Spring 2013
Stat 8004: Statistical Methods II
Lecture 2

Lecturer: Jichun Xie

1 One-Way ANOVA

1.1 An Example

An individual's critical flicker frequency is the highest frequency at which the flicker in a flickering light source can be detected. At frequencies above the critical frequency, the light source appears to be continuous even though it is actually flickering. This investigation recorded critical flicker frequency and iris colour of the eye for 19 subjects. Please use a linear regression to study whether there is any association between eye color and flicker frequency.

1.2 Model and Formulation

1.2.1 Model 1

$$y_{ij} = \beta_i + e_{ij}, \quad \text{where } e_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2), \quad i = 1, \ldots, 3; \ j = 1, \ldots, n_i. \tag{1}$$

Hypothesis: $\mathrm{H}_0 : \ \beta_1 = \beta_2 = \beta_3$.

How to formulate in the form $\mathbf{K}^{\mathrm{T}}\boldsymbol{\beta} = \mathbf{m}$?

Under Model (1), the estimators of the parameters are

$$\hat{\beta}_i = \bar{y_{i\cdot}} \sim N(\mu, \sigma^2/n),$$
$$\hat{\sigma}^2 = s^2 = \frac{RSS}{p(n-1)},$$

where

$$RSS = \sum_{i=1}^{p} \sum r = 1^n (y_{ir} - \bar{y}_{i\cdot})^2 \sim \sigma^2 \chi^2(p(n-1), 0).$$

Further $\hat{\beta}_i$, $i = 1, \ldots, 3$ and RSS are mutually independent.

1.2.2 Model 2

$$y_{ij} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e_{ij}, \quad \text{where } e_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2), \quad i = 1, \ldots, 3;\ j = 1, \ldots, n_i. \tag{2}$$

Here the baseline is the "brown", X_1 is the indicator of "Green", and X_2 is the indication of "Blue". X_1 and X_2 are called *dummy variables*. They are helpful when there are multiple levels of the factor of interest.

Hypothesis: $\mathrm{H}_0 : \ \beta_1 = \beta_2 = 0$.

How to formulate in the form $\mathbf{K}^{\mathrm{T}}\boldsymbol{\beta} = \mathbf{m}$?

1.2.3 Model 3

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad \text{where } e_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2), \quad \sum_{i=1}^{3} \alpha_i = 0, \quad i = 1, \ldots, 3,\ j = 1, \ \ldots, n_i. \tag{3}$$

If written in matrix form $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, what would each term be?

For non-full rank design matrix $\mathbf{X}$, hypothesis testing problem is sometimes tricky. Some hypotheses *cannot* be tested.

Examples: Can the following hypotheses be tested?

$$\begin{aligned}
&\mathrm{H}_0 : \mu = 0 \\
&\mathrm{H}_0 : \alpha_1 = \alpha_2 = \alpha_3 \\
&\mathrm{H}_0 : \alpha_1 + \alpha_2 = 3 \\
&\mathrm{H}_0 : \alpha_1 - \alpha_2 = 3 \\
&\mathrm{H}_0 : 2\mu + \alpha_2 + \alpha_3 = 0
\end{aligned}$$

The hypothesis $\mathrm{H}_0 : \alpha_1 = \alpha_2 = \alpha_3$ can be tested. In the filcker example, this hypothesis helps to test whether there is any assorication between eye color and flikcer frequency. Why?

The above model is called a one-way ANOVA model. The general form is

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad \text{where } e_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2), \quad \sum_{i=1}^{a} \alpha_i = 0, \quad i = 1, \ldots, a, \; j = 1, \ldots, n_i. \tag{4}$$

In order to test weather the outcome is associated with different factor levels (a levels), we build up the hypothesis:

$$\mathrm{H}_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_a.$$

1.3 ANOVA Table

You can use either model to do the testing. They are essentially equivalent. Model 3 is more commonly used since it is easy to interpretate. Table 1 summarizes how to conduct the test.

Source	SS	d.f.	MS	F Statistic
Between	$SSB = \sum_{i=1}^{a} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$	$a-1$	$SSB/(a-1)$	$F = \frac{SSB/(a-1)}{SSW/(n-a)}$
Within	$SSW = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2$	$n-a$	$SSW/(n-a)$	$\overset{\text{under } \mathrm{H}_0}{\sim} F(a-1, n-a)$
Total	$SST_m = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{\cdot\cdot})^2$	$n-1$	$SST_m/(n-1)$	

Table 1: One-way ANOVA table

What does each term mean?

