What to do today (March 7, 2023)?

Part 1. Introduction and Review (Chp 1-5)

Part 2. Basic Statistical Inference (Chp 6-9)

Part 3. Important Topics in Statistics (Chp 10-13)

§3.1. Analysis of Variance (ANOVA, Chp 10-11)

§3.1.1 Introduction

§3.1.2 One-Factor ANOVA (Chp 10)

§3.1.3 Multi-Factor ANOVA (Chp 11)

§3.1.4 Further Topics on ANOVA

§3.2 Introduction to Regression Analysis (Chp 12-13)

Part 4. Further Topics (Selected from Chp 14-16)

Some Logistics.

- Homework 7 has been assigned. It's due on Monday March 13.
- Midterm 2 will be on Friday March 10, to cover Chp 6-10.

§3.1.2B More on Multiple Comparisons

(i) Tukey's Procedure

After One-factor ANOVA answers yes/no difference among I groups, Tukey's procedure identifies which two groups are different, using simultaneous CI of $\mu_i - \mu_j$, for any i, j

- *Constructing CI:* $\mu_i - \mu_j$**'s** $1-\alpha$ **CI** is

$$(\bar{X}_{i.} - \bar{X}_{j.}) \pm W_{ij}$$

  $W_{ij} = Q_\alpha(I, n_T - I)\sqrt{\frac{MSS_e}{2}(\frac{1}{n_i} + \frac{1}{n_j})}$ with $Q_\alpha(I, n_T - I)$ the upper-tail α critical value of the studentized range distn.

- *Making inference:* If $|\bar{X}_{i.} - \bar{X}_{j.}| > W_{ij}$, conclude μ_i and μ_j are significantly different at level α

§3.1.2B More on Multiple Comparisons

(ii). CI of Other Parameters in one-factor ANOVA
For example, ...

- to estimate $\theta = \sum_{i=1}^{I} c_i \mu_i$?

$$\hat{\theta} = \sum_{i=1}^{I} c_i \hat{\mu}_i = \sum_{i=1}^{I} c_i \bar{X}_{i.}$$

$$var\left(\hat{\theta}\right) = \sum_{i=1}^{I} c_i^2 var\left(\hat{\mu}_i\right) = \sum_{i=1}^{I} c_i^2 \frac{\sigma^2}{n_i} = \sum_{i=1}^{I} \frac{c_i^2}{n_i} \hat{\sigma}^2$$

$\Longrightarrow$ $(1-\alpha)100\%$ CI of $\theta = \sum_{i=1}^{I} c_i \mu_i$:

$$\hat{\theta} \mp ME(\hat{\theta}) = \sum_{i=1}^{I} c_i \bar{X}_{i.} \mp t_{\alpha/2}(n_T - I) \sqrt{\sum_{i=1}^{I} \frac{c_i^2}{n_i} MSE}$$

§3.1.3 Two-Factor ANOVA (Chp 11)

- **Setting.**
  - (i) A study is concerned with two factors A, B with I, J levels, to answer whether the outcomes are closely associated with the factors, jointly or individually.
    e.g., body weights according to gender (F,M: I=2) and age (E,M,Y: J=3)?
  - (ii) Observations: kth obs in (i, j)th group X_{ijk}, $k = 1, \ldots, n_{ij}$ and $i = 1, \ldots, I$, $j = 1, \ldots, J$ with $n_T = \sum_i \sum_j n_{ij}$. ($n_{ij} \equiv n$, balanced study)

- **Formulation (I).**

$$X_{ijk} = \mu_{ij} + \epsilon_{ijk}, \quad \epsilon_{ijk} \sim N(0, \sigma^2) \text{ iid}$$

To test $H_0 : \mu_{ij} = \mu$ for all i, j vs $H_1 : \textit{otherwise}$

$\Longrightarrow$ **one-factor ANOVA**: when $n_T > IJ$ (at least one $n_{ij} > 1$)

