

One-Way Analysis of Variance

Applied Regression and Other Multivariable Methods
Sections 17-1 - 17-6

One-way ANOVA Overview

- Interested in comparing **several** population **means**
- Commonly framed in terms of comparing
 - Several treatments/factors
 - Several levels of one treatment/factor
- If k treatments, could do $k(k - 1)/2$ t-tests
 - Does not test equality of all means at once
 - Multiple tests → More chance of Type I error
 - Constant variance? Error df?
- ANOVA provides method of joint inference
 $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$
- ANOVA considered special case of regression
- Trts compared through dummy variable regression
- Examples
 - SMK - smokers vs nonsmokers on SBP
 - SPECIES - species A vs Species B on CAL

Example

- Consider smoking history with categories
 1. Have never smoked
 2. Quit smoking more than a two years ago
 3. Other
- Define dummy variables to be

$$Z_1 = \begin{cases} 1 & \text{if Category 1} \\ 0 & \text{otherwise} \end{cases} \quad Z_2 = \begin{cases} 1 & \text{if Category 2} \\ 0 & \text{otherwise} \end{cases}$$
- Fit the regression model

$$Y_i = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + E_i$$
- Can show

$$\hat{\beta}_0 = \bar{Y}_3, \quad \hat{\beta}_1 = \bar{Y}_1 - \bar{Y}_3, \quad \hat{\beta}_2 = \bar{Y}_2 - \bar{Y}_3$$
- $H_0 : \beta_2 = 0$ tests equality of category 2 vs 3 means
- $H_0 : \beta_1 = 0$ tests equality of category 1 vs 3 means
- $H_0 : \beta_1 - \beta_2 = 0$ tests equality of 1 vs 2 means
- $H_0 : \beta_1 = \beta_2 = 0$ tests equality of **all means**

Analysis of Variance Table

- $H_0 : \beta_1 = \beta_2 = 0$ is simply overall F test
- Can use ANOVA from regression output
- Model SS ↔ Between SS, Resid SS ↔ Within SS

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Between	SS(Trt)	$k - 1$	MST	MST/MSE
Within	SSE	$N - k$	MSE	
Total	SS(Total)	$N - 1$		

- SS calculations given dummy variable regressors

If balanced (equal n):

$$SS(\text{Trt}) = n \sum (\bar{Y}_i - \bar{Y})^2$$

$$SSE = (n - 1) \sum S_i^2$$

If unbalanced (various n_i):

$$SS(\text{Trt}) = \sum (T_i^2 / n_i) - G^2 / N$$

$$SSE = \sum \sum Y_{ij}^2 - \sum (T_i^2 / n_i)$$

$$T_i = n_i \bar{Y}_i \text{ and } G = \sum T_i$$

If $F > F_{\alpha, k-1, N-k}$ then reject H_0

Numeric Demonstration

Consider this SMK comparison with $n = 3$ people per category

Category 1	Category 2	Category 3
20	24	26
23	26	27
22	25	27

$$\begin{aligned}\bar{Y}_1 &= 21.67 & \bar{Y}_2 &= 25.00 & \bar{Y}_3 &= 26.67 \\ S_1^2 &= 2.33 & S_2^2 &= 1.00 & S_3^2 &= 0.33 \\ & & \bar{Y} &= 24.44 & & \end{aligned}$$

