MATH 503: Mathematical Statistics

Lecture 9: Analysis of Variance

Reading: CB Sections 11.1-11.2

(or HMC Sections 9.1-9.2, 9.5)

Kimberly F. Sellers

Department of Mathematics and Statistics

Today's Topics

- Quadratic forms
- One-way ANOVA
- Two-way ANOVA
  - Without interaction
  - With interaction

What is a quadratic form?

- A homogenous polynomial of degree 2 in n variables
- A real quadratic form is one where the variables and coefficients are real
- Examples:
  - $X_1^2 + X_1X_2 + X_2^2$ is a quadratic form in X_1, X_2
  - $X_1^2 + X_2^2 + X_3^2 - 2X_1X_2$ is a quadratic form in X_1, X_2, X_3
  - $(n-1)S^2$is a quadratic form in $X_1, X_2, \dots, X_n$

Theorem

Let $Q = Q_1 + Q_2 + \cdots + Q_{k-1} + Q_k$, where $Q, Q_1, \ldots, Q_k$ are $k + 1$ rv's that are real quadratic forms in n indpt rvs which are $\mathrm{N}(\mu, \sigma^2)$ distributed. Let $\frac{Q}{\sigma^2}, \frac{Q_1}{\sigma^2}, \ldots, \frac{Q_{k-1}}{\sigma^2}$ have χ^2 distributions with df $r, r_1, \ldots, r_{k-1}$, resp. Let Q_k be nonnegative. Then:

(a) $Q_1, \ldots, Q_k$ are independent, and hence

(b) $\frac{Q_k}{\sigma^2}$ has a χ^2 dist. with $r - (r_1 + \cdots + r_{k-1}) = r_k$ df

Notation

$$\bar{X}_{\cdot\cdot} = \frac{X_{11} + \cdots + X_{1b} + \cdots + X_{a1} + \cdots + X_{ab}}{ab} = \frac{\sum_{i=1}^{a} \sum_{j=1}^{b} X_{ij}}{ab}$$

$$\bar{X}_{i\cdot} = \frac{X_{i1} + \cdots + X_{ib}}{b} = \frac{\sum_{j=1}^{b} X_{ij}}{b}, \qquad i = 1, \dots, a$$

$$\bar{X}_{\cdot j} = \frac{X_{1j} + \cdots + X_{aj}}{a} = \frac{\sum_{i=1}^{a} X_{ij}}{a}, \qquad j = 1, \dots, b$$

Quadratic Form Notation (cont.)

Total SS:

$$Q = (ab-1)S^2 = \sum_{j=1}^{b}\sum_{i=1}^{a}\left(X_{ij} - \bar{X}_{..}\right)^2$$

Within row SS:

$$Q_1 = \sum_{j=1}^{b}\sum_{i=1}^{a}\left(X_{ij} - \bar{X}_{i.}\right)^2$$

Among/across rows SS:

$$Q_2 = \sum_{j=1}^{b}\sum_{i=1}^{a}(\bar{X}_{i.} - \bar{X}_{..})^2$$

Quadratic Form Notation (cont.)

Within column SS:

$$Q_3 = \sum_{j=1}^{b} \sum_{i=1}^{a} \left(X_{ij} - \bar{X}_{.j}\right)^2$$

Among/across columns SS:

$$Q_4 = \sum_{j=1}^{b} \sum_{i=1}^{a} \left(\bar{X}_{.j} - \bar{X}_{..}\right)^2$$

Another quadratic term/SS:

$$Q_5 = \sum_{j=1}^{b} \sum_{i=1}^{a} \left(X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..}\right)^2$$

Example

Show that $(ab-1)S^2$ can be represented in the form $Q = Q_1 + Q_2$ where $\frac{Q}{\sigma^2}$ and $\frac{Q_1}{\sigma^2}$ are χ^2 distributions with $ab-1$ and $a(b-1)$ df, resp. What can you say about the distribution of Q_2?

Example

Show that $(ab-1)S^2$ can be represented in the form $Q = Q_3 + Q_4$ where $\frac{Q}{\sigma^2}$ and $\frac{Q_3}{\sigma^2}$ are χ^2 distributions with $ab-1$ and $b(a-1)$ df, resp. What can you say about the distribution of Q_4?

Example

Show that $(ab-1)S^2$ can be represented in the form $Q = Q_2 + Q_4 + Q_5$ where $\frac{Q}{\sigma^2}, \frac{Q_2}{\sigma^2}$ and $\frac{Q_4}{\sigma^2}$ are χ^2 distributions with $ab-1$, $a-1$ and $b-1$ df, resp. What can you say about the distribution of Q_5?

