Statistics

Random Samples, Statistics and Sampling Distributions

Shiu-Sheng Chen

Department of Economics
National Taiwan University

Fall 2019

Section 1

Random Samples and Descriptive Statistics

Random Samples

Definition (Random Samples)

A random sample with size n, $\{X_i\}_{i=1}^n = \{X_1, X_2, \ldots, X_n\}$, is a set of i.i.d. random variables.

- Random samples are also called I.I.D. samples.
- Notation

$$\{X_i\}_{i=1}^n \sim^{\text{i.i.d.}} (\mu, \sigma^2)$$

- Properties

$$E(X_1) = E(X_2) = \cdots = E(X_n) = \mu$$

$$Var(X_1) = Var(X_2) = \cdots = Var(X_n) = \sigma^2$$

$$E(X_iX_j) = E(X_i)E(X_j) \text{ for any } i \neq j$$

Descriptive Statistics

- Frequency Distribution Table
- Empirical Density Function (Histogram)
- Empirical Distribution Function
- Statistics and Sampling Distribution

Example: Statistics Midterm Exam

- 1st Midterm Exam Scores of 167 students in 2018
- 69 5 66 88 73 96 88 92 67 79 74 72 73 63 66 73 60 78 50 86 64 69 40 59 71 32 74 72 87 83 71 87 90 79 57 84 67 78 71 80 51 70 56 99 61 31 46 96 87 73 72 81 72 84 77 75 38 91 82 15 69 75 49 62 13 58 74 79 44 72 84 70 68 37 57 61 43 71 71 36 48 36 35 65 83 69 63 59 46 79 58 82 81 68 50 88 35 55 80 71 59 76 87 71 50 65 76 29 37 68 40 72 47 39 84 58 49 43 83 55 44 73 54 53 56 54 59 79 61 98 69 84 82 74 59 85 64 70 85 78 84 78 63 59 85 57 25 80 69 63 45 84 87 97 98 86 100 100 79 56 91 69 78 72 71 77

R Code

R Example (Data Loading and Frequency Distribution Table)

```
## 讀取資料
dat = read.csv(’2018Midterm1.csv.csv’, header=TRUE)
Midterm = dat$Midterm

## 建構次數分配表
breaks = seq(0, 100, by=5)
Midterm.cut = cut(Midterm, breaks, right=FALSE)
Midterm.freq = table(Midterm.cut)
Midterm.freq
```

Frequency Distribution Table

```
Midterm.cut
   [0,5)   [5,10)  [10,15)  [15,20)  [20,25)  [25,30)  [30,35)
       0        1        1        1        0        2        2
 [35,40)  [40,45)  [45,50)  [50,55)  [55,60)  [60,65)  [65,70)
       8        6        7        7       17       11        16
[70,75)  [75,80)  [80,85)  [85,90)  [90,95) [95,100)
     27       17       18       13        4        6
>
```

Empirical Density Function

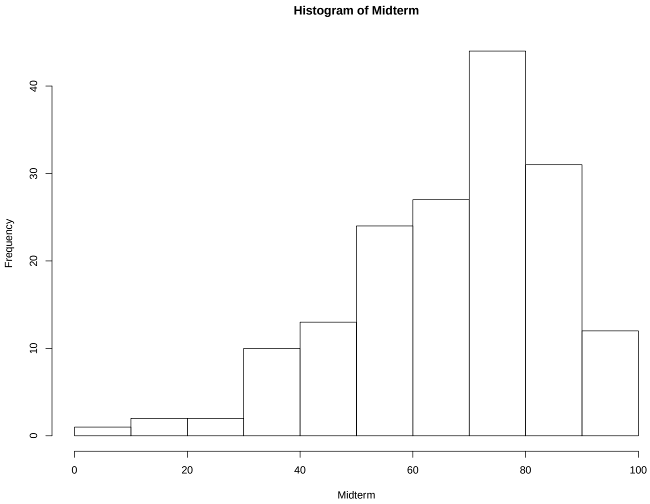
- Empirical Density Function
  - Histogram
  - Relative frequency distribution

