What to do today (March 14, 2023)?

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)

§3.2.1 Introduction

§3.2.2 Simple Linear Regression (Chp 12)

§3.2.3 More Advanced Topics (Chp 13)

Some Logistics.

- Homework 8 has been assigned. It's due on Monday March 20.
- Marked Midterm 2 papers will be distributed at next week's tutorial.

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}.

§3.1.4 Further Topics on ANOVA

3.1.4A Multi-factor ANOVA

For example, a study on how adult body weights relate to (A) gender (f,m), (B) age (y,m,e) and (C) education (lh, h, u, pu)
$\Longrightarrow$ a 3-factor study: $I = 2$, $J = 3$ and $K = 4$.

what to consider: (i) main effects of A, B, C? (ii) two-factor interactions: AB,BC,AC? *(iii) three-factor interactions: ABC?*

observations. X_{ijkl}: lth obs from group (i, j, k), $l = 1, \ldots, n_{ijk}$ with $i = 1, \ldots, I$, $j = 1, \ldots, J$ and $k = 1, \ldots, K$.

3-factor ANOVA model.

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

$\mu_{ijk} = \mu + \alpha_i + \beta_i + \gamma_k + (\alpha\beta)_{ij} + (\beta\gamma)_{jk} + (\alpha\gamma)_{ik} + (\alpha\beta\gamma)_{ijk}$ with constraints $\sum_i \alpha_i = 0$, ...

hypotheses to test?

- **Set 1. on main effects**

  $H_{0A} : \alpha_i = 0$ vs H_{1A} : *otherwise*;

  $H_{0B} : \beta_j = 0$ vs H_{1B} : *otherwise*;

  $H_{0C} : \gamma_k = 0$ vs H_{1C} : *otherwise*

- **Set 2. on two factor interactions**

  $H_{0AB} : (\alpha\beta)_{ij} = 0$ vs H_{1AB} : *otherwise*;

  $H_{0BC} : (\beta\gamma)_{jk} = 0$ vs H_{1BC} : *otherwise*;

  $H_{0AC} : (\alpha\gamma)_{ik} = 0$ vs H_{1AC} : *otherwise*

- **Set 3. on three factor interactions**

  $H_{0ABC} : (\alpha\beta\gamma)_{ijk} = 0$ vs H_{1ABC} : *otherwise*.

variation decomposition. Only if $n_{ijk} \equiv n > 1$

$$SS_T = SS_A + SS_B + SS_C + SS_{AB} + SS_{BC} + SS_{AC} + SS_{ABC} + SS_e$$

testing procedures. Test statistics: for example

- $F_A = \frac{MSS_A}{MSS_e} = \frac{SS_A/(I-1)}{SS_e/(n_T-IJK)} \sim F(I-1, (n-1)IJK)$ under H_{0A}.
- $F_{AB} = \frac{MSS_{AB}}{MSS_e} = \frac{SS_{AB}/(I-1)(J-1)}{SS_e/(n_T-IJK)} \sim F((I-1)(J-1), (n-1)IJK)$ under H_{0AB}
- $F_{ABC} = \frac{MSS_{ABC}}{MSS_e} = \frac{SS_{ABC}/(I-1)(J-1)(K-1)}{SS_e/(n_T-IJK)} \sim$ $F((I-1)(J-1)(K-1), (n-1)IJK)$ under H_{0ABC}

3.1.4B* ANOVA with Random (Mixed) Effects

For example, in a study with one factor A: I is large.

- **One-factor random effect ANOVA model.**

$$X_{ij} = \mu_i + \epsilon_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

$\alpha_i \sim N(0, \sigma_\alpha^2)$ iid, $i = 1, \ldots, I$; $\alpha_i \perp \epsilon_{ij}$;
$\epsilon_{ij} \sim N(0, \sigma^2)$ iid, $j = 1, \ldots, n_i$ and $i = 1, \ldots, I$.
$\Longrightarrow$ *The value of α_i is not of the primary interest but the patterns of α_i,* $i = 1, \ldots, I$

- **F-test.**

$$SS_T = SS_{tr} + SS_e; \quad E(SS_{tr}) = \sigma^2 + \frac{1}{I-1}(n_T - \frac{\sum_i n_i^2}{n_T})\sigma_\alpha^2$$

$H_0 : \sigma_\alpha^2 = 0$ vs $H_1 : otherwise$

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

under H_0

For another example, in 2-factor study: effect of A is random, effect of B is fixed
$\Longrightarrow$ **a mixed-effects ANOVA model**
e.g., a new drug's efficacy – 30 hospitals (sites) participate in the trial and both male and female subjects are enrolled

