Procedure

Dependent Variable: lifetime

Source	DF	Sum of 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

Source	DF	Sum of 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:

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

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.

Summary

- Use t-tests for simple hypothesis tests and CIs
- Use F-tests for composite hypothesis tests
- Think about tests as comparing models
- For F-test, fit both models and compute the pvalue