Source of Variation	df	SS	MSS	F-value
treatment	IJ-1	SS_{tr}	$\frac{SS_{tr}}{(IJ-1)}$	$F = \frac{MSS_{tr}}{MSS_e}$
error	$n_T - IJ$	SS_e	$\frac{SS_e}{(n_T-IJ)}$	
total	$n_T - 1$	SS_T	$\frac{SS_T}{(n_T-1)}$	

$$SS_T = SS_{tr} + SS_e$$

$$SS_T = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}} (X_{ijk} - \bar{X}_{\dots})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}} X_{ijk}^2 - n_T\bar{X}_{\dots}^2$$

$$SS_e = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}} (X_{ijk} - \bar{X}_{ij.})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}} X_{ijk}^2 - \sum_{i=1}^{I}\sum_{j-1}^{J} n_{ij}\bar{X}_{ij.}^2$$

How about factor A's (or B's) individual effect? How about factors A and B's interaction?

- **Formulation (II).** Consider

$$\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

α_i: factor A's main effect with $\sum_{i=1}^{I} \alpha_i = 0$,
β_j: factor B's main effect with $\sum_{j=1}^{J} \beta_j = 0$,
$(\alpha\beta)_{ij}$: factors A and B's interaction with
$\sum_{i=1}^{I} (\alpha\beta)_{ij} = \sum_{j=1}^{J} (\alpha\beta)_{ij} = 0$.

Two-factor ANOVA model.

$$X_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \quad \epsilon_{ijk} \sim N(0, \sigma^2) \text{ iid}$$

$k = 1, \ldots, n_{ij}$ and $i = 1, \ldots, I$, $j = 1, \ldots, J$.

To test the following 3 sets of hypotheses:
$H_{0A} : \alpha_i = 0$ for all i vs H_{1A} : *otherwise*
$H_{0B} : \beta_j = 0$ for all j vs H_{1B} : *otherwise*
$H_{0AB} : (\alpha\beta)_{ij} = 0$ for all i, j vs H_{1AB} : *otherwise*

To consider individual variation:

$$X_{ijk} - \bar{X}_{...} = (\bar{X}_{i..} - \bar{X}_{...}) + (\bar{X}_{.j.} - \bar{X}_{...}) + (\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}_{...}) + (X_{ijk} - \bar{X}_{ij.})$$

How about total variation decomposition?

- **Sum of Squares.**
  $n_T = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$, the total number of study observations.

[a]. Sum squares total: (Variation total)

$$SS_T = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} (X_{ijk} - \bar{X}_{...})^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} X_{ijk}^2 - n_T \bar{X}_{...}^2$$

[b]. Sum squares error: (Variation due to individual fluctuation)

$$SS_e = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} (X_{ijk} - \bar{X}_{ij.})^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} X_{ijk}^2 - \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \bar{X}_{ij.}^2$$

[c]. Sum squares factor A/B: (Variation due to factor A/B's diff levels)

$$SS_A = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}}(\bar{X}_{i..} - \bar{X}_{...})^2 = \sum_{i=1}^{I}(\sum_{j=1}^{J} n_{ij})\bar{X}_{i..}^2 - n_T\bar{X}_{...}^2$$

$$SS_B = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}}(\bar{X}_{.j.} - \bar{X}_{...})^2 = \sum_{j=1}^{J}(\sum_{i=1}^{I} n_{ij})\bar{X}_{.j.}^2 - n_T\bar{X}_{...}^2$$

[d]. Sum squares A,B interaction: (Variation due to factors A and B's interaction)

$$SS_{AB} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}}(\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}_{...})^2$$

with

$$\begin{aligned} & \bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}_{...} \\ = \ & [\bar{X}_{ij.} - \bar{X}_{...}] - [\bar{X}_{i..} - \bar{X}_{...}] - [\bar{X}_{.j.} - \bar{X}_{...}] \\ \approx \ & (\mu_{ij} - \mu) - \alpha_i - \beta_i = (\alpha\beta)_{ij} \end{aligned}$$

