Chapter 7: Sampling Distributions

Yu Yang

School of Statistics
University of Minnesota

October 3, 2022

Recall From Chapter 1

- parameter: a number that describes a population
- statistic: a number that describes a sample
- inference: drawing conclusions about a population based on information from a sample.

Understand Sampling Distribution

- We use a sample statistic to estimate the unknown parameter, and we use sampling distribution to characterize the sample statistic.
- The value of a statistic will vary from sample to sample. Therefore, statistics have their own distributions.

Sampling Distribution

The sampling distribution of a statistic is the probability distribution that describes the possible values the statistic can take and assigns probabilities for those values

Oct 3 Lecture Stopped Here

Example 7.2

A roulette wheel in Las Vegas has 38 slots. 18 of 38 numbers on a roulette wheel are red. Adam decides to play roulette. He bet his money on red. The probability distribution of one single output is:

Result	Win (W)	Lose (L)
Probability	9/19	10/19

We can see that the population proportion is $p = 9/19$.

(a) Adam will play two games. List all possible outcomes and for each outcome calculate $\hat{p}$, the proportion of winnings.

Example 7.2

A roulette wheel in Las Vegas has 38 slots. 18 of 38 numbers on a roulette wheel are red. Adam decides to play roulette. He bet his money on red. The probability distribution of one single output is:

Result	Win (W)	Lose (L)
Probability	9/19	10/19

We can see that the population proportion is $p = 9/19$.

(a) Adam will play two games. List all possible outcomes and for each outcome calculate $\hat{p}$, the proportion of winnings.
Possible samples: {LL, LW, WL, WW}
LL: $\hat{p} = 0$; WW: $\hat{p} = 2/2 = 1$;
WL: $\hat{p} = 1/2$; LW: $\hat{p} = 1/2$

Example 7.2 (Continued)

(b) Use your answer to (a) to find the sampling distribution for the sample proportion, $\hat{p}$.

Sampling Distribution:

$\hat{p}$	0	1/2	1
probability	100/361	180/361	81/361

$P(\hat{p} = 0) = P(LL) = 10/19 \times 10/19 = 100/361$
$P(\hat{p} = 1/2) = P(WL \text{ or } LW) = 9/19 \times 10/19 + 10/19 \times 9/19 = 180/361$
$P(\hat{p} = 1) = P(WW) = 9/19 \times 9/19 = 81/361$

Example 7.2 (Continued)

(c) Find the mean and standard deviation of this samping distribution.

Example 7.2 (Continued)

(c) Find the mean and standard deviation of this samping distribution.

$$\mu_{\hat{p}} = 0\left(\frac{100}{361}\right) + \left(\frac{1}{2}\right)\left(\frac{180}{361}\right) + 1\left(\frac{81}{361}\right)$$

$$= \frac{9}{19}$$

$$\sigma^2_{\hat{p}} = \left(0 - \frac{9}{19}\right)^2\left(\frac{100}{361}\right) + \left(\frac{1}{2} - \frac{9}{19}\right)^2\left(\frac{180}{361}\right) + \left(1 - \frac{9}{19}\right)^2\left(\frac{81}{361}\right)$$

$$= \frac{45}{361}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{45}{361}}$$

Although a sampling distribution can be constructed for **any** statistic, we will focus on the sampling distributions for the sample mean and sample proportion.

Recall

- A proportion is a value between 0 and 1. When multiplied by 100, it can be interpreted as a percentage.

  Example: The proportion of students that are late, that get A's etc.

- A sample mean is the average of a set of data

  Example: The mean number of hours per week spent on studying, the mean amount of money per week spent on food, etc.

The Sampling Distribution of a Sample Mean

Notation:

$$\mu = \text{population mean}$$
$$\sigma = \text{population standard deviation}$$
$$\bar{x} = \text{sample mean}$$

Statistical inference: Use $\bar{x}$ to estimate μ.

EXAMPLE 7.3

Suppose we want to determine the average income of the students in this class but do not have time to ask every single person. On average, would we rather make our estimate based on a sample of 2 students or a sample of 10 students?

The Sampling Distribution of a Sample Mean

Notation:

μ = population mean

σ = population standard deviation

$\bar{x}$ = sample mean

Statistical inference: Use $\bar{x}$ to estimate μ.

EXAMPLE 7.3

Suppose we want to determine the average income of the students in this class but do not have time to ask every single person. On average, would we rather make our estimate based on a sample of 2 students or a sample of 10 students?

