Spring 2013
Stat 8004: Statistical Methods II
Lecture 3

Lecturer: Jichun Xie

1 Two-Way ANOVA

1.1 Example

Example: Mental hospital admissions during full moons.

Larsen and Marx (1986) wrote:

In folklore, the full moon is often portrayed as something sinister, a kind of evil force possessing the power to control our behaviour. Over the centuries, many prominent writers and philosophers have shared this belief. Milton, in Paradise Lost, refers to

```
Demoniac frenzy, moping melancholy
And moon-struck madness.
```

And Othello, after the murder of Desdemona, laments

```
It is the very error of the moon She comes more near the earth
than she was want And makes men mad
```

On a more scholarly level, Sir William Blackstone, the renowned eighteenth centure English barrister, defined a "lunatic" as

```
one who hath ... lost the use of his reason and who hath lucid intervals,
sometimes enjoying his senses and sometimes not,
and that frequently depending upon changes of the moon
```

The data give the admission rates to the emergency room of a Virginia mental health clinic before, during and after the 12 full moons from August 1971 to July 1972.

Model:

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, \tag{1}$$

where $e_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2)$, $i = 1, \ldots, a;\ j = 1, \ldots, b$.

In this specific example, $a = 3$ stands for the phase of the moon, $b = 12$ stands for the month.

Parameter	Interpolation
μ	common effect
α_i	effect due to the ith phase of the moon
β_j	effect due to the jth month

How to write the model in the matrix form?

Now we use R to display and analyze the data.

```
Analysis of Variance Table

Response: Admission
          Df  Sum Sq Mean Sq F value    Pr(>F)
Moon       2   41.51  20.757  3.5726   0.04533 *
Month     11 455.58  41.417  7.1285 5.076e-05 ***
```

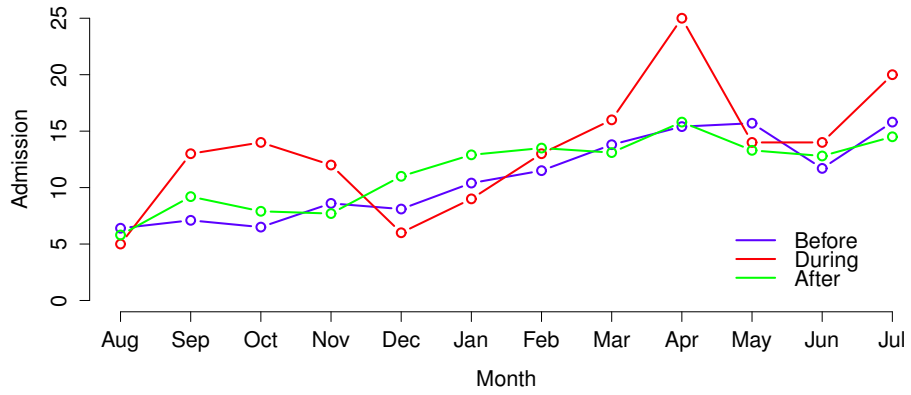

Figure 1: Mental hospital admissions vs. the month

```
Residuals 22 127.82   5.810
---
Signif. codes:  0 '***' 0.001 '**' 0.01'*' 0.05 '.'
 0.1 ' ' 1
```

The ANOVA table shows that the assciation between the mental hospital admissions and the month is highly significant. The asscoation between the admissions and the phase of moon is also significant.

1.2 Hypothesis Testing and ANOVA Table (Single Observation)

How to conduct hypothesis testing about the effect of factor A and factor B?

Let

$$SSA = b\sum_{i=1}^{a}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$$

$$SSB = a\sum_{j=1}^{b}(\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2$$

$$SSE = \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot})^2$$

1.2.1 Testing Factor A

The following three hypotheses are equivalent.

$$\begin{aligned}
&\mathrm{H}_0 : \alpha_i = \alpha_j,\ \forall 1 \le i < j \le a.\\
&\mathrm{H}_0 : \alpha_i = \alpha_{i+1},\ i = 1, \ldots, a-1.\\
&\mathrm{H}_0 : \alpha_i = 0,\ i = 1, \ldots, a.
\end{aligned}$$

Let

$$F_A = \frac{SSA/(a-1)}{SSE/\{(a-1)(b-1)\}}.$$

Under H_0, $F_A \sim F(a-1, (a-1)(b-1))$.

