Statistical Methods

Nels Grevstad

Metropolitan State University of Denver

ngrevsta@msudenver.edu

October 16, 2019

Topics

Objectives

Objectives:

- State the random effects version of the one-factor ANOVA model.
- State the treatment effects version of the two-factor ANOVA model when $K = 1$.
- Carry out two-factor ANOVA F tests for the effects of Factors A and B when $K = 1$.

A Random Effects Model

- Up to now, the **levels** of the **factor** were assumed to have been **hand-picked**.

A Random Effects Model

- Up to now, the **levels** of the **factor** were assumed to have been **hand-picked**.

  For example, the five labs were singled out because they specifically were of interest to the inspectors.

A Random Effects Model

- Up to now, the **levels** of the **factor** were assumed to have been **hand-picked**.

  For example, the five labs were singled out because they specifically were of interest to the inspectors.

- In some studies, the **levels** of the **factor** are selected **randomly** from a *population of levels*.

A Random Effects Model

- Up to now, the **levels** of the **factor** were assumed to have been **hand-picked**.

  For example, the five labs were singled out because they specifically were of interest to the inspectors.

- In some studies, the **levels** of the **factor** are selected **randomly** from a *population of levels*.

  For example, five labs could be **randomly** selected from a **population** of **labs**.

- In the first case, the so-called ***fixed effects*** model was

$$X_{ij} \;=\; \mu + \alpha_i + \epsilon_{ij}\,,$$

where the **effects** $\alpha_1, \alpha_2, \ldots, \alpha_I$ are unknown **constants** that sum to zero:

$$\sum_{i=1}^{I} \alpha_i \;=\; 0\,.$$

- In the second case, the appropriate model is the so-called ***random effects*** model:

$$X_{ij} \;=\; \mu + A_i + \epsilon_{ij} \,,$$

where the **effects** $A_1, A_2, \ldots, A_I$ are **random variables** with expected value zero:

$$E(A_i) \;=\; 0 \qquad\qquad \text{for } i = 1, 2 \ldots, I \,.$$

One-factor ANOVA Random Effects Model:

$$X_{ij} \;=\; \mu + A_i + \epsilon_{ij}\,, \tag{1}$$

where

μ is a constant called the ***true grand mean***

A_i is the ***random treatment effect*** for ith treatment, and the A_i's are iid $N(0, \sigma_A)$.

ϵ_{ij} are iid $N(0, \sigma)$ ***random errors***, independent of the A_i's.

- In terms of the **random effects ANOVA model**, the hypotheses are:

$$
\begin{aligned}
H_0\colon \sigma_A &= 0 \\
H_a\colon \sigma_A &> 0
\end{aligned}
$$

The **null hypothesis** says there's **no variation** in the treatment effects (i.e. the effects are all the same).

- In terms of the **random effects ANOVA model**, the hypotheses are:

$$
\begin{aligned}
H_0\colon \sigma_A &= 0 \\
H_a\colon \sigma_A &> 0
\end{aligned}
$$

The **null hypothesis** says there's **no variation** in the treatment effects (i.e. the effects are all the same).

The **alternative** says the the effects **vary** (i.e. they're *not* all the same).

- Although the *hypotheses* in the one-factor **fixed** and **random effects models** are different, *they're tested in exactly the same way*:

- Although the *hypotheses* in the one-factor **fixed** and **random effects models** are different, *they're tested in exactly the same way*:

P-value = Tail area under the $F_{I-1,I(J-1)}$ distribution to the right of $F = \text{MSTr}/\text{MSE}$.

Two-Factor ANOVA with $K = 1$

Introduction

- We're sometimes interested in *simultaneously* testing for the effects of *two* factors on a response variable.

Example

A study was carried out ascertain the stability of vitamin C in reconstituted frozen orange juice stored in a refrigerator for up to one week.

Example

A study was carried out ascertain the stability of vitamin C in reconstituted frozen orange juice stored in a refrigerator for up to one week.