Relationships: Only in a balanced study ($n_{ij} \equiv n$)

$$SS_T = SS_A + SS_B + SS_{AB} + SS_e$$

with $SS_{tr} = SS_A + SS_B + SS_{AB}$ (*Otherwise, no such decomposition and a different method needed*)

Further

$$E\{MSS_e\} = E\{\frac{SS_e}{n_T - IJ}\} = \sigma^2$$

with $n_T = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}$: it's nIJ if $n_{ij} \equiv n$.

$$E\{MSS_{tr}\} = E\{\frac{SS_{tr}}{IJ - 1}\} = \sigma^2 + \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}(\mu_{ij} - \mu)^2}{IJ - 1}$$

$$E\{MSS_A\} = E\{\frac{SS_A}{I - 1}\} = \sigma^2 + \frac{nJ}{I - 1}\sum_{i=1}^{I}\alpha_i^2$$

$$E\{MSS_B\} = E\{\frac{SS_B}{J - 1}\} = \sigma^2 + \frac{nI}{J - 1}\sum_{j=1}^{J}\beta_j^2$$

$$E\{MSS_{AB}\} = \sigma^2 + \frac{n}{(I - 1)(J - 1)}\sum_{i=1}^{I}\sum_{j=1}^{J}(\alpha\beta)_{ij}^2$$

- **Two-factor ANOVA Table.** (in a balanced study: $n_T = IJn$)

Source of Variation	df	SS	MSS	F-value
A	I-1	SS_A	$\frac{SS_A}{(I-1)}$	$F_A = \frac{MSS_A}{MSS_e}$
B	J-1	SS_B	$\frac{SS_B}{(J-1)}$	$F_B = \frac{MSS_B}{MSS_e}$
AB	(I-1)(J-1)	SS_{AB}	$\frac{SS_{AB}}{(I-1)(J-1)}$	$F_{AB} = \frac{MSS_{AB}}{MSS_e}$
error	$n_T - IJ$	SS_e	$\frac{SS_e}{(n_T-IJ)}$	
total	$n_T - 1$	SS_T		

- **F-Test (on Factor A).**
  $H_{0A} : \alpha_i = 0$ vs H_{1A} : *otherwise*

$$F_A = \frac{SS_A/(I-1)}{SS_e/(n_T - IJ)} \sim F(I-1, n_T - IJ)$$

  under H_{0A}.
  At α-significance level, $c_A = f_\alpha(I-1, n_T - IJ)$ such that $P_{H_{0A}}(F_A > c_A) = \alpha$; reject H_{0A} if $F_{A,obs} > c_A$.

- **F-Test (on Factor B).**
  $H_{0B} : \beta_j = 0$ vs H_{1B} : *otherwise*

  $$F_B = \frac{SS_B/(J-1)}{SS_e/(n_T - IJ)} \sim F(J-1, n_T - IJ)$$

  under H_{0B}.
  At α-significance level, $c_B = f_\alpha(J-1, n_T - IJ)$ such that $P_{H_{0B}}(F_B > c_B) = \alpha$; reject H_{0B} if $F_{B,obs} > c_B$.

- **F-Test (on Factors A,B interaction).**
  $H_{0AB} : (\alpha\beta)_{ij} = 0$ vs H_{1AB} : *otherwise*

  $$F_{AB} = \frac{SS_{AB}/(I-1)(J-1)}{SS_e/(n_T - IJ)} \sim F((I-1)(J-1), n_T - IJ)$$

  under H_{0AB}.
  At α-significance level, $c_{AB} = f_\alpha((I-1)(J-1), n_T - IJ)$ such that $P_{H_{0AB}}(F_{AB} > c_{AB}) = \alpha$; reject H_{0AB} if $F_{AB,obs} > c_{AB}$.