1.2.2 Testing Factor B

The following three hypotheses are equivalent.

$$\begin{aligned}
&\mathrm{H}_0 : \beta_i = \beta_j,\ \forall 1 \le i < j \le b.\\
&\mathrm{H}_0 : \beta_i = \beta_{i+1},\ i = 1, \ldots, b-1.\\
&\mathrm{H}_0 : \beta_i = 0,\ i = 1, \ldots, b.
\end{aligned}$$

Let

$$F_B = \frac{SSB/(b-1)}{SSE/\{(a-1)(b-1)\}}.$$

Under H_0, $F_B \sim F(b-1, (a-1)(b-1))$.

1.2.3 Testing Factor A and Factor B Jointly

The following three hypotheses are equivalent.

$$\begin{aligned}
&\mathrm{H}_0 : \alpha_{i_1} = \alpha_{i_2},\ \forall 1 \le i_1 < i_2 \le a; \quad \beta_{j_1} = \beta_{j_2},\ \forall 1 \le j_1 < j_2 \le b.\\
&\mathrm{H}_0 : \alpha_i = \alpha_{i+1},\ i = 1, \ldots, a-1; \quad \beta_j = \beta_{j+1},\ j = 1, \ldots, b-1.\\
&\mathrm{H}_0 : \alpha_i = 0,\ i = 1, \ldots, a; \quad \beta_j = 0,\ j = 1, \ldots, b.
\end{aligned}$$

Let

$$F_{A+B} = \frac{(SSA + SSB)/(a+b-2)}{SSE/\{(a-1)(b-1)\}}.$$

Under H_0, $F_{A+B} \sim F(a+b-2, (a-1)(b-1))$.

The results are summarized in Table 1.

Source	SS	d.f.	MS	F Statistic
Factor A	SSA	$a-1$	$SSA/(a-1)$	$F_A = MSA/MSE$
Factor B	SSB	$b-1$	$SSB/(b-1)$	$F_B = MSB/MSE$
Error	SSE	$(a-1)(b-1)$	$\frac{SSE}{(a-1)(b-1)}$	
Total	SST_m	$ab-1$	$SST_m/(n-1)$	

Table 1: Two-way ANOVA table

1.3 Another Example

Example: Age and Memory

Why do older people often seem not to remember things as well as younger people? Do they not pay attention? Do they just not process the material as thoroughly? One theory regarding memory is that verbal material is remembered as a function of the degree to which is was processed when it was initially presented. Eysenck (1974) randomly assigned 50 younger subjects and 50 older (between 55 and 65 years old) to one of five learning groups. The Counting group was asked to read through a list of words and count the number of letters in each word. This involved the lowest level of processing. The Rhyming group was asked to read each word and think of a word that rhymed with it. The Adjective group was asked to give an adjective that could reasonably be used to modify each word in the list. The Imagery group was instructed to form vivid images of each word, and this was assumed to require the deepest level of processing. None of these four groups was told they would later be asked to recall the items. Finally, the Intentional group was asked to memorize the words for later recall. After the subjects had gone through the list of 27 items three times they were asked to write down all the words they could remember.

Variable	Description
Age	Younger or Older
Process	The level of processing: Counting, Rhyming, Adjective, Imagery or Intentional
Words	Number of words recalled

Model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}, \tag{2}$$

where $e_{ijk} \overset{i.i.d.}{\sim} N(0, \sigma^2)$, $i = 1, \ldots, a;\ j = 1, \ldots, b; \quad k = 1, \ldots, n.$

```
Analysis of Variance Table

Response: Words
          Df  Sum Sq Mean Sq F value    Pr(>F)
Age        1  240.25  240.25  24.746 2.943e-06 ***
Process    4 1514.94  378.73  39.011 < 2.2e-16 ***
Residuals 94  912.60    9.71
---
Signif. codes:  0 '***' 0.001 '**' 0.01'*' 0.05 '.'
 0.1 ' ' 1
```

Let

$$SSA = bn \sum_{i=1}^{a} (\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots})^2$$

$$SSB = an \sum_{j=1}^{b} (\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots})^2$$

$$SSE = \sum_{k=1}^{n} \sum_{i=1}^{a} \sum_{j=1}^{b} (y_{ijk} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots})^2$$

Source	SS	d.f.	MS	F Statistic
Factor A	SSA	$a-1$	$\frac{SSA}{a-1}$	$F_A = MSA/MSE$
Factor B	SSB	$b-1$	$\frac{SSB}{b-1}$	$F_B = MSB/MSE$
Factor A+B	$SSA+SSB$	$a+b-2$	$\frac{SSA+SSB}{a+b-2}$	$F_{A+B} = MS(A+B)/MSE$
Error	SSE	$N-a-b+1$	$\frac{SSE}{N-a-b+1}$	
Total	SST_m	$N-1$	$\frac{SST_m}{N-1}$	

Table 2: Two-way ANOVA table without interaction. Here $N = abn$ is the total sample size.