Three **brands** of orange juice were tested at three **storage times** (in days after the orange juice was blended).

Example

A study was carried out ascertain the stability of vitamin C in reconstituted frozen orange juice stored in a refrigerator for up to one week.

Three **brands** of orange juice were tested at three **storage times** (in days after the orange juice was blended).

The response variable is milligrams of ascorbic acid (vitamin C) per liter. The data are on the next slide.

Factor B: Time

Factor A: Orange Juice Brand	0 Days ($j = 1$)	3 Days ($j = 2$)	7 Days ($j = 3$)	
Richfood ($i = 1$)	54.2	42.8	47.6	$\bar{X}_{1\cdot} = 48.2$
Sealed-Sweet ($i = 2$)	56.0	44.0	42.0	$\bar{X}_{2\cdot} = 47.3$
Minute Maid ($i = 3$)	53.6	48.0	43.3	$\bar{X}_{3\cdot} = 48.3$
	$\bar{X}_{\cdot 1} = 54.6$	$\bar{X}_{\cdot 2} = 44.9$	$\bar{X}_{\cdot 3} = 44.3$	$\bar{X}_{\cdot\cdot} = 47.9$

The study was designed to find out:

The study was designed to find out:

- Does the **brand** of orange juice affect the vitamin C content (i.e. is there a ***factor A main effect***)?

The study was designed to find out:

- Does the **brand** of orange juice affect the vitamin C content (i.e. is there a ***factor A main effect***)?
- Does **storage time** affect vitamin C content (i.e. is there a ***factor B main effect***)?

- We'll refer to each **combination** of **levels** of the **two factors** as a ***treatment group***.

- We'll refer to each **combination** of **levels** of the **two factors** as a ***treatment group***.

  Example: In study of vitamin C in orange juice, there were nine treatment groups (each consisting of a single observation).

- We'll refer to each **combination** of **levels** of the **two factors** as a ***treatment group***.

  Example: In study of vitamin C in orange juice, there were nine treatment groups (each consisting of a single observation).

- We'll start by focusing on the case in which there's only **one observation per group**.

- **Notation**:

- **Notation**:

$\boldsymbol{I}$ = The number of levels of Factor A.
$\boldsymbol{J}$ = The number of levels of Factor B.
$\boldsymbol{K}$ = The common sample size in each of the IJ groups (combinations of levels of Factors A and B).

For now, we'll assume $K = 1$.

$\boldsymbol{X_{ij}}$ = The (single) observation at the ith level of Factor A and jth level of Factor B (i, jth group).

- **Notation**:

$\boldsymbol{I}$ = The number of levels of Factor A.
$\boldsymbol{J}$ = The number of levels of Factor B.
$\boldsymbol{K}$ = The common sample size in each of the IJ groups (combinations of levels of Factors A and B).

For now, we'll assume $K = 1$.

$\boldsymbol{X_{ij}}$ = The (single) observation at the ith level of Factor A and jth level of Factor B (i, jth group).

(In practice, the sample sizes *don't* all have to be the same).

- The data can be laid out in a table as below:

- The data can be laid out in a table as below:

Factor B

		Level $j = 1$	Level $j = 2$	$\cdots$	Level $j = J$	
Factor A	Level $i = 1$	X_{11}	X_{12}	$\cdots$	X_{1J}	$\bar{X}_{1\cdot}$
	Level $i = 2$	X_{21}	X_{22}	$\cdots$	X_{2J}	$\bar{X}_{2\cdot}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	Level $i = I$	X_{I1}	X_{I2}	$\cdots$	X_{IJ}	$\bar{X}_{I\cdot}$
		$\bar{X}_{\cdot 1}$	$\bar{X}_{\cdot 2}$	$\cdots$	$\bar{X}_{\cdot J}$	$\bar{X}_{\cdot\cdot}$

