Intermediate Probability and Statistics

X. Joan Hu

Department of Statistics and Actuarial Science
Simon Fraser University

Spring 2023

What to do today (Friday Jan 13, 2023)?

§1.1 Introduction

§1.2 Review 1 on Chp 1-5: Basic Concepts

§1.3 Review 2 on Chp 1-5: Sampling Distributions

§2.1 Point Estimation

§2.1.1 Some General Concepts

§1.3 Review 2: Sampling Distributions

Statistics and their distributions:

- **statistic**: a function of r.v.s. Its distribution is obtained case by case.

Example 3: Consider **iid sample** of $X \sim F_X(\cdot)$ with mean μ and variance σ^2.

$$X_1, \ldots, X_n \text{ are independent with each other and } \sim F_X(\cdot)$$

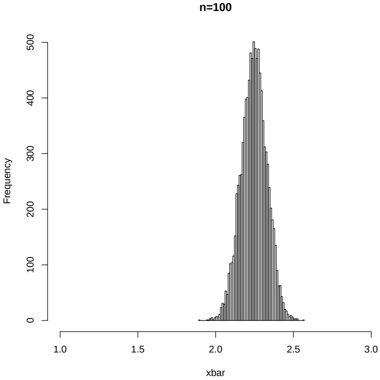
The sample mean $\bar{X} = (X_1 + \ldots + X_n)/n$'s distribution?

- $E(\bar{X}) = \mu$ and $V(\bar{X}) = \sigma^2/n$.
- If $X \sim N(\mu, \sigma^2)$, $\bar{X} \sim N(\mu, \sigma^2/n)$

Central Limit Theorem. Provided that $X_1, \ldots, X_n$ are iid with mean μ and variance σ^2.

- The distribution of $X_1 + \ldots + X_n$ is approximately $N(n\mu, n\sigma^2)$, if $n >> 1$.
- The distribution of $\bar{X}$ is approximately $N(\mu, \sigma^2/n)$, if $n >> 1$.

To motivate the CLT, let's consider the sample mean $\bar{X}$ of a random sample $\{X_1, \cdots, X_n\}$ from the distn given in the table: $E(X) = 9/4$, $V(X) = 11/16$ *What is the distn of $\bar{X}$ when $n = 100$?* The histogram of $\bar{X}$ based on $m = 10^5$ repetitions: Almost $N(9/4, 11/1600)$!



Review2: Sampling Distributions (cont'd)

- **Normal Distribution**: $X \sim N(\mu, \sigma^2)$
  - to calculate $P(a < X < b)$ with any given a, b?

to standardize r.v. X:

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

Denote $P(Z \leq z)$ by $\Phi(z)$.

$$\begin{aligned} P(a < X < b) &= P(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}) \\ &= \Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma}) \end{aligned}$$

How to obtain the values of $\Phi(\cdot)$?

- *The standard normal distribution table:* Table A.3 Standard Normal Curve Areas
- *Alternatively, using* `R` *function*:
  `pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)`

Review2: Sampling Distributions (cont'd)

- **Some distributions derived from $N(\mu, \sigma^2)$**

Chi-Square Distribution. Suppose $Z_1, \ldots, Z_K$ are i.i.d. with $N(0,1)$. Let W be $W = Z_1^2 + \ldots + Z_K^2$. The distribution of W is the chi-square distribution with the degrees of freedom (df) K, denoted by $W \sim \chi^2(K)$.

Properties:

(i) $\mathrm{E}(W) = K$.

(ii) $\mathrm{V}(W) = 2K$.

(iii) If $W_1 \sim \chi^2(K_1)$, $W_2 \sim \chi^2(K_2)$ and W_1 and W_2 are independent, then $W_1 + W_2 \sim \chi^2(K_1 + K_2)$. (*why?*)

How to obtain relevant values of $\chi^2(\cdot)$?

- *χ^2-distribution table:* Table A.7 Critical Values of Chi-Square Distribution
- *Alternatively, using* `R` *function*:

```
pchisq(q, df, ncp = 0, lower.tail = TRUE, log.p =
FALSE)
```

Student's t-Distribution. Suppose $Z \sim N(0,1)$ and $W \sim \chi^2(K)$, and $Z \perp W$. Let T be

$$T = \frac{Z}{\sqrt{W/K}}.$$

The distribution of T is the t-distribution with K degrees of freedom (df): $T \sim t(K)$. *It was initially derived by Gosset (1908).*

Properties:

(i) $\mathrm{E}(T) = 0$.

(ii) $\mathrm{V}(T) = K\big/(K-2)$, if $K > 2$.

(iii) If $T \sim t(K)$ with $K >> 1$, T's distribution is approximately $N(0,1)$. That is $t(\infty) = N(0,1)$. (*why?*)

How to obtain relevant values of $t(\cdot)$?

- *Student's t-distribution table.* Table A.5 Critical Values for t-Distributions
- *Alternatively, using* `R` *function*:
  `pt(q, df, ncp, lower.tail = TRUE, log.p = FALSE)`

§2.1 Point Estimation (Chp6)

§2.1.1 Some General Concepts

2.1.1A. What does point estimation do?

Suppose r.v. $X \sim F(\cdot;\theta)$ (**population**) with unknown θ (**parameter**).

- Use the available information (data, a sample from the population) to compute a 'good guess' (**point estimate**) for the true value of θ
  - The formula used to obtain a point estimate is called the **point estimator** of θ, denoted by $\hat{\theta}$.
  - A point estimator is a suitable statistic: it is often referred to as a realization or an evaluation of the corresponding point estimator.

Example 2.1 (Devore 9th: p249) An automobile manufacturer has developed a new type of bumper. The manufacturer has used this bumper in a sequence of 25 controlled crashes against a wall at 10 mph, using one of its compact car models. The parameter to be estimated is p, the proportion of all such crashes that result in no damage:

$$p = P(\text{no damage in a single crash}).$$

Let X the number of crashes that result in no visible damage: X observed to be $x = 15$.

$$\text{estimator } \hat{p} = \frac{X}{n}; \qquad \text{estimate } \hat{p}_{obs} = \frac{x}{n} = \frac{15}{25} = 0.60.$$

Why to use X/n as $\hat{p}$? $X \sim B(25, p)$, so $E(X) = 25p$ and is approximated by the observed $x = 15$

an alternative solution:
Suppose iid $Y_1, \ldots, Y_{25}$ drawn from population $Y \sim B(1, p)$, $Y = 1$ if no crash and $= 0$ if crashed.

Note $\sum_{i=1}^{25} Y_i = X$:

$$E(Y) = p = P(Y = 1)$$

$$\hat{p} = \bar{Y} = \frac{1}{25}\big(Y_1 + \ldots + Y_{25}\big) = \frac{X}{25}$$

Note X is observed as 15. Thus,

$$\hat{p}_{obs} = 15/25$$

What will we do next?

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

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

2.1 Point Estimation

2.2 Confidence Interval

2.3 One-Sample Test

2.4 Inference Based on Two-Samples

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

3.1 One-Factor Analysis of Variance

3.2 Multi-Factor ANOVA

3.3 Simple Linear Regression Analysis

3.4 More on Regression

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

Remarks:

- *Homework 1 is due on Monday Jan 16 by 5:00pm: please submit it via the course canvas page.*
- *Classroom for Tue lecture 10:30-12:20 is now AQ5006; for Mon 1st Tutorial D101, BLU10901.*