2 Two-Way ANOVA

2.1 Example

Example: Mental hospital admissions during full moons.

Figure 1: Vampire and wolfman 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:

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, \tag{5}$$

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 ***
Residuals 22 127.82   5.810
```

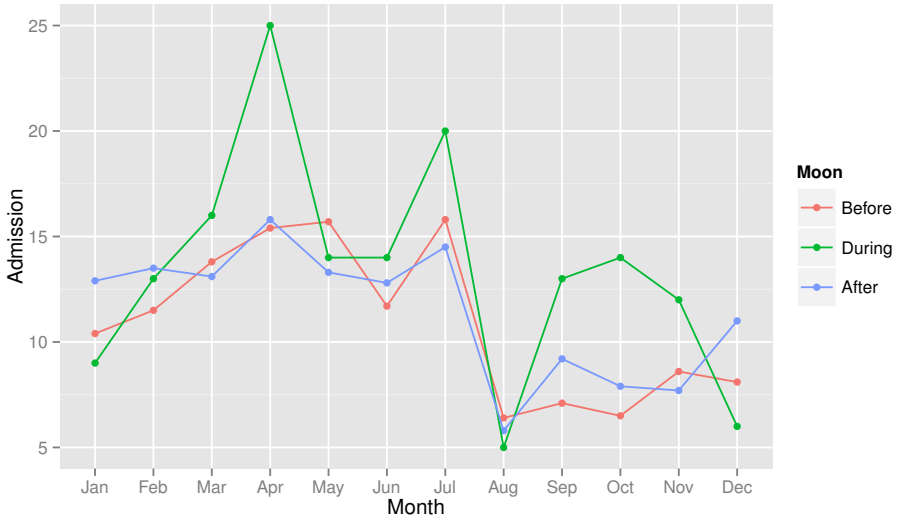

Figure 2: Mental hospital admissions vs. the month

```
---
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.

2.2 Hypothesis Testing and ANOVA Table (Single Observation)

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

Let

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

2.2.1 Testing Factor A

The following three hypotheses are equivalent.

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

Let

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

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

2.2.2 Testing Factor B

The following three hypotheses are equivalent.

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

Let

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

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

2.2.3 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

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

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

The results are summarized in Table 3.

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

Table 2: Two-way ANOVA table

2.3 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:

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ijk}, \tag{6}$$

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

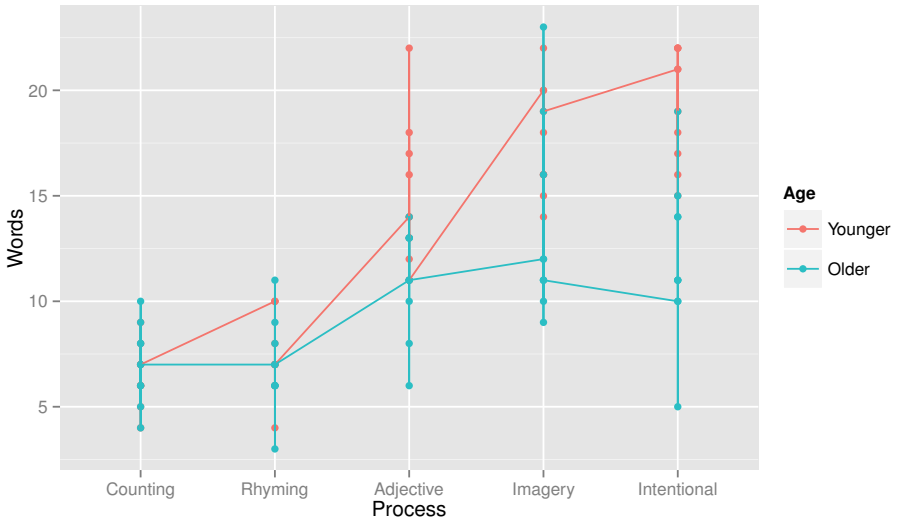

Figure 3: Memory vs. age

```
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

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

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

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

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	$N-1$	$SST_m/(N-1)$	

Table 3: Two-way ANOVA table