What to do today (March 3, 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 6 has been assigned. It's due on Monday March 6.
- Midterm 2 will be on Friday March 10, to cover Chp 6-11.

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

§3.1.1 Introduction

- multiple comparisons (multiple-sample problems)
- why ANOVA?

§3.1.2 One-Factor ANOVA (Chp 10)

§3.1.3 Two-Factor ANOVA (Chp 11)

§3.1.4 Discussion

3.1.2 One-Factor ANOVA

- **Setting.**

(i) A study is concerned with one factor with I levels, to answer whether the outcomes are closely associated with the factor.

- e.g., to study voting trend according to education with 4 levels: less than high-school, high-school graduate, colledge/univeristy degree, post-graduate degree

(ii) Observations: jth obs in ith group X_{ij}, $j = 1, \dots, n_i$ and $i = 1, \dots, I$ ($n_i \equiv n$, balanced study)

	1	2	3	...	I	Total
	X_{11}	X_{21}	X_{31}	...	X_{I1}	
	X_{12}	X_{22}	X_{32}	...	X_{I2}	
	...	...	...	...	...	
	...	...	...	...	...	
	X_{1n}	X_{2n}	X_{3n}	...	X_{In}	
Total	$X_{1.}$	$X_{2.}$	$X_{3.}$	...	$X_{I.}$	$X_{..}$
Mean	$\bar{X}_{1.}$	$\bar{X}_{2.}$	$\bar{X}_{3.}$	...	$\bar{X}_{I.}$	$\bar{X}_{..}$

- **Formulation.**

(i) $X_1, \ldots, X_I$ are independent, and associated with the I groups' outcomes: $X_i \sim N(\mu_i, \sigma^2)$

(ii) To test $H_0 : \mu_1 = \ldots = \mu_I$ vs H_1 : *otherwise*

- **One-Factor ANOVA Model.**

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

$j = 1, \ldots, n_i$ and $i = 1, \ldots, I$

- **Summary Statistics.**

(i) *Sample Means.*

[i.a]. Sample mean of the ith group: $i = 1, \ldots, I$,

$$\bar{X}_{i.} = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i}; \quad E(\bar{X}_{i.}) = \mu_i$$

[i.b]. Overall (grand) sample mean:

$$\bar{X}_{..} = \frac{\sum_{i=1}^{I} \sum_{j=1}^{n_i} X_{ij}}{\sum_{i=1}^{I} n_i}; \quad E(\bar{X}_{..}) = \frac{\sum_{i=1}^{I} n_i \mu_i}{\sum_{i=1}^{I} n_i}$$

$E(\bar{X}_{..}) = \mu_.$ is $\sum_{i=1}^{I} \mu_i / I$ if $n_i \equiv n$

(ii) *Sum of Squares.*

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

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

$n_T = \sum_{i=1}^{I} n_i$, the total number of study observations.

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

$$SS_e = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i.})^2 = \sum_{i=1}^{I}\left\{ \sum_{j=1}^{n_i} X_{ij}^2 - n_i \bar{X}_{i.}^2 \right\}$$

[ii.c]. Sum squares treatment: (Variation due to factor's diff levels)

$$\begin{aligned} SS_{tr} &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}(\bar{X}_{i.} - \bar{X}_{..})^2 = \sum_{i=1}^{I} n_i(\bar{X}_{i.} - \bar{X}_{..})^2 \\ &= \sum_{i=1}^{I} n_i \bar{X}_{i.}^2 - n_T \bar{X}_{..}^2 \end{aligned}$$

(iii) *Relationship.*

$$X_{ij} = \mu_i + \epsilon_{ij} : \quad \epsilon_{ij} \sim N(0, \sigma^2)$$

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

$$\Longrightarrow \sum_i \sum_j (X_{ij} - \bar{X}_{..})^2 = \sum_i \sum_j (X_{ij} - \bar{X}_{i.})^2 + \sum_i n_i(\bar{X}_{i.} - \bar{X}_{..})^2$$

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

$$E\big\{SS_e/(n_T - I)\big\} = \sigma^2$$

with $n_T = \sum_{i=1}^{I} n_i$: it's nI if $n_i \equiv n$.

$$E\big\{SS_{tr}/(I-1)\big\} = \sigma^2 + \frac{\sum_{i=1}^{I} n_i(\mu_i - \mu_.)^2}{I-1}$$

- **F-Test.** $H_0 : \mu_1 = \ldots = \mu_I$ vs H_1 : *otherwise*

$$F = \frac{SS_{tr}\big/(I-1)}{SS_e\big/(n_T - I)} \sim F(I-1, n_T - I)$$

under H_0.

At α-significance level, $c = f_\alpha(I-1, n_T - I)$ such that $P_{H_0}(F > c) = \alpha$; reject H_0 if $F_{obs} > c$.

- **ANOVA Table.**

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

Example 6.1

- **Study.** to compare 5 brands of automobile oil filters (I=5)
- **Data.** $n_i = 9$ outcomes with each brand
- **Formulation.** indpt $X_i \sim N(\mu_i, \sigma^2)$, $i = 1, \ldots, 5$; to test $H_0 : \mu_1 = \ldots = \mu_5$ vs H_1 : *otherwise* at level $\alpha = 5\%$
- **ANOVA table.**

Source of Variation	df	SS	MSS	F-value
factor	5-1	13.32	3.33	$F_{obs} = 37.84$
error	$45 - 5$	3.53	0.088	
total	44	16.85		

- Making inference.
  $f_\alpha(4, 40) = 2.60 < F_{obs} \Longrightarrow$ reject H_0.

§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 $\theta = \sum_{i=1}^{I} c_i \mu_i$:

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

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)