What to do today (March 28, 2023)?

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

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

§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)

3.2.3A Multiple linear regression

3.2.3B Regression with transformed variables

3.2.3C Regression with categorical predictors

3.2.3D Discussion

3.2.3E Comprehensive Example

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

§4.1 Distribution-Free Procedures (Chp15.1, 15.2)

§4.2 Quality Control Methods (Chp16.1, 16.2, 16.3)

For another example, to consider how Y is associted with Factors A and B: Factor A with 3 levels, Factor B with 4 levels, n_{ij} can be different.

"Two-Factor Study with Unbalanced Data"!

- $3-1$ dummy variables for Factor A: X_{A1}, X_{A2}; $4-1$ dummy variables for Factor B: X_{B1}, X_{B2}, X_{B3}
- Consider a multiple linear regression model

$$Y = \beta_0 + \beta_1 X_{A1} + \beta_2 X_{A2} + \gamma_1 X_{B1} + \gamma_2 X_{B2} + \gamma_3 X_{B3} + \epsilon$$

  or a regression model with cross-product terms

$$Y = \beta_0 + \beta_1 X_{A1} + \ldots + \gamma_3 X_{B3} + \phi_1 X_{A1} X_{B1} + \ldots + \epsilon$$

Balanced data type is not required in the regression analysis.

3.2.3D Discussion

- What does a regression analysis do?
- Variable selection, model selection in regression analysis
- What if Y is categorical?
  $\Longrightarrow$ Generalized Linear Models. *Categorical Data Analysis*
- What if $E(Y|X)$ is not specified into a linear form, or a parametric form?
- Model checking in (linear) regression analysis
  What if $Y_1, \ldots, Y_n$ are not indpt?
  e.g. they're market prices of a stock recorded on days 1, 2, ..., n? $\Longrightarrow$ Time Series: how to deal with?

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)

4.1 Distribution-Free Procedures (Chp15.1, 15.2)

4.2 Quality Control Methods (Chp16.1, 16.2, 16.3)

§4.1 Distribution-Free Procedures (Nonparametric Methods)

§4.1.1 Basic Concepts

- **order statistics.**

Definition. Suppose $X_1, \ldots, X_n$ are iid observations from a continuous r.v. $X \sim f(\cdot)$ with cdf $F(\cdot)$. The **order statistics** of the random sample are $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$: $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$.

$$X_{(1)} = \text{the smallest value of } X_1, \ldots, X_n,$$
$$X_{(2)} = \text{the 2nd smallest value of } X_1, \ldots, X_n, \text{ ... ...},$$
$$X_{(n)} = \text{the largest value of } X_1, \ldots, X_n.$$

Distribution. $X_{(k)} \sim \frac{n!}{(k-1)!(n-k)!} F(x)^{k-1}(1 - F(x))^{n-k} f(x)$ for $k = 1, \ldots, n$.

Example 10.1 Realizations of 5 iid observations $X_1, \ldots, X_5$ from a population are given in the table below.

x_1	x_2	x_3	x_4	x_5
0.62	0.98	0.31	0.81	0.53

The order statistics? The rank statistics?

- **rank statistics.**

Definition. The rank of X_k, the kth observation in a random sample of size n, is r_k such that $X_k = X_{(r_k)}$, for $k = 1, \ldots, n$.

- **percentiles/quantiles.**

Definition. Suppose r.v. $X \sim f(\cdot)$ with a random sample $X_1, \ldots, X_n$.
Population percentitles: π_p is the $(100p)$th percentile of the population if $P(X \leq \pi_p) = p$. That is, $\int_{-\infty}^{\pi_p} f(x)dx = p$.

Sample percentiles: Let $X_{(1)}, \ldots, X_{(n)}$ be the order statistics. Then $X_{(r)}$ is the $(r/n)100$th (or $(r/n+1)100$th) sample percentile.

- e.g., If $p = 0.5$, the population median m is the $(100p)$th population percentile.

  The order statistic $X_{(n+1/2)}$ is the sample median when n is odd; all values in between $X_{(n/2)}$ and $X_{(n/2+1)}$ are the sample median when n is even.