Example 6.3

- **Study.** to compare yields of 3 different tomato varieties and 4 different plant densities, each combination of variety is used in 3 plots
- **Formulation.** two-factor ANOVA model; to test H_{0A}, H_{0B}, H_{0AB} at level $\alpha = 5\%$
- **Data.** $I = 3$, $J = 4$ and $n = 3$: x_{ijk}

A	B: 1	2	3	4	$x_{i..}$	$\bar{x}_{i..}$
1	$x_{111}, x_{112}, x_{113}$	...	...		...	...
2	$x_{211}, x_{212}, x_{213}$	...	...		...	...
3	... ...					
$x_{.j.}$	... ...				$x_{...} = 500$	
$\bar{x}_{.j.}$	... ...					$\bar{x}_{...} = 13.89$

- **ANOVA table.**

Source of Variation	df	SS	MSS	F-value
A	$3-1$	327.60	163.8	$F_{A,obs} = 103.02$
B	$4-1$	86.69	28.9	$F_{B,obs} = 18.18$
AB	$(3-1)(4-1)$	8.03	1.34	$F_{AB,obs} = .84$
error	24	38.04	1.59	
total	35	460.36		

- **Making inference.**

$f_{0.01}(2,24) = 5.61 < F_{A,obs} \Longrightarrow$ reject H_{0A}.
$f_{0.01}(3,24) = 4.72 < F_{B,obs} \Longrightarrow$ reject H_{0B}.
$f_{0.01}(6,24) = 3.67 > F_{AB,obs} \Longrightarrow$ don't reject H_{0AB}.

Note that *roughly* $\mu_{ij} = \mu + \alpha_i + \beta_j$: *e.g.* $\mu_{1j} - \mu_{2j} = \alpha_1 - \alpha_2$; $\mu_{i2} - \mu_{i3} = \beta_2 - \beta_3$.

What if $n \not> 1$ (i.e. $n = 1$)?

Example 6.4 (p438)

- **Study.** to remove marks on fabrics from erasable pens with A: brand of pen and B: wash treatment.
- **Data.** overall specimen color change (lower, better): $I = 3$, $J = 4$ and $n = 1$

	B					
A	1	2	3	4	total	average
1	.97		... ...		2.39	.598
2	.77		... ...		1.38	.345
3	.67		... ...		1.82	.455
total	2.41	1.01	1.27	.90	5.59	
average		... ...				.466

- **To test on** H_{0A}, H_{0B} and H_{0AB}?
  two factor study but $n = 1$: in 2-factor ANOVA table $n_T = IJ$
  $\Longrightarrow$ consider $\mu_{ij} = \mu + \alpha_i + \beta_j$.

- **ANOVA table.**

Source of Variation	df	SS	MSS	F-value
A	3-1	0.128	...	$F_{A,obs} = 4.43$
B	4-1	0.480	...	$F_{B,obs} = 11.05$
error	$(3-1)(4-1)$	0.087	...	
total	12-1	0.695	...	

- **Making inference.**

$f_\alpha(2,6) = 5.14 > F_{A,obs} \Longrightarrow$ don't reject H_{0A}.
$f_\alpha(3,6) = 4.76 > F_{B,obs} \Longrightarrow$ reject H_{0B}.

How about 3-factor ANOVA? How about multi-factor ANOVA?

To be discussed next ... ...

What will we study next?

Part 1. Introduction and Review (Chp 1-5)

Part 2. Basic Statistical Inference (Chp 6-9)

Part 3. Important Topics in Statistics (Chp 10-13)

- *3.1A One-Factor Analysis of Variance (Chp 10)*
- *3.1B Multi-Factor ANOVA (Chp 11)*
- **3.2A Simple Linear Regression Analysis (Chp 12)**
- **3.2B More on Regression (Chp 13)**

Part 4. Further Topics (Selected from Chp 14-16)