R Code

R Example (Histogram)

```
## 繪製直方圖
hist(Midterm, breaks=10, right=FALSE, xlab=’Midterm’,
main=’Histogram of Midterm’)
```

Empirical Density Function



Empirical Distribution Function

Definition (Empirical Distribution Function)

Given random sample $\{X_i\}_{i=1}^n \sim^{i.i.d.} F_X(x)$, the empirical distribution function (EDF) is defined as

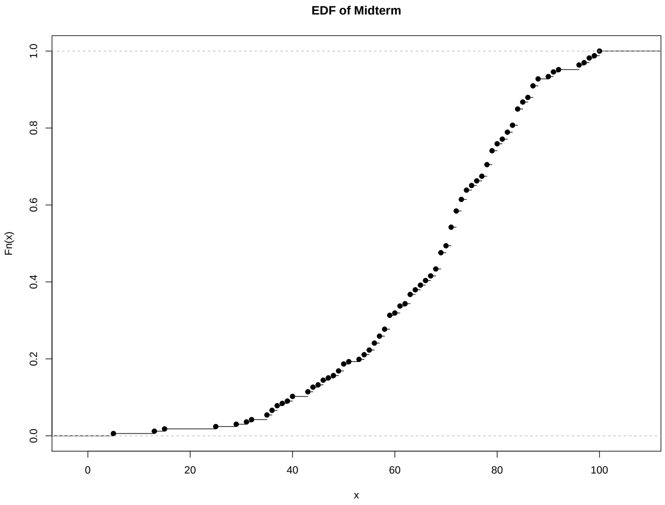
$$\hat{F}_n(x) = \frac{\text{number of elements in the sample} \leq x}{n} = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \leq x\}}$$

R Code

R Example (Empirical Density Function)

```
## 繪製 EDF
medf <- ecdf(Midterm)
plot(medf, main=’EDF of Midterm’)
```

Empirical Distribution Function



Section 2

Statistics

Statistics

Definition (Statistic)

Any function of the random sample is called a statistic:

$$T_n = T(X_1, X_2, \ldots, X_n).$$

- A statistic does not contain unknown parameters.
- The subscript n indicates the sample size.

Examples of Statistics

- Sample mean:

$$\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$$

- Sample variance:

$$S_n^2 = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{n-1}$$

- Sample r-th moments:

$$m_r = \frac{1}{n} \sum_{i=1}^n X_i^r$$

- Sample covariance/correlation coefficient:

$$S_{XY} = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{n-1}, \qquad r_{XY} = \frac{S_{XY}}{S_X S_Y}$$

Sampling Distributions

Definition (Sampling Distribution)

Let random variable $T_n = T(X_1, X_2, \ldots, X_n)$ be a function of random sample, then the distribution of T_n is called the sampling distribution.

Example 1

- If $\{X_i\}_{i=1}^n$ is a random sample from Bernoulli(p), then

$$T_n = \sum_{i=1}^n X_i \sim \text{Binomial}(n, p).$$

That is, Binomial distribution is the sampling distribution of T_n, which is a function of the Bernoulli random sample,$\{X_i\}_{i=1}^n$.

Example 2

- If $\{X_i\}_{i=1}^n$ is a random sample from $N(\mu, \sigma^2)$, then

$$T_n = \sum_{i=1}^n X_i \sim N\left(n\mu, n\sigma^2\right).$$

$$T_n = \underbrace{\frac{1}{n}\sum_{i=1}^n X_i}_{\bar{X}_n} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Example 3

- Let $\{X_i\} \sim^{i.i.d.} N(\mu, \sigma^2)$,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

- Then it can be show that $\bar{X}_n \perp S_n^2$, and

$$\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1), \qquad \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1), \qquad \frac{\bar{X}_n - \mu}{\frac{S_n}{\sqrt{n}}} \sim t(n-1)$$