One-way ANOVA

- Consider b indpt normal rv's with $\mu_1, \mu_2, \dots, \mu_b$ unknown, and common unknown σ^2.
- For each j, $X_{1j}, X_{2j}, \dots, X_{aj} \sim \mathrm{N}(\mu_j, \sigma^2)$ iid
- Consider the model:

$$X_{ij} = \mu_j + e_{ij} = \mu + \beta_j + e_{ij}$$

$$i = 1, \dots, a; \;\; j = 1, \dots, b$$

where $e_{ij} \sim \mathrm{N}(0, \sigma^2)$

- Hypothesis test:

$H_0\!: \mu_1 = \mu_2 = \cdots = \mu_b = \mu$ vs. H_1: otherwise

(alternatively, $H_0\!: \beta_j = 0 \;\; \forall j$ vs. H_1: otherwise)

The One-way ANOVA Construct

- Use the likelihood ratio test where

  $\Omega = \{(\mu_1, \ldots, \mu_b, \sigma^2) : -\infty < \mu_j < \infty,\ 0 < \sigma^2 < \infty\}$, and

  $\omega = \{(\mu_1, \ldots, \mu_b, \sigma^2) : -\infty < \mu_1 = \cdots = \mu_b = \mu < \infty,\ 0 < \sigma^2 < \infty\}$

The One-way ANOVA Construct (cont.)

$$\Lambda = \frac{L(\hat{\omega})}{L(\hat{\Omega})} = \left[\frac{\sum_{j=1}^{b}\sum_{i=1}^{a}\left(X_{ij}-\bar{X}_{.j}\right)^2}{\sum_{j=1}^{b}\sum_{i=1}^{a}\left(X_{ij}-\bar{X}_{..}\right)^2}\right]^{ab/2}$$

$$= \left(\frac{Q_3}{Q}\right)^{ab/2} = \left(\frac{Q_3}{Q_3+Q_4}\right)^{ab/2} = \left(\frac{1}{1+\frac{Q_4}{Q_3}}\right)^{ab/2}$$

where $F = \frac{Q_4/(b-1)}{Q_3/[b(a-1)]} = \frac{Q_4/[\sigma^2(b-1)]}{Q_3/[\sigma^2 b(a-1)]} \sim F_{b-1,b(a-1)}$

One-Way ANOVA Table

Source of Variation	df	Sum of Squares (SS)	Mean Square (MS)	F-ratio
Between treatments (ie columns)	b-1	$SS_{Treat} = Q_4$	$MS_{Treat} = SS_{Treat}/(b-1)$	$F = MS_{Treat}/MSE$
Error (within treatments)	ab-b = b(a-1)	$SSE = Q_3$	$MSE = SSE/[b(a-1)]$	
Total	ab-1	$SST = Q_4+Q_3 = Q$		

Note: One-way ANOVA

- This test allows for different sample sizes for each of the *b* normal distributions, i.e. we can generalize to consider the model

$$X_{ij} = \mu_j + e_{ij} \quad i = 1, \dots, a_j; \; j = 1, \dots, b$$

where $e_{ij} \sim \mathrm{N}(0, \sigma^2)$

Example

The driver of a diesel-powered automobile decided to test the quality of three types of diesel fuel sold in the area based on mpg. Test the null hypothesis that the three means are equal using the following data. Make the usual assumptions and take $\alpha = 0.05$.

Brand A	38.7	39.2	40.1	38.9	
Brand B	41.9	42.3	41.3		
Brand C	40.8	41.2	39.5	38.9	40.3

Example, cont. (SAS code)

```
data diesel;
input mpg fuel $;
datalines;
38.7    a
39.2    a
      :
      :
      :
40.3    c
;
proc print; run;

proc anova data=diesel;
  class fuel;
  model mpg = fuel;
;
run;
```

SAS Output

Obs	mpg	fuel
1	38.7	a
2	39.2	a
3	40.1	a
4	38.9	a
5	41.9	b
6	42.3	b
7	41.3	b
8	40.8	c
9	41.2	c
10	39.5	c
11	38.9	c
12	40.3	c

The ANOVA Procedure
Class Level Information

Class	Levels	Values
fuel	3	a b c

Number of Observations Read	12
Number of Observations Used	12

SAS Output (cont.)