Note that in this example, we have same numbers of observations in each category. This is called *balanced design*. In some situations, we might have varying numbers of observations in each category. Suppose in the ith level of

factor A and in the jth level of factor B, there are n_{ij} observations. Under the following conditions, the analysis of unbalanced data would be exactly like the balanced data:

$$n_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}}, \quad \forall i, j,$$

where $n_{i\cdot} = \sum_{j=1}^{b} n_{ij}$ and $n_{\cdot j} = \sum_{i=1}^{a} n_{ij}$. The condition is called *proportional frequencies condition.*

For example, the following design is proportional.

Factor	B1	B2	B3	
A1	2	3	4	9
A2	6	9	12	27
A3	4	6	8	18
	12	18	24	54

Why do we care about proportional frequency condition or the balanced condition? Because under such conditions, the contrasts to test factor A and factor B are orthogonal. Suppose $\mathbf{q}_i^{\mathrm{T}}\boldsymbol{\beta}$ and $\mathbf{q}_j^{\mathrm{T}}\boldsymbol{\beta}$ are too estimable functions. Then

$$\mathrm{Cov}(\mathbf{q}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}, \mathbf{q}_j^{\mathrm{T}}\hat{\boldsymbol{\beta}}) = \sigma^2 \mathbf{q}_i^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{q}_j.$$

Definition 1. $\mathbf{q}_i^{\mathrm{T}}\boldsymbol{\beta}$ *and* $\mathbf{q}_j^{\mathrm{T}}\boldsymbol{\beta}$ *are called orthogonal contrasts if and only if* $\mathbf{q}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}} \perp\!\!\!\perp \mathbf{q}_j^{\mathrm{T}}\hat{\boldsymbol{\beta}}$*, i.e.*

$$\mathrm{Cov}(\mathbf{q}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}, \mathbf{q}_j^{\mathrm{T}}\hat{\boldsymbol{\beta}}) = \sigma^2 \mathbf{q}_i^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{q}_j = 0.$$

For orthogonal contrasts, the sum squares for the contrasts can be added up in the ANOVA table.

2 Two-Way ANOVA with Interactions

2.1 Revisit the Memory Example

Now let's consider changing a model.

In Figure 2, these two lines are not parallel. What does it mean? The memory difference between the younger group and the older group are different across

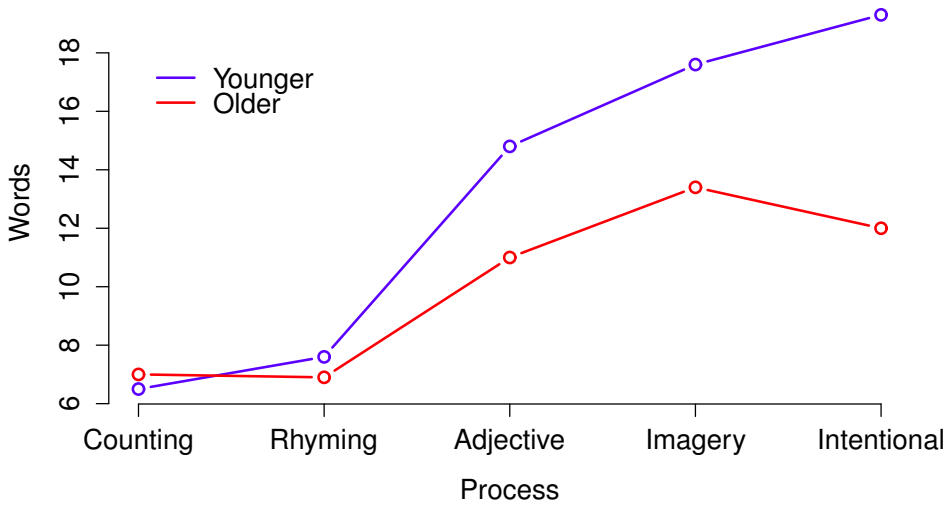

Figure 2: Memory versus Age

processes. For example, the difference between two gropus in the counting process (low level process) is much smaller than the difference in the intentional process (high level process).