Example 6.6 (p457)

- **Study.** two potential causes of electric motor vibration: A. the materialused for the motor casing; B. the supply source of bearings used in the motor.
- **Data.** on amount of vibration: I=3, J=5, K=2
- **Model.** (two-factor mixed effects model with α_i as fixed effect, β_j and $(\alpha\beta)_{ij}$ random): $X_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$, $\epsilon_{ijk} \sim N(0, \sigma^2)$ iid, $\beta_j \sim N(0, \sigma^2_\beta)$ and $(\alpha\beta)_{ij} \sim N(0, \sigma^2_{\alpha\beta})$; $\beta_j \perp (\alpha\beta)_{ij} \perp \epsilon_{ijk}$
- **Test** $H_{0A} : \alpha_i = 0$, $H_{0B} : \sigma^2_\beta = 0$ and $H_{0AB} : \sigma^2_{\alpha\beta} = 0$?
- **ANOVA table.** (using the MINITAB output)

Source of Variation	df	SS	MSS	F-value	p-value
A	3-1	0.7047	...	$F_{A,obs} = 0.24$	.790
B	5-1	36.6747	...	$F_{B,obs} = 6.32$	.013
AB	(3-1)(5-1)	11.6053	...	$F_{AB,obs} = 13.05$	$< .001$
error	$(3)(5)(2-1)$	1.67	...		
total	30-1	50.6547	...		

- **Making inference.**

  Using significance-level $\alpha = .05$,

  $\Longrightarrow$ reject H_{0B} and reject H_{0AB}.

What will we study next?

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

§3.2.1 Introduction

§3.2.2 Simple Linear Regression (Chp 12)

3.2.2A modeling
3.2.2B estimation of model parameters
3.2.2C additional inferences
3.2.2D residual analysis

§3.2.3 More Advanced Topics (Chp 13)

3.2.3A multiple linear regression
3.2.3B regression with transformed variables
3.2.3C categorical predictors
3.2.3D discussion

§3.2. Introduction to Regression Analysis

§3.2.1 Introduction. (*Why and What?*)

- Recall a function in math

$$x \longrightarrow y : y = f(x)$$

  e.g. $y = mileage, x = time \Longrightarrow y = ax$ with $a = speed$ if the speed over $[0, x]$ is uniform.

- In reality, examples of "given x, is y fully determined"?
  e.g. $x = height, y = weight$: can we have $y = f(x)$?

- What if it is of interest to establish how a variable Y depends on another variable X?
  $\Longrightarrow$ **Regression Analysis**.

- *Key idea*: focus on studying $E(Y|X) = f(X)$

$$Y = f(X) + \epsilon, \quad E(\epsilon) = 0$$

  What is $f(\cdot)$?

  - to start with $f(\cdot)$ is a linear function:

    $$Y = \beta_0 + \beta_1 X + \epsilon$$

    $\Longrightarrow$ **Simple Linear Regression Analysis (Chp 12)**.

  - if $f(\cdot)$ is not linear? $\Longrightarrow$ **Nonlinear Regression Analysis**.

- What if it is of interest to establish how a variable Y depends on several variables $X_1, \ldots, X_K$?
  $\Longrightarrow$ **Multiple Linear (Nonlinear) Regression Analysis (Chp 13)**.

§3.2.2 Simple Linear Regression (Chp 12)

§3.2.2A Modeling

Goal. to establish how r.v. Y depends on a variable X linearly

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Y: response variable, dependent variable
- X: explanatory variable, independent variable, predictor
- ϵ: random error $E(\epsilon) = 0$ and $V(\epsilon) = \sigma^2$, and $X \perp\!\!\!\perp \epsilon$
- parameters β_0 and β_1: intercept and slope

Data. Consider a study of n independent units,
with the values of the predictor X are $x_1, \ldots, x_n$, corresponding to the observed responses $y_1, \ldots, y_n$, respectively.

i (obs)	1	2	...	...	n	total	squares total
x_i	x_1	x_2	...	...	x_n	$\sum x_i$	$\sum x_i^2$
y_i	y_1	y_2	...	...	y_n	$\sum y_i$	$\sum y_i^2$
	cross-product total: $\sum x_i y_i$						

Simple Linear Regression Model
Given data from independent units: $\{(X_i, Y_i) : i = 1, \ldots, n\}$

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

with ϵ_i independent and $E(\epsilon_i) = 0$ and $V(\epsilon_i) = \sigma^2$.