Unless we take a very unrepresentative sample, the average income of 10 students will better approximate the overall average income than will the average income of just 2 students.

The Law of Large Numbers (LLN)

Theorem: The Law of Large Numbers
Let $x_1, x_2, \ldots, x_n$ be independent observations from **any** population with mean μ. Then, as the sample size n increases.

$$\bar{x} \to \mu$$

Interpretation: The larger the sample, the better $\bar{x}$ approximates μ.

NOTE: The LLN only tells us that $\bar{x}$ gets closer to μ as the sample size increases. In order to understand the variability in $\bar{x}$ from sample to sample we need to study its sampling distribution.

Mean and Standard Deviation of the Sampling Distribution of $\bar{x}$

Suppose $x_1, x_2, \ldots, x_n$ is a random sample from **any** population with mean μ and standard deviation σ.

Proposition:
The mean of the sampling distribution of $\bar{x} = \mu_{\bar{x}} = \mu$

Interpretation: If we take a LOT of samples and calculate $\bar{x}$ for each sample, the average of the $\bar{x}$'s will be close to μ.

What do I mean?

Suppose I want to know the *true* mean height of all North American women. I get a sample of heights of women from STAT 3011:

$$62, 66, 69, 61, 65: \ \bar{x} = 64.6$$

We went out and get more samples, say from other statistics courses:

$$61, 66, 65, 68, 72: \ \bar{x} = 66.4$$
$$59, 63, 68, 67, 69: \ \bar{x} = 65.2$$
$$58, 62, 63, 67, 68: \ \bar{x} = 63.6$$
$$62, 65, 64, 68, 68: \ \bar{x} = 65.4$$

The mean also has a distribution! We call that the sampling distribution!
(Mean of 64.6, 66.4, 65.2, 63.6, 65.4) = $65.04 \approx \mu$ if sample size was large (here it is only 5)

Notes:

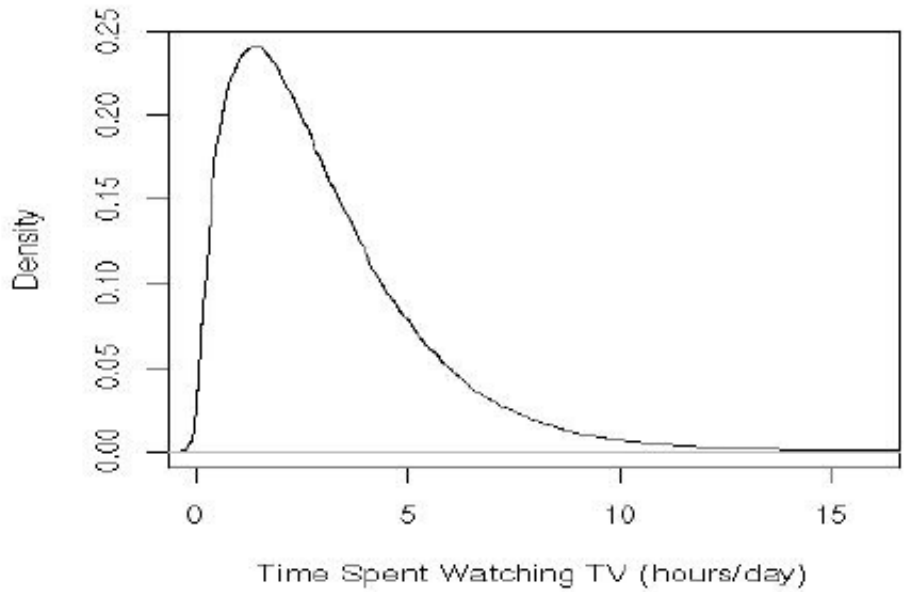
1. $\mu_{\bar{x}} = \mu \Rightarrow$ sample averages have the same mean as individual observations

2. $\sigma_{\bar{x}} = \sigma/\sqrt{n} \Rightarrow$ sample averages are less variable than individual observations.

3. $\sigma_{\bar{x}} = \sigma/\sqrt{n} \Rightarrow$ as sample size n increases, the variability of $\bar{x}$ from sample to sample decreases.