Consider the following model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \quad i = 1, \ldots, a, \ j = 1, \ldots, b, \ k = 1, \ldots, n, \quad (3)$$

where $\epsilon_{ijk} \sim N(0, \sigma^2)$ *i.i.d.*.

Compared with Note that this model has more parameters involved. γ_{ij} are called interaction terms. In this example, $a = 2$, $b = 5$, $n = 10$.

parameter	Interpretation
μ	the common effect
α_i	the main effect of the ith group (younger or older)
β_j	the main effect of the jth process
γ_{ij}	the interaction term

The estimable functions under this model will be linear combinations of

$$\mathbb{E}(Y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij} = \mu_{ij}.$$

What should Figure 2 look like if there are no interaction terms?

When we can observe "main effects" but no iteractions:

- Age has an impact on memory.
- Different processes make an impact on memory.
- The memory differences between the older group and the young group are the same across different processes.

2.2 Inference and Analysis

Parameter estimation:

$$\begin{aligned}\hat{\mu} &= \bar{y}_{...} \\ \hat{\alpha}_i &= \bar{y}_{i..} - \bar{y}_{...} \\ \hat{\beta}_j &= \bar{y}_{.j.} - \bar{y}_{...} \\ \hat{\gamma}_{ij} &= \bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}\end{aligned}$$

```
fit2 <- lm(Words ~ Age*Process, data=fm)
summary(fit2)
anova(fit2)
```

will yield

```
Analysis of Variance Table

Response: Words
             Df  Sum Sq Mean Sq F value    Pr(>F)
Age           1  240.25  240.25 29.9356 3.981e-07 ***
Process       4 1514.94  378.73 47.1911 < 2.2e-16 ***
Age:Process   4  190.30   47.58  5.9279 0.0002793 ***
Residuals    90  722.30    8.03
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
  0.1 ' ' 1
```

And

```
aov(formula = Words ~ Age*Process, data=fm)
```

will yield similar output

```
Call:
   aov(formula = Words ~ Age * Process, data = fm)

Terms:
                     Age Process Age:Process
                      Residuals
Sum of Squares    240.25 1514.94      190.30
  722.30
Deg. of Freedom        1       4           4
  90

Residual standard error: 2.832941
Estimated effects are balanced
```

Remark: In R, to incorporate both main effects and interaction terms in the model, we just need to write the interaction terms in the formula. As long as the interaction term is there, the main effects are automatically included.

How to write out the hypotheses to test the interaction terms?

$$\mathrm{H}_0: \ \gamma_{ij} = 0, \quad \forall i = 1, \ldots, a; \ j = 1, \ldots, b.$$

Equivalently, it is testing

$$\begin{aligned} \mathrm{H}_0 : \mu_{a,1} - \mu_{a-1,1} &= \ldots = \mu_{a,b} - \mu_{a-1,b}, \\ \mu_{a-1,1} - \mu_{a-2,1} &= \ldots = \mu_{a-1,b} - \mu_{a-2,b} \\ \mu_{2,1} - \mu_{1,1} &= \ldots = \mu_{2,b} - \mu_{1,b} \end{aligned}$$

Let

$$SSA = bn \sum_{i=1}^{a} (\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots})^2$$

$$SSB = an \sum_{j=1}^{b} (\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots})^2$$

$$SS(A*B) = n \sum_{i=1}^{a} \sum_{j=1}^{b} (\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots})^2$$

$$SSE = \sum_{k=1}^{n} \sum_{i=1}^{a} \sum_{j=1}^{b} (y_{ijk} - \bar{y}_{ij\cdot})^2$$

Source	SS	d.f.	MS	F Statistic
Factor A	SSA	$a-1$	$\frac{SSA}{a-1}$	$F_A = MSA/MSE$
Factor B	SSB	$b-1$	$\frac{SSB}{b-1}$	$F_B = MSB/MSE$
Factor A*B	$SS(A*B)$	$(a-1)(b-1)$	$\frac{SS(A*B)}{(a-1)(b-1)}$	$F_{A*B} = MS(A*B)/MSE$
Error	SSE	$ab(n-1)$	$\frac{SSE}{ab(n-1)}$	
Total	SST_m	$abn-1$	$\frac{SST_m}{abn-1}$	

Table 3: Two-Way ANOVA table with Interactions