- (cont'd)

$$
\begin{aligned}
\bar{X}_{i\cdot} &= \text{The } i\text{th } \underline{\textbf{\textit{Factor A level mean}}} \text{ of all observations at level } i \text{ of Factor A.} \\
\bar{X}_{\cdot j} &= \text{The } j\text{th } \underline{\textbf{\textit{Factor B level mean}}} \text{ of all observations at level } j \text{ of Factor B.} \\
\bar{X}_{\cdot\cdot} &= \text{The } \underline{\textbf{\textit{grand mean}}} \text{ of } \textit{all } IJ \text{ observations.}
\end{aligned}
$$

- Note: $\bar{X}_{..}$ can be obtained as:

- Note: $\bar{X}_{..}$ can be obtained as:
    - The average of the I Factor A level means.
    - The average of the J Factor B level means.

The Two-Factor ANOVA Model (Two Versions)

- On the next slide is one **statistical model** for describing data from a two-factor study.

Two-factor ANOVA Model (Group Means Version):

$$X_{ij} = \mu_{ij} + \epsilon_{ij} \, , \tag{2}$$

where

μ_{ij} is the ***true mean*** response to level i of Factor A and level j of Factor B.

ϵ_{ij} are iid $N(0, \sigma)$ ***random errors***.

Two-Factor ANOVA Model (Additive Effects Version)

- When there's only one observation per cell, there aren't enough data to estimate all the parameters in the model (2):

Two-Factor ANOVA Model (Additive Effects Version)

- When there's only one observation per cell, there aren't enough data to estimate all the parameters in the model (2):

  we could estimate each μ_{ij} by X_{ij}, but that would "use up" all the data and there'd be none left over to estimate σ.

Two-Factor ANOVA Model (Additive Effects Version)

- When there's only one observation per cell, there aren't enough data to estimate all the parameters in the model (2):

  we could estimate each μ_{ij} by X_{ij}, but that would "use up" all the data and there'd be none left over to estimate σ.

- Furthermore, it's preferable to use a model like the one on the next slide that has parameters representing the **effects** of the two factors.

Two-factor ANOVA Model (Additive Effects Version): Assume the existence of I parameters $\alpha_1, \alpha_2, \ldots, \alpha_I$ and J parameters $\beta_1, \beta_2, \ldots, \beta_J$ such that

$$X_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \tag{3}$$

so that

$$\mu_{ij} = \mu + \alpha_i + \beta_j \,,$$

where

$$\sum_{i=1}^{I} \alpha_i = 0 \quad \text{and} \quad \sum_{j=1}^{J} \beta_j = 0 \,,$$

In this model,

μ is a constant called the ***true grand mean***.
α_i is the ***effect*** of the ith level of **Factor** A.
β_j is the ***effect*** of the jth level of **Factor** B.
ϵ_{ij} are iid $N(0, \sigma)$ ***random errors***.

- Some comments:

- Some comments:
    - The two models are equivalent **if it's reasonable to assume that the μ_{ij}'s satisfy the *additivity structure***:

$$\mu_{ij} = \mu + \alpha_i + \beta_j \,. \tag{4}$$

- (cont'd)
  - The parameters μ, $\alpha_1, \alpha_2, \ldots, \alpha_I$, and $\beta_1, \beta_2, \ldots, \beta_J$ **woudn't** be uniquely defined without imposing the constraints:

$$\sum_i \alpha_i = 0 \qquad \text{and} \qquad \sum_j \beta_j = 0.$$

    For example, adding a constant c to μ and subtracting c from each α_i (or each β_j) would lead to the *same* value of μ_{ij}.

- (cont'd)
  - With these constraints, it can be shown (by summing both sides of (4) over i and j) that
  $$\mu = \frac{\sum_i \sum_j \mu_{ij}}{IJ},$$
  i.e. the **true grand mean** the **average** of the IJ **group means**), ...