What are β_0, β_1 and σ^2?
$\Longrightarrow$ **to estm the parameter with the data ...**
that is, to fit the regression model

§3.2.2B Estimation of model parameters

Recall the **Simple Linear Regression Model**:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

for $i = 1, \ldots, n$, ϵ_i's are independent and $E(\epsilon_i) = 0$, $V(\epsilon_i) = \sigma^2$.

(i) to estm β_0, β_1.

thinking ... If b_0 and b_1 are good estimates for β_0 and β_1, $Y_i - (b_0 + b_1 X_i)$ should be small for all i:

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n} \left\{ Y_i - (\beta_0 + \beta_1 X_i) \right\}^2$$

should be small at (b_0, b_1).

Least Squares Estimation (LSE). The estimators $\hat{\beta}_0, \hat{\beta}_1$ are the LSE of β_0, β_1, if

$$L(\hat{\beta}_0, \hat{\beta}_1) = \min_{\beta_0, \beta_1} L(\beta_0, \beta_1).$$

The LSE of β_0, β_1 is the solution to

$$\begin{cases} \frac{\partial L(\beta_0,\beta_1)}{\partial \beta_0} = \sum_{i=1}^n 2(Y_i - \beta_0 - \beta_1 X_i) = 0 \\ \frac{\partial L(\beta_0,\beta_1)}{\partial \beta_1} = \sum_{i=1}^n 2(Y_i - \beta_0 - \beta_1 X_i)X_i = 0 \end{cases}$$

$\bar{X} = \sum_i X_i/n$, $\bar{Y} = \sum_i Y_i/n$.

$\Longrightarrow$

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \end{aligned}$$

Often denote $\hat{\beta}_1 = S_{XY}/S_{XX}$ with

$$S_{XY} = \sum_i (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_i X_i Y_i - n\bar{X}\bar{Y}$$

and

$$S_{XX} = \sum_i (X_i - \bar{X})(X_i - \bar{X}) = \sum_i X_i^2 - n\bar{X}^2.$$

Properties of the LSE $\hat{\beta}_0$ and $\hat{\beta}_1$:

- **linear.** (linear functions of Y_i's)

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \bar{X}) Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2} = \sum_i \frac{(X_i - \bar{X})}{S_{XX}} Y_i \\ \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} = \sum_i \left[\frac{1}{n} - \frac{(X_i - \bar{X})\bar{X}}{S_{XX}}\right] Y_i\end{aligned}$$

- **unbiased.** Note that

$$E(\hat{\beta}_1) = \frac{E(S_{XY})}{S_{XX}},$$

$$E(S_{XY}) = \sum_i (X_i - \bar{X})(\beta_0 + \beta_1 X_i - [\beta_0 + \beta_1 \bar{X}]) = \beta_1 \sum_i (X_i - \bar{X})^2$$

$$\Longrightarrow E(\hat{\beta}_1) = \beta_1.$$

- **variance.**

$$\begin{aligned} V(\hat{\beta}_1) &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 V(Y_i)}{S_{XX}^2} = \frac{\sigma^2}{S_{XX}} \\ V(\hat{\beta}_0) &= \sigma^2[\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}] \end{aligned}$$

Study designs: to have a large S_{XX}. (Why?)

- **the best unbiased linear estimator**

$$V(\hat{\beta}_1) \leq V(\tilde{\beta}_1); V(\hat{\beta}_0) \leq V(\tilde{\beta}_0)$$

if $\tilde{\beta}_0, \tilde{\beta}_1$ are unbiased linear estimators of β_0, β_1.

(ii) to estm σ^2

thinking ... $\sigma^2 = V(\epsilon)$ can be estimated by the sample variance $\sum_{i=1}^{n}(\epsilon_i - \bar{\epsilon})^2/(n-1)$ if $\epsilon_i = Y_i - [\beta_0 + \beta_1 X_i]$ were observable.

How about using $e_i = Y_i - \hat{Y}_i = Y_i - [\hat{\beta}_0 + \hat{\beta}_1 X_i]$?

$\Longrightarrow$ an unbiased estimator of σ^2:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SS_e}{n-2} = MSS_e$$

Further inferences, e.g. CI and testing? What are the distributions of the LSE $\hat{\beta}_0$ and $\hat{\beta}_1$?

$\Longrightarrow$ to be studied later ... ...

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)