Intermediate Probability and Statistics

X. Joan Hu

Department of Statistics and Actuarial Science
Simon Fraser University

Spring 2023

What to do today (Tuesday Jan 10, 2022)?

§1.1 Introduction

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

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

Review 1A: Probability and Statistics

- **Probability**
  - definitions and properties
    - classical, frequentist, Bayesian
    - the three axioms in Kolmogorov definition and extensions
  - conditional probability and independence

$$P(A \mid B) = \frac{P(AB)}{P(B)}$$

$$A \perp B \quad \text{iff} \quad P(AB) = P(A)P(B)$$

Review 1A: Probability and Statistics (cont'd)

- **Statistics**
  - population vs sample
    - population: the target to make inference on
    - sample: subset of population with available information to be used to make inference
    - random sample: a set of independent identically distributed observations from the population
  - descriptive statistics: functions/tables/plots of data
    - sample mean and variance, etc
    - contingency table
    - histogram, boxplot, scatterplot, etc

Review 1B: Random Variable and Probability Distribution

- random variable:

$$X : S \to (-\infty, \infty)$$

- discrete distributions: (*probability mass function*)

  For example,

  - discrete uniform: $P(X = a_k) = 1/K, \quad a_1, \ldots, a_K.$
  - binomial: $Bin(n, p)$
  - Poisson: $Poisson(\lambda)$

- continuous distributions: (*probability density function*)

  For example,

  - continuous uniform: $Unif(a, b)$
  - normal: $N(\mu, \sigma^2)$
  - exponential: $NE(\lambda)$

Review 1B: Random Variable and Probability Distribution (cont'd)

- cumulative distribution:

  $$F(x) = P(X \leq x),$$

  valued in [0, 1] with $F(-\infty) = 0$ and $F(\infty) = 1$.

- joint distribution

  $$F(x, y) = P(X \leq x, Y \leq y).$$

  - it's $F_X(x)F_Y(y)$ iff $X \perp Y$.
  - relation to pmf, pdf

Review1C: Expectation, Variance, Covariance and Correlation

"Did you hear about the politician who promised that, if he was elected, he'd make certain that everybody would get an above average income?"

- **Expectation**
  - definition
    - discrete r.v. X: $E(X) = \sum_{\text{all } x} xp(x)$
    - continuous r.v. X: $E(X) = \int_{-\infty}^{\infty} xf(x)dx$

    In general, a r.v. X: $E(X) = \int_{-\infty}^{\infty} xdF(x)$
  - properties:

$$E(aX + bY) = aE(X) + bE(Y)$$

Review1C: Expectation, Variance, Covariance and Correlation (cont'd)

- **Variance**: $V(X) = E(X - EX)^2$
  - $V(X) = E(X^2) - (EX)^2$
  - $V(aX + b) = a^2 V(X)$
- **Covariance**: $Cov(X, Y) = E(X - EX)(Y - EY)$

$$Cov(X, Y) = E(XY) - (EX)(EY)$$
$$Cov(aX + b, cY + d) = acCov(X, Y)$$

$\Longrightarrow$

$$Var(aX + bY) = a^2 V(X) + b^2 V(Y) + 2abCov(X, Y)$$

Review1C: Expectation, Variance, Covariance and Correlation (cont'd)

- **Correlation Coefficient:**

$$corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{V(X)V(Y)}}$$

$-1 \leq corr(X, Y) \leq 1$

Remark: Population Quantity vs Sample Quantity

- expectation (population mean) vs sample mean
- population variance vs sample variance
- population covariance vs sample covariance

§1.3 Review 2: Sampling Distributions

Statistics and their distributions:

- **statistic**: a function of r.v.s.
  How about its distribution?
  to obtain it case by case.

Example 1: Consider r.v. $X \sim F_X(\cdot)$, $Y = g(X)$'s distribution?

eg, $Y = 1/X$ with $X > 0$: if $y > 0$,

$$P(Y \leq y) = P(X \geq 1/y) = 1 - F_X(1/y)$$

Example 2: Consider r.v.s. X_1 and X_2, $Y = h(X_1, X_2)$'s distribution?

eg, if $X_1 \perp X_2$, $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$:

$$Y = aX_1 + bX_2 : \quad Y \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2).$$

Review2: Sampling Distributions (cont'd)

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 $X \sim N(\mu, \sigma^2)$, $\bar{X} \sim N(\mu, \sigma^2/n)$

What if $X \nsim N(\mu, \sigma^2)$?

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: $E(X) = 9/4$, $V(X) = 11/16$

x	1	2	3
$p_X(.)$	1/4	1/4	1/2

n=2

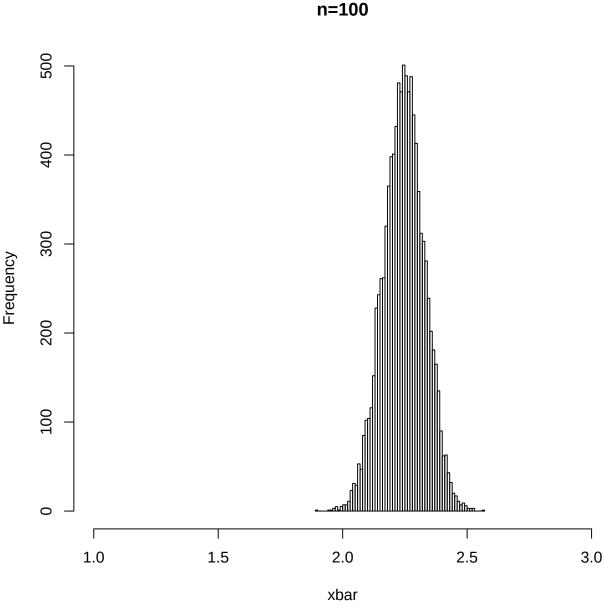
$\bar{x}$	2/2	3/2	4/2	5/2	6/2
$p_{\bar{X}}(.)$	1/16	2/16	5/16	4/16	4/16

n=3

$\bar{x}$	3/3	4/3	5/3	6/3	7/3	8/3	9/3
$p_{\bar{X}}(.)$	1/64	3/64	9/64	13/64	18/64	12/64	8/64

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)`

"*Statistics is the science of learning from data.*"

- By processing/summarizing the data: tabulating/plotting
- By making inferences with the data, to understand uncertainties using the limited information

Statistical Thinking ("The Basic Practice of Statistics", 6th Edn, by Moore et al.)

- Data are numbers with a context.
- Where the data come from matters.
- Always look at the data.
- Beware the lurking variable.
- Variation is everywhere.
- Conclusions are not certain.

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)

Homework 1 is due on Monday 16 by 5:00pm.