- (cont'd)
  - With these constraints, it can be shown (by summing both sides of (4) over i and j) that

    $$\mu = \frac{\sum_i \sum_j \mu_{ij}}{IJ},$$

    i.e. the **true grand mean** the **average** of the IJ **group means**), ...

    and (by summing both sides of (4) first over i and then over j) that the **effects** $\boldsymbol{\alpha_1, \alpha_2, \ldots, \alpha_I}$ and $\boldsymbol{\beta_1, \beta_2, \ldots, \beta_J}$ are

    $$\alpha_i = \mu_{i\cdot} - \mu \qquad \text{and} \qquad \beta_j = \mu_{\cdot j} - \mu,$$

    where $\boldsymbol{\mu_{\cdot j}}$ and $\boldsymbol{\mu_{i\cdot}}$ are the **true Factor A** and **B level means**, defined as the average of the μ_{ij}'s in the ith row or jth column:

    $$\mu_{i\cdot} = \frac{\sum_j \mu_{ij}}{J} \qquad \text{and} \qquad \mu_{\cdot j} = \frac{\sum_i \mu_{ij}}{I}.$$

- We'll want to test **two sets** of **hypotheses**:

- We'll want to test **two sets** of **hypotheses**:

Null and Alternative Hypotheses:

- **Factor A main effect**:

$$\begin{aligned} H_{0A} &: \quad \alpha_1 = \alpha_2 = \ldots = \alpha_I = 0 \\ H_{aA} &: \quad \text{not all } \alpha_i\text{'s equal zero} \end{aligned} \tag{5}$$

- **Factor B main effect**:

$$\begin{aligned} H_{0B} &: \quad \beta_1 = \beta_2 = \ldots = \beta_J = 0 \\ H_{aB} &: \quad \text{not all } \beta_j\text{'s equal zero} \end{aligned} \tag{6}$$

- H_{0A} says Factor A doesn't have any effect, and H_{aA} says it does.

- H_{0A} says Factor A doesn't have any effect, and H_{aA} says it does.

  H_{0B} says Factor B doesn't have any effect, and H_{aB} says is does.

Sums of Squares and the ANOVA Partition

- We can ***partition*** the **total variation** in the data into three parts:

Sums of Squares and the ANOVA Partition

- We can ***partition*** the **total variation** in the data into three parts:
  - One reflecting variation **between** the levels of **Factor A**.
  - Another reflecting variation **between** the levels of **Factor B**.
  - The other reflecting variation **within** the groups.

Sums of Squares and the ANOVA Partition

- We can ***partition*** the **total variation** in the data into three parts:
  - One reflecting variation **between** the levels of **Factor A**.
  - Another reflecting variation **between** the levels of **Factor B**.
  - The other reflecting variation **within** the groups.

The ***ANOVA*** F ***tests*** are based on the amount of variation **between** levels of the factor relative to the amount of variation **within** groups.

- The **partition** will involve the following ***sums of squares*** (shown with their **df**):

- The **partition** will involve the following ***sums of squares*** (shown with their **df**):
  - **SST** is the ***total sum of squares***, defined as

$$\text{SST} = \sum_{i=1}^{I} \sum_{j=1}^{J} (X_{ij} - \bar{X}_{\cdot\cdot})^2 \qquad \boldsymbol{df = IJ - 1}$$

which measures the **total** variation in the X_{ij}'s.

- (cont'd):
  - **SSA** is the ***Factor A sum of squares***, defined as

$$\begin{aligned} \text{SSA} &= \sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2 \\ &= J\sum_{i=1}^{I}(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2 \qquad \boldsymbol{df = I - 1} \end{aligned}$$

    which measures variation between the **levels** of **Factor A** due to both the **Factor A effect** and **random error**.

- (cont'd):
  - **SSB** is the ***Factor B sum of squares***, defined as

$$\text{SSB} = \sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{X}_{\cdot j} - \bar{X}_{\cdot\cdot})^2$$

$$= I\sum_{i=1}^{I}(\bar{X}_{\cdot j} - \bar{X}_{\cdot\cdot})^2 \qquad \boldsymbol{df = J - 1}$$

which measures variation between the **levels** of **Factor B** due to both the **Factor B effect** and **random error**.

