Chapter 7: Sampling Distribution

Lianming Wang

University of South Carolina

January 4, 2023

Section 7.1-7.2: Point estimation and sampling distribution

- A parameter is a numerical descriptive measure of a population.
- A sample statistic is a numerical descriptive measure of a sample. It is calculated from the observations in the sample.
- Sample statistics themselves are random variables. The **sampling distribution** of a sample statistic is the probability distributions of the statistic.
- Note: it is assumed that the sample is a random sample, which is representative of the population.

Example:

Consider two independent observations (X_1, X_2) from a Binomial distribution $B(n = 2, p = 0.5)$.

x	0	1	2
$p(x)$	.25	.5	.25

- Define $Y = \frac{(X_1+X_2)}{2}$. Find the sampling distribution of Y.
- Define $Z = \max(X_1, X_2)$. Find the sampling distribution of Z.

List all the possible values of (X_1, X_2).

(x_1, x_2)	Probability	$y = \frac{(x_1+x_2)}{2}$	$z = \max(x_1, x_2)$
(0,0)	.25*.25	0	0
(0,1)	.25*.5	.5	1
(0,2)	.25*.25	1	2
(1,0)	.5*.25	.5	1
(1,1)	.5*.5	1	1
(1,2)	.5*.25	1.5	2
(2,0)	.25*.25	1	2
(2,1)	.25*.5	1.5	2
(2,2)	.25*.25	2	2

Example Continued.

- The sampling distribution of $Y = \frac{(X_1+X_2)}{2}$

y	0	.5	1	1.5	2
$p(y)$	.0625	.25	.375	.25	.0625

- The sampling distribution of $Z = \max(X_1, X_2)$

z	0	1	2
$p(z)$	.0625	.5	.4375

- The expected value of Z is
  $\mu_z = \sum zp(z) = 0 * 0.0625 + 1 * .5 + 2 * .4375 = 1.375$
- The variance of Z is
  $\sigma_z^2 = \sum z^2 p(z) - \mu_z^2 =$
  $[0^2 * 0.0625 + 1^2 * .5 + 2^2 * .4375] - 1.375^2 = 0.3594$

Point estimator

- Suppose $X_1, \cdots, X_n$ are a random sample. There are two implications:
  - $X_1, \cdots, X_n$ are independent of each other.
  - $X_1, \cdots, X_n$ have the same distribution.
- Statistic $\hat{\theta} = \hat{\theta}(X_1, \cdots, X_n)$ is a point estimator of θ.
- Commonly seen examples:
  - We can use sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ to estimate the population mean μ.
  - We can use sample variance $s^2 = \frac{1}{n-1}[\sum_{i=1}^{n} X_i^2 - n(\bar{X})^2]$ to estimate the population variance σ^2.
  - We can use sample proprotion $\hat{p}$ to estimate the population proportion p.

Theorem: If $X_1, X_2, \cdots, X_n$ are iid from $N(\mu, \sigma^2)$, then $\overline{X} = \frac{\sum_i X_i}{n} \sim N(\mu, \sigma^2/n)$ for any positive integer n.

This theorem implies the following properties

- The sampling distribution of $\overline{X}$ is a normal distribution.
- The expected value of $\overline{X}$ is μ. That is $\mu_{\overline{X}} = \mu$.
- The variance of $\overline{X}$ is $\frac{\sigma^2}{n}$. That is, $\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$.

Section 7.3: Reporting point estimators

Unbiased estimator

- Suppose statistic $\hat{\theta}$ is an estimator of a population parameter θ. $\hat{\theta}$ is said to be unbiased for θ if $E(\hat{\theta}) = \theta$.
- Example: suppose we have an independently and identically distributed (iid) sample $(X_1, X_2, \cdots, X_n)$ from $N(\mu, 1)$, i.e., $X_i \sim N(\mu, 1)$.
  - We can use $\overline{X} = \frac{1}{n}\sum_i X_i$ to estimate μ, and we have $E(\overline{X}) = \mu$.
  - We can just use the first observation as an estimator, i.e., X_1. It is also an unbiased estimator.

How do we evaluate or compare different unbiased estiamtors?

- The standard error of an estimator $\hat{\theta}$ is its standard deviation, i.e., $\sigma_{\hat{\theta}} = \sqrt{var(\hat{\theta})}$.
- If the standard error involves unknown parameters that can be estimated, substitution of those values into $\sigma_{\hat{\theta}}$ produces an estimated standard error, denoted by $\hat{\sigma}_{\hat{\theta}}$.
- When reporting a point estimate from a sample, we usually also need to report its (estimated) standard error.
- When multiple unbiased estimators are available, the best one is the one with smallest standard error (or variance).

The Central Limit Theorem:

A random sample of n observations $(X_1, X_2, \cdots, X_n)$ is selected from a population (any population) with mean μ and variance σ^2. Then when sample size n is sufficiently large ($n \geq 30$), the sampling distribution of $\overline{X}$ will be approximately a normal distribution with mean μ and variance σ^2/n.

Exercise 1.

The lifetime of batteries produced in a local plant is an approximately normal random variable with mean equal to 10 hours and standard deviation equal to 2 hours.

a. If one battery is randomly selected from the plant, what is the probability that the battery will last longer than 12 hours?

b. If two batteries are randomly selected from the plant, what is the probability that their average lifetimes is over 12 hours?

c. If two batteries are randomly selected from the plant, what is the probability that at least one of them last longer than 12 hours?

Exercise 2.

Suppose the mean and the standard deviation of the amount of soda in a large drink is 27.5oz and 1.5oz, respectively. If a random sample of 30 drinks is collected, what is the probability that

a. they will average between 27 and 28 oz?

b. the average amount of these 30 drinks is over 28 oz?

Exercise 3.

Unaltered bitumen, as commonly found in lead-zine deposits, have atomic hydrogen/carbon (H/C) ratios that average 1.4 with standard deviation 0.25. Find the probability that the average H/C ratio is less than 1.3 if we randomly select 35 bitumen samples.

Approximation of a Binomial distribution with a normal distribution

- Use tables for Binomial probabilities when n is small.
- No Binomial tables available for large n, e.g., $n > 30$.
- We can use normal approximation:
  - Suppose $X \sim B(n, p)$ for some large n. Then $\mu = np$ and $\sigma = \sqrt{np(1-p)}$.
  - $P(X \leq a) \approx P(z \leq \frac{a+0.5-\mu}{\sigma})$, where a is an integer between 0 and n.
  - Adding 0.5 in the above normal approximation is called correction for continuity.

Exercise 4.

Accordingn to recent studies, 1% of all patients who undergo laser surgery (i.e., LASIK) to correct their vision have serious post-laser problems. In a random sample of 100,000 LASIK patients, let X be the number who experience serious post-laser vision problems.

a. What is the exact distribution of X?

b. Find the expected value μ and standard deviation σ of X.

c. Find the probability that fewer than 950 patients in a sample of 100,000 will experience serious post-laser vision problems.