The ANOVA Procedure

Dependent Variable: mpg

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	11.78300000	5.89150000	10.22	0.0048
Error	9	5.18616667	0.57624074		
Corrected Total	11	16.96916667			

R-Square	Coeff Var	Root MSE	mpg Mean
0.694377	1.885585	0.759105	40.25833

Source	DF	Anova SS	Mean Square	F Value	Pr > F
fuel	2	11.78300000	5.89150000	10.22	0.0048

Example, cont. (Solution in R)

```
> diesel <- read.table("C:/diesel.txt",header=TRUE)
> summary(aov(mpg ~ factor(fuel), data=diesel))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
factor(fuel)	2	11.783	5.891	10.22	0.00482 **
Residuals	9	5.186	0.576		

Exercise: Verify the ANOVA table by hand.

Two-way ANOVA

- Now consider two factors A and B with levels a and b, respectively
- X_{ij} = response for Factor A at level i, Factor B at level j; $i = 1, \ldots, a$ and $j = 1, \ldots, b$
- Total sample size, $n = ab$
- $X_{ij} \sim \mathrm{N}(\mu_{ij}, \sigma^2)$ indpt

Two-way ANOVA (cont.)

Consider the two-way, main-effects model

$$\mu_{ij} = \bar{\mu}_{..} + (\bar{\mu}_{i.} - \bar{\mu}_{..}) + \left(\bar{\mu}_{.j} - \bar{\mu}_{..}\right)$$

$$= \mu + \alpha_i + \beta_j, \qquad i = 1, \ldots, a; \quad j = 1, \ldots, b$$

where $\sum_{i=1}^{a} \alpha_i = 0$ and $\sum_{j=1}^{b} \beta_j = 0$.

Consider the hypotheses:

$H_{0A}: \alpha_1 = \cdots = \alpha_a = 0$ vs. $H_{1A}: \alpha_i \neq 0$, for some i,

and

$H_{0B}: \beta_1 = \cdots = \beta_b = 0$ vs. $H_{1B}: \beta_j \neq 0$, for some j

Two-way ANOVA (cont.)

- To consider H_{0B} vs H_{1B}, the LRT uses the quadratic forms

$$(ab-1)S^2 = \sum_{j=1}^{b}\sum_{i=1}^{a}(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2 + \sum_{j=1}^{b}\sum_{i=1}^{a}\left(\bar{X}_{\cdot j} - \bar{X}_{\cdot\cdot}\right)^2$$
$$+ \sum_{j=1}^{b}\sum_{i=1}^{a}\left(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_{\cdot\cdot}\right)^2,$$

i.e. $Q = Q_2 + Q_4 + Q_5$.

- Thus, Λ is monotone wrt

$$F = \frac{Q_4/(b-1)}{Q_5/[(a-1)(b-1)]} \sim F_{(b-1),[(a-1)(b-1)]}$$

- Decision rule: reject H_{0B} if $F \geq c$, where $\alpha = P_{H_{0B}}(F \geq c)$.

Two-way ANOVA (cont.)

- To consider H_{0A} vs H_{1A}, the LRT uses the quadratic forms

$$(ab-1)S^2 = \sum_{j=1}^{b}\sum_{i=1}^{a}(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2 + \sum_{j=1}^{b}\sum_{i=1}^{a}\left(\bar{X}_{\cdot j} - \bar{X}_{\cdot\cdot}\right)^2$$
$$+ \sum_{j=1}^{b}\sum_{i=1}^{a}\left(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_{\cdot\cdot}\right)^2,$$

i.e. $Q = Q_2 + Q_4 + Q_5$.

- Thus, Λ is monotone wrt

$$F = \frac{Q_2/(a-1)}{Q_5/[(a-1)(b-1)]} \sim F_{(a-1),[(a-1)(b-1)]}$$

- Decision rule: reject H_{0A} if $F \geq c$, where

$\alpha = P_{H_{0A}}(F \geq c)$.

Two-Way ANOVA Table (w/o interaction)

Source of Variation	df	SS	MS	F-ratio
Between Columns	b-1	SS_{Col}	$MS_{Col} = SS_{Col}/(b-1)$	$F = MS_{Col}/MSE$
Between Rows	a-1	SS_{Row}	$MS_{Row} = SS_{Row}/(a-1)$	$F = MS_{Row}/MSE$
Error	(a-1)(b-1)	SSE	$MSE = SSE/[(a-1)(b-1)]$	
Total	ab-1	SST = Q		

Example

Data selected from Graybiel et al. (1975, *Aviation Space Environ. Med*,46: 1107-1118, cited by Brown in *Statistics: A Biomedical Introduction*) concern the decrease in motion sickness induced by rotation following three treatments: Scopolamine, Dimenhydrinate, and Amphetimine. The data in **Br10-Ta12.txt** are measurements (units not cited) on 10 patients, each of whom was given each of the three drugs. Are the treatments different?

Example (solution in R)