- (cont'd):
  - **SSE** is the ***error sum of squares***, defined as

$$\text{SSE} = \sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij}-\bar{X}_{i\cdot}-\bar{X}_{\cdot j}+\bar{X}_{\cdot\cdot})^2 \qquad \boldsymbol{df = (I-1)(J-1)}$$

which measures variation of the X_{ij}'s **within** treatment groups due to **random error**.

Proposition

ANOVA Partition of the Total Variation: It can be shown that

$$\text{SST} = \text{SSA} + \text{SSB} + \text{SSE}\,.$$

Additive Property of Degrees of Freedom:

$$\text{df for SST} = \text{df for SSA} + \text{df for SSB} + \text{df for SSE}$$

Mean Squares

- The ***Factor A mean square***, ***Factor B mean square***, and ***mean squared error*** are:

$$\begin{aligned} \text{MSA} &= \frac{\text{SSA}}{I-1} \\ \text{MSB} &= \frac{\text{SSB}}{J-1} \\ \text{MSE} &= \frac{\text{SSE}}{(I-1)(J-1)} \end{aligned}$$

The Two-Factor ANOVA F Tests

Two-Factor ANOVA F Test Statistics:

$$F_A = \frac{\text{MSA}}{\text{MSE}} \quad \text{and} \quad F_B = \frac{\text{MSB}}{\text{MSE}}$$

- F_A reflects variation **between** levels of Factor A (**MSA**) relative to **within**-groups variation (**MSE**).

- F_A reflects variation **between** levels of Factor A (**MSA**) relative to **within**-groups variation (**MSE**).

  F_A will be **large** when there's substantial variation in $\bar{X}_{1\cdot}, \bar{X}_{2\cdot}, \ldots, \bar{X}_{I\cdot}$, which are estimates of the true level means $\mu_{1\cdot}, \mu_{2\cdot}, \ldots, \mu_{I\cdot}$.

- F_A reflects variation **between** levels of Factor A (**MSA**) relative to **within**-groups variation (**MSE**).

  F_A will be **large** when there's substantial variation in $\bar{X}_{1\cdot}, \bar{X}_{2\cdot}, \ldots, \bar{X}_{I\cdot}$, which are estimates of the true level means $\mu_{1\cdot}, \mu_{2\cdot}, \ldots, \mu_{I\cdot}$.

- F_B reflects variation **between** levels of Factor B (**MSB**) relative to **within**-groups variation (**MSE**).

- F_A reflects variation **between** levels of Factor A (**MSA**) relative to **within**-groups variation (**MSE**).

  F_A will be **large** when there's substantial variation in $\bar{X}_{1\cdot}, \bar{X}_{2\cdot}, \ldots, \bar{X}_{I\cdot}$, which are estimates of the true level means $\mu_{1\cdot}, \mu_{2\cdot}, \ldots, \mu_{I\cdot}$.

- F_B reflects variation **between** levels of Factor B (**MSB**) relative to **within**-groups variation (**MSE**).

  F_B will be **large** when there's substantial variation in $\bar{X}_{\cdot 1}, \bar{X}_{\cdot 2}, \ldots, \bar{X}_{\cdot J}$, which are estimates of the true level means $\mu_{\cdot 1}, \mu_{\cdot 2}, \ldots, \mu_{\cdot J}$.

- F_A reflects variation **between** levels of Factor A (**MSA**) relative to **within**-groups variation (**MSE**).

  F_A will be **large** when there's substantial variation in $\bar{X}_{1\cdot}, \bar{X}_{2\cdot}, \ldots, \bar{X}_{I\cdot}$, which are estimates of the true level means $\mu_{1\cdot}, \mu_{2\cdot}, \ldots, \mu_{I\cdot}$.