- e.g., If $X \sim N(0,1)$, $Z_{0.95}$ is X's 95th population percentile if $P(X \leq Z_{0.95}) = 0.95$.

  X's right-tailed critical value $z_{0.025}$, i.e., $P(X > z_{0.025}) = 0.025$, is its $(1 - 0.025)100$th percentile.

- **empirical distribution function.**

Definition. Suppose r.v. $X \sim F(\cdot)$ with a random sample $X_1, \ldots, X_n$. Its **empirical distribution** is defined as

$$\hat{F}_n(x) = \frac{1}{n}\#\{X_i : X_i \leq x; i = 1, \ldots, n\}, \quad x \in (-\infty, \infty).$$

That is, $\hat{F}_n(x) = 0$ if $x < X_{(1)}$; k/n, if $X_{(k)} \leq x < X_{(k+1)}$ when $1 \leq k \leq n-1$; 1, if $x \geq X_{(n)}$.

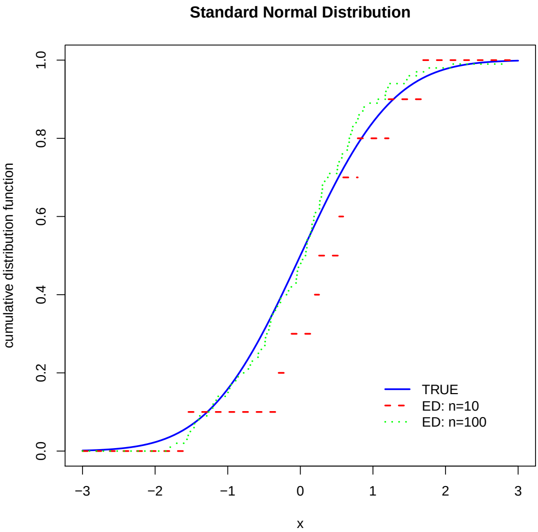
For a fixed x,

- $E[\hat{F}_n(x)] = F(x)$
- $Var[\hat{F}_n(x)] = F(x)[1 - F(x)]/n$.

- **empirical distribution function.**

Definition. Suppose r.v. $X \sim F(\cdot)$ with a random sample $X_1, \ldots, X_n$. Its **empirical distribution** is defined as

$$\hat{F}_n(x) = \frac{1}{n}\#\{X_i : X_i \le x; i = 1, \ldots, n\}, \quad x \in (-\infty, \infty).$$



§4.1.2 Nonparametric Testing Procedures

§4.1.2A Binomial test: (*the sign test*)

- **setting.** r.v. $X \sim f(\cdot)$ with a random sample $X_1, \ldots, X_n$. To test H_0 : its population median $m = m_0$ vs H_1 : otherwise.

- **test statistic.** $S_i = 1$ or 0 if $X_i - m_0 \geq 0$ or not.

$$S = \sum_{i=1}^{n} S_i = \#\{i : X_i - m_0 \geq 0; i = 1, \ldots, n\}$$

(i) $S \sim B(n, 1/2)$ under H_0.
(ii) if $n \gg 1$, $Z = \frac{S - n/2}{\sqrt{n/4}} \sim N(0, 1)$ approximately under H_0.

- **making inference.**

  (i) Obtain the two critical values c_1 and c_2 from the binomial table: $P_{H_0}(S > c_2) = \alpha/2$; $P_{H_0}(S < c_1) = \alpha/2$.
  Reject H_0 if $S_{obs} > c_2$ or $S_{obs} < c_1$. [the exact approach]

  (ii) If $n \gg 1$, reject H_0 if $|Z_{obs}| > z_{\alpha/2}$. [the approximate approach]

How about to test $H_0 : \pi_p = \pi_p^*$ for $0 < p < 1$?

Are there any other test procedures using more information?

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)

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

§4.1 Distribution-Free Procedures (Chp15.1, 15.2)

§4.1.1 Basic Concepts

§4.1.2 Nonparametric Testing Procedures

§4.2 Quality Control Methods (Chp16.1, 16.2, 16.3)

§4.2.1 Introduction

§4.2.2 Examples of Control Charts