Example 3

Theorem (Daly's Theorem)

Let $\{X_i\}_{i=1}^n \sim^{i.i.d.} N(\mu, \sigma^2)$, *and* $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. *Suppose that* $g(X_1, X_2, \ldots, X_n)$ *is translation invariant, that is,* $g(X_1 + c, X_2 + c, \ldots, X_n + c) = g(X_1, X_2, \ldots, X_n)$ *for all constant* c. *Then* $\bar{X}_n$ *and* $g(X_1, X_2, \ldots, X_n)$ *are independent.*

- Proof: omitted here.

Section 4

Biased Samples

Biased Samples

- Ideally, we would like our data to be a random sample from the target population. In practice, samples can be tainted by a variety of biases.
- Two typical biases:
  - Selection bias
  - Survivor bias
- Reading: Gary Smith (2014) 'Garbage In, Gospel Out' in *Standard Deviations: Flawed Assumptions, Tortured Data, and Other Ways to Lie with Statistics*. Overlook Press.

Selection Bias

Definition

Selection bias occurs when the results are distorted because the sample systematically excludes or under-represents some elements of the population.

- This particular kind of selection bias is also known as self-selection bias because people choose to be in the sample.
- We should be careful making comparisons to people who made different choices.

Self-Selection Bias

Example 1

- Scott Geller, a psychology professor at Virginia Tech, studied drinking in three bars near campus. He found that a drinker consumes more than twice as much beer if it comes in a pitcher than in a glass or bottle.

- Hence, he argues that banning pitchers in bars could make a dent in the drunken driving problem.

Self-Selection Bias

Example 2

- A study found that Harvard freshmen who had not taken SAT preparation courses scored an average of 63 points higher on the SAT than did Harvard freshmen who had taken such courses.
- Harvard's admissions director said that this study suggested that SAT preparation courses are ineffective.

Survivor Bias

Definition

Survivor bias is that when we choose a sample from a current population to draw inferences about a past population, we leave out members of the past population who are not in the current population: We look at only the survivors.

- Prospective study vs. Retrospective study

Survivor Bias

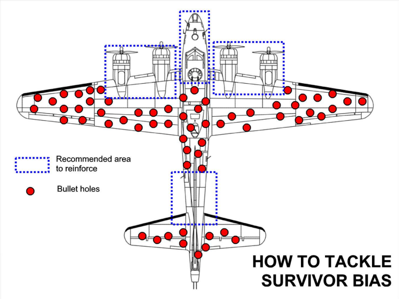
Example 1: Which Places Need Protection?

- In World War II, the British Royal Air Force (RAF) planned to attach heavy plating to its airplanes to protect them from German fighter planes and land-based antiaircraft guns. The protective plates weighed too much to cover an entire plane, so the RAF collected data on the location of bullet and shrapnel holes on planes that returned from bombing runs.

Survivor Bias

Example 1: Which Places Need Protection?

- Most holes on the wings and rear of the plane, and very few on the cockpit, engines, or fuel tanks



- Conclusion: the protective plates should be put on the wings and rear. Do you agree?

Survivor Bias

Example 1: Which Places Need Protection?

- Abraham Wald had recognized that these data suffered from survivor bias.
- During World War II, Wald was a member of the Statistical Research Group (SRG) at Columbia University, where he applied his statistical skills to various wartime problems.

Survivor Bias

Example 2: Success Secrets

- In writing his bestselling book Good to Great, Jim Collins and his research team spent five years looking at the forty-year history of 1,435 companies and identified 11 stocks that clobbered the average stock.
- After scrutinizing these eleven great companies, Collins identified several common characteristics and attached catchy names to each, like Level 5 Leadership – leaders who are personally humble, but professionally driven to make their company great.
- The problem, of course, is that this is a backward-looking study undermined by survivor bias.