```
> pdata <- read.table("C:/Br10-Ta12.txt")
> colnames(pdata) <- c("outcome","medicine","patient")
> summary(aov(outcome ~ factor(medicine) + factor(patient),data=pdata))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
factor(medicine)	2	10.95	5.473	0.399	0.677
factor(patient)	9	108.33	12.036	0.878	0.561
Residuals	18	246.68	13.704		

Example (SAS Code)

```
data medicine;
input outcome drug $ patient;
cards;
 9.8      scopolomine     1
 4.0      scopolomine     2
 1.6      scopolomine     3
                 .
                 .
                 .
 1.0      amphetamine     9
 2.0      amphetamine     10
;
proc print; run;

proc anova data=medicine;
  class drug patient;
  model outcome = drug patient;
;

run;
```

Example (SAS Output)

The ANOVA Procedure

Class Level Information

Class	Levels	Values
drug	3	amphetam dimenhyd scopolom
Patient	10	1 2 3 4 5 6 7 8 9 10

Number of Observations Read	30
Number of Observations Used	30

Example (SAS Output cont.)

The ANOVA Procedure
Dependent Variable: outcome

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	119.2713333	10.8428485	0.79	0.6468
Error	18	246.6806667	13.7044815		
Corrected Total	29	365.9520000			

R-Square	Coeff Var	Root MSE	outcome Mean
0.325921	151.7195	3.701956	2.440000

Source	DF	Anova SS	Mean Square	F Value	Pr > F
drug	2	10.9460000	5.4730000	0.40	0.6765
patient	9	108.3253333	12.0361481	0.88	0.5612

Two-way ANOVA w/ Interaction

- $X_{ijk} \sim \mathrm{N}(\mu_{ij}, \sigma^2)$, $i = 1, \ldots, a$; $j = 1, \ldots, b$; $k = 1, \ldots, c$ indpt.
- Consider the model

$$\begin{aligned} X_{ijk} &= \mu_{ij} + e_{ijk} \\ &= \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}, \end{aligned}$$

where $\mu = \bar{\mu}_{..}$, $\alpha_i = \bar{\mu}_{i.} - \bar{\mu}_{..}$, $\beta_j = \bar{\mu}_{.j} - \bar{\mu}_{..}$, $\gamma_{ij} = \mu_{ij} - \bar{\mu}_{i.} - \bar{\mu}_{.j} + \bar{\mu}_{..}$, and

$\sum_{i=1}^{a} \alpha_i = 0, \sum_{j=1}^{b} \beta_j = 0, \sum_{i=1}^{a} \gamma_{ij} = \sum_{j=1}^{b} \gamma_{ij} = 0.$

Two-way ANOVA w/ Interaction

Hypothesis test for interaction term:

$H_{0AB}: \gamma_{ij} = 0 \ \forall i, j$ vs $H_{1AB}: \gamma_{ij} \neq 0$ for some i, j

$$\begin{aligned}
\Rightarrow \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} \left(X_{ijk} - \bar{X}_{...}\right)^2 = {} & bc \sum_{i=1}^{a} (\bar{X}_{i..} - \bar{X}_{...})^2 \\
& + ac \sum_{j=1}^{b} \left(\bar{X}_{.j.} - \bar{X}_{...}\right)^2 \\
& + c \sum_{i=1}^{a} \sum_{j=1}^{b} \left(\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}_{...}\right)^2 \\
& + \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} \left(X_{ijk} - \bar{X}_{ij.}\right)^2
\end{aligned}$$

Two-way ANOVA w/ Interaction

- Decision rule: reject H_{0AB} if $F \geq c$, where c st. $P(F \geq c) = \alpha$, and

$$F = \frac{c\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}_{...}\right)^2 \Big/ [(a-1)(b-1)]}{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\left(X_{ijk} - \bar{X}_{ij.}\right)^2 \Big/ [ab(c-1)]}$$

- If we fail to reject H_{0AB}, we can still perform tests regarding the main effects

Two-Way ANOVA Table (w/ interaction)

Source of Variation	df	SS	MS	F-ratio
Between Columns	b-1	SS_{Col}	$MS_{Col} = SS_{Col}/(b-1)$	$F = MS_{Col}/MSE$
Between Rows	a-1	SS_{Row}	$MS_{Row} = SS_{Row}/(a-1)$	$F = MS_{Row}/MSE$
Interaction	(a-1)(b-1)	SS_{Int}	$MS_{Int} = SS_{Int}/[(a-1)(b-1)]$	$F = MS_{Int}/MSE$
Error	ab(c-1)	SSE	$MSE = SSE/[ab(c-1)]$	
Total	abc-1	SST = Q		