- F_B reflects variation **between** levels of Factor B (**MSB**) relative to **within**-groups variation (**MSE**).

  F_B will be **large** when there's substantial variation in $\bar{X}_{\cdot 1}, \bar{X}_{\cdot 2}, \ldots, \bar{X}_{\cdot J}$, which are estimates of the true level means $\mu_{\cdot 1}, \mu_{\cdot 2}, \ldots, \mu_{\cdot J}$.

- In other words, F_A will be **large** when **Factor A** has an **effect**, and F_B will be **large** when **Factor B** has an **effect**.

Large values of F_A provide **evidence against H_0 in favor of H_a : Not all of the α_i's are zero.**

Large values of F_B provide **evidence against H_0 in favor of H_a : Not all of the β_j's are zero.**

- Suppose data in a two-factor study follow the **two-factor ANOVA model**, where the error terms ϵ_{ijk} are iid $N(0, \sigma)$.

Sampling Distributions of the Test Statistics Under H_0:

1. If F_A is the F test statistic for Factor A, then when

$$H_{0A} : \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$$

   is true,

$$F \sim F(I-1, (I-1)(J-1)).$$

2. If F_B is the tF test statistic for Factor B, then when

$$H_{0B} : \beta_1 = \beta_2 = \cdots = \beta_J = 0$$

   is true,

$$F \sim F(J-1, (I-1)(J-1)).$$

- The $F(I-1,\ (I-1)(J-1))$ and $F(J-1,\ (I-1)(J-1))$ curves give us:

- The $F(I-1, \ (I-1)(J-1))$ and $F(J-1, \ (I-1)(J-1))$ curves give us:
    - The ***rejection regions*** as the **extreme largest 100α% of F values**.

- The $F(I-1,\ (I-1)(J-1))$ and $F(J-1,\ (I-1)(J-1))$ curves give us:
  - The ***rejection regions*** as the **extreme largest 100α% of F values**.
  - The ***p-values*** as the **tail areas to the right of the observed F_A and F_B values**.

The ANOVA Table

- ANOVA results are summarized in an ***ANOVA table***:

The ANOVA Table

- ANOVA results are summarized in an ***ANOVA table***:

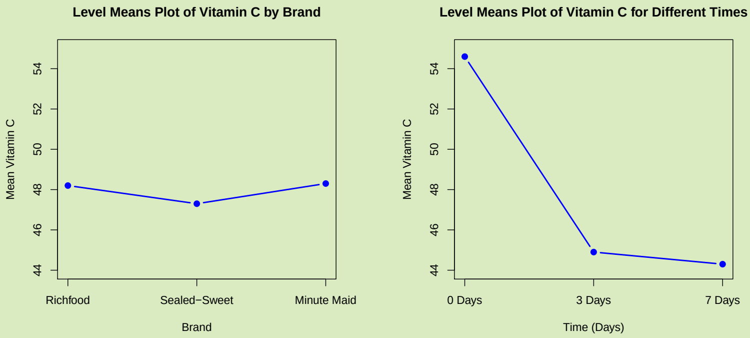
Source of Variation	**df**	**Sum of Squares**	**Mean Square**	**f**	**P-value**
Factor A	$I-1$	SSA	MSA = SSA/$(I-1)$	F_A =MSA/MSE	p
Factor B	$J-1$	SSB	MSB = SSB/$(J-1)$	F_B =MSB/MSE	p
Error	$(I-1)(J-1)$	SSE	MSE = SSE/$(I-1)(J-1)$		
Total	$IJ-1$	SST			

Exercise

For the study of the effects of **brand** and **storage time** on vitamin C in orange juice, the **ANOVA table** is below.

Source of Variation	df	Sum of Squares	Mean Square	f	P-value
Brand	2	1.70	0.85	0.101	0.9058
Time	2	199.94	99.97	11.961	0.0205
Error	4	33.43	8.36		
Total	8	235.06			

Here are so-called ***level means plots***.



What conclusions can be drawn?