Category 1 vs 3 : Difference is -5.00

Category 2 vs 3 : Difference is -1.67

Category 1 vs 2 : Difference is -3.33

Sum of squares **between** categories is

$$n \sum (\bar{Y}_i - \bar{Y})^2 = 3((21.67 - 24.44)^2 + (25.00 - 24.44)^2 + (26.67 - 24.44)^2) = 38.89$$

$$MST = \frac{38.89}{3-1} = 19.44$$

Sum of squares **within** categories is

$$(n - 1) \sum S_i^2 = 2(2.33 + 1.00 + 0.33) = 7.33$$

$$MSE = \frac{7.33}{3(3-1)} = 1.22$$

Assuming equal variance, can compute std error for diff in means

$$S_{\bar{Y}_i - \bar{Y}_j} = \sqrt{2(1.22)/3} = 0.903$$

18-4

Using SAS

```
data one;
input smk response @@;
cards;
1 20 1 23 1 22
2 24 2 26 2 25
3 26 3 27 3 27
;
```

```
data two;
set one;
if smk = 1 then x1=1;
else x1=0;
if smk = 2 then x2=1;
else x2=0;
```

```
proc reg data=two;
model response=x1 x2;
```

```
-----
Analysis of Variance

Source      DF      Sum of      Mean      F Value      Prob>F
Model       2      38.88889    19.44444    15.909      0.0040
Error       6       7.33333     1.22222
C Total     8      46.22222

Root MSE    1.10554    R-square    0.8413
Dep Mean    24.44444    Adj R-sq    0.7885
C.V.        4.52267
```

```
Parameter Estimates

Variable    DF      Parameter      Standard      T for H0:      Prob > |T|
INTERCEP   1      26.666667      0.63828474    41.779      0.0001
X1          1      -5.000000      0.90267093    -5.539      0.0015
X2          1      -1.666667      0.90267093    -1.846      0.1144
```

18-5

Summary

- Since ANOVA special case of regression, assumes
 - Independence
 - Normal Errors
 - Constant variance
- Must do the same diagnostic checks as before
- Equal variance implies $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma^2$
- Joint estimate of σ^2 is $MSE=1.22$
- $\bar{\beta}_1$ and $\bar{\beta}_2$ compare means
- $S_{\bar{\beta}_1} = S_{\bar{\beta}_2} = S_{\bar{Y}_i - \bar{Y}_j}$
- $H_0 : \beta_1 = \beta_2 = 0$ comparable to $H_0 : \mu_1 = \mu_2 = \mu_3$
- Hypothesis tested using multiple F test

$$F = \frac{MST}{MSE} = \frac{38.88/2}{7.33/6} = 15.91$$

18-6

Fixed vs Random Factors

- Very important to distinguish
- Random factor
 - Levels in experiment considered a random sample from larger population of levels
 - Effect of level can be considered a random process (e.g., genetic effect)
 - Want to draw inference on larger population of levels
- Fixed factor
 - Levels in experiment selected by some non-random process
 - Restrict inference to just the levels in study

18-7

Examples

[Exp 1] To study the effects of pesticides on a bird's calcium content, an experimenter randomly (and equally) allocated sixty-five chicks to five diets (a control and four with a different pesticide included). After a month each chick's calcium content (mg) in one cm length of bone was measured.

- Y = calcium content
- Chick - Random Factor
- Diet - Fixed Factor

[Exp 2] A psychologist is interested in studying the mean and variance in IQs of 1st grade children from the low income areas of several major cities. Six grade schools were randomly chosen (from the low income areas) and from each of these schools, five 1st grade children were randomly chosen and had their IQs measured.

- Y = IQ score
- School - Random Factor
- Student - Random Factor

18-8

Model Framework

Fixed Effects

$$Y_{ij} = \mu + \alpha_i + E_{ij} \quad \begin{cases} i = 1, 2, \dots, k \\ j = 1, 2, \dots, n_i \end{cases}$$

μ - grand/overall mean

α_i - i th trt effect (difference from overall)

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

μ_i - i th trt mean ($\mu + \alpha_i$)

Consider restriction $\sum \alpha_i = 0$

Testing

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

$$H_A : \text{at least one mean different}$$

Comparable to testing

$$H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_k = 0$$

$$H_A : \alpha_i \neq 0 \text{ for at least one } i$$

18-9

Model Framework

Random Effects

$$Y_{ij} = \mu + A_i + E_{ij} \quad \begin{cases} i = 1, 2, \dots, k \\ j = 1, 2, \dots, n_i \end{cases}$$

μ - grand/overall mean

$$A_i \sim N(0, \sigma_A^2)$$

$$E_{ij} \sim N(0, \sigma^2)$$

A_i 's and E_{ij} 's independent

- $\text{Var}(Y_{ij}) = \sigma_A^2 + \sigma^2$
- $\text{Corr}(Y_{ij}, Y_{ij'}) = \sigma_A^2 / (\sigma_A^2 + \sigma^2)$
- Known as intra-class correlation

Testing

$$H_0 : \sigma_A^2 = 0$$

$$H_A : \sigma_A^2 > 0$$

18-10

Why use F ?

- Under usual assumptions, can show

- Fixed Effects

$$E(\text{MSE}) = \sigma^2$$

$$E(\text{MST}) = \sigma^2 + \sum n_i \alpha_i^2 / (k - 1)$$

- Random Effects

$$E(\text{MSE}) = \sigma^2$$

$$E(\text{MST}) = \sigma^2 + n_0 \sigma_A^2$$

- Under H_0 , MSE and MST are unbiased estimates of σ^2 (ratio should be near 1)
- F is a ratio of "variance" estimates
- If statistic deviates from one, then reject H_0
- When F large, at least one $\alpha_i \neq 0$ or $\sigma_A^2 > 0$
- Can also show F same as t-test when $k = 2$

18-11

Similarity With t-test

- Consider the square of the t-test statistic

$$\begin{aligned}
 T^2 &= \left(\frac{\bar{Y}_1 - \bar{Y}_2}{S_p \sqrt{2/n}} \right)^2 \\
 &= \left(\frac{(\bar{Y}_1 - \bar{Y}) - (\bar{Y}_2 - \bar{Y})}{S_p \sqrt{2/n}} \right)^2 \\
 &= \left(\frac{2(\bar{Y}_1 - \bar{Y})}{S_p \sqrt{2/n}} \right)^2 \\
 &= \left(\frac{4(\bar{Y}_1 - \bar{Y})^2}{S_p^2 (2/n)} \right) \\
 &= \frac{2((\bar{Y}_1 - \bar{Y})^2 + (\bar{Y}_2 - \bar{Y})^2)}{S_p^2 (2/n)} \\
 &= \frac{n((\bar{Y}_1 - \bar{Y})^2 + (\bar{Y}_2 - \bar{Y})^2)}{S_p^2} \\
 &= \frac{MS(\text{Between})}{MS(\text{Within})} = \frac{MST}{MSE}
 \end{aligned}$$

- When $k = 2$, $T^2 = F$
- F-test gives identical results as t-test $H_A \neq$

18-12

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	

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

$$T_1 = 33, T_2 = 75, \text{ and } T_3 = 48$$

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

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

$$SSE = 2274 - (33^2/3 + 75^2/5 + 48^2/4) = 210$$

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

$$P\text{-value} > 0.20 \text{ (DNR)}$$

18-13

Using SAS (lambs.sas)

```

option nocenter ps=65 ls=75;

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
;

/* Generate side by side boxplots */
symbol1 bwidth=5 i=box; axis1 offset=(5);
proc gplot; plot wtgain*diet / frame haxis=axis1;

/* Use GLM instead of REG */
proc glm;
class diet;
model wtgain=diet;
output out=diag r=res p=pred;

/* Generate a residual boxplots */
proc gplot; plot res*diet /frame haxis=axis1;

/* Generate residual plot and horizontal line at zero */
proc sort; by pred;
symbol1 v=circle i=sm50;
proc gplot; plot res*pred / haxis=axis1;
run;

```

18-14

Output

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	36.00000000	18.00000000	0.77	0.4907

Source	DF	Type III SS	Mean Square	F Value	Pr > F
diet	2	36.00000000	18.00000000	0.77	0.4907

18-15