Example 7.4

Let X = number of hours an American watches TV on an average day and suppose it is known that X has a mean of 3 with a standard deviation of 2.25. The following is a density curve for random variable X:



Example 7.4

This means: $\mu = 3$ and $\sigma = 2.25$

(a) Take a sample of 10 Americans. What are the mean and standard deviation *of the sampling distribution of* $\bar{X}$, the average number of hours per day spent watching TV for these 10 people.

Example 7.4

This means: $\mu = 3$ and $\sigma = 2.25$

(a) Take a sample of 10 Americans. What are the mean and standard deviation *of the sampling distribution of* $\bar{X}$, the average number of hours per day spent watching TV for these 10 people.

$$\mu_{\bar{x}} = \mu = 3$$
$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 2.25/\sqrt{10} = 0.71$$

(b) What if we take a sample of 100 people?

Example 7.4

This means: $\mu = 3$ and $\sigma = 2.25$

(a) Take a sample of 10 Americans. What are the mean and standard deviation *of the sampling distribution of* $\bar{X}$, the average number of hours per day spent watching TV for these 10 people.

$$\mu_{\bar{x}} = \mu = 3$$
$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 2.25/\sqrt{10} = 0.71$$

(b) What if we take a sample of 100 people?

$$\mu_{\bar{x}} = \mu = 3$$
$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 2.25/\sqrt{100} = 0.225$$

$\bar{x}$ is less variable than when we $n = 100$ than when $n = 10$.

Example 7.4

So as the size of each sample increases, the standard deviation of the sampling distribution decreases.

(c) How large of a sample would we need in order to obtain an estimate, $\bar{x}$, with a standard deviation less than or equal to .05?

Example 7.4

So as the size of each sample increases, the standard deviation of the sampling distribution decreases.

(c) How large of a sample would we need in order to obtain an estimate, $\bar{x}$, with a standard deviation less than or equal to .05? Want

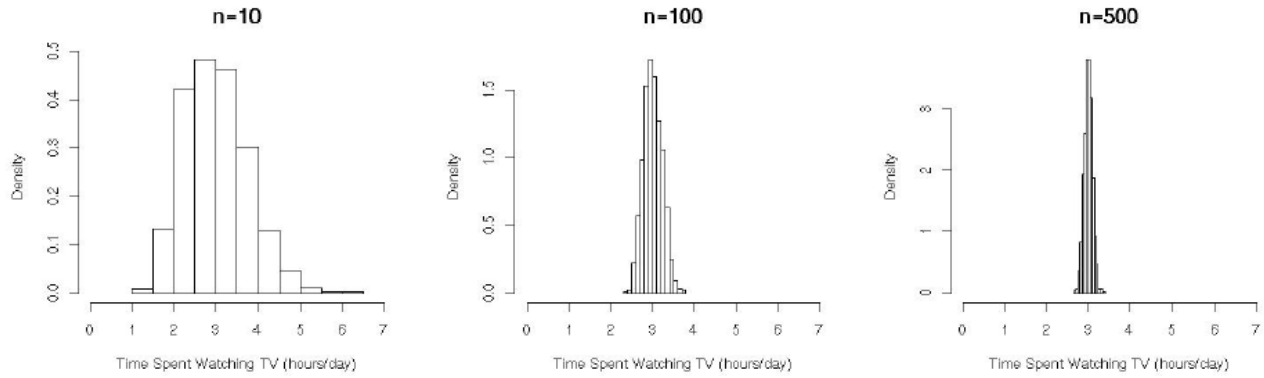
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.25}{\sqrt{n}} \leq .05$$

$$\frac{2.25}{\sqrt{n}} \leq .05 \Rightarrow \sqrt{n} \geq \frac{2.25}{.05} = 45$$

$$\Rightarrow n \geq 45^2 = 2025$$

Shape of the sampling distribution

For the above example, we plot histograms for 1000 samples, each of size: $n = 10$, $n = 100$, and $n = 500$:



Observations:

1. the sampling distributions have different shapes than the population distribution (which we saw a few slides ago)
2. bell-shaped, symmetric (appear normal)
3. centered around the true mean (3)
4. become much less variable as the sample size increases.

Sampling Distribution of $\bar{x}$

1. **Normal Distribution**
   If $x_1, x_2, \ldots x_n$ is a random sample from $N(\mu, \sigma)$, then

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

2. **General: Central Limit Theorem (CLT)**
   Suppose $x_1, x_2, \ldots x_n$ is a random sample from **any** population with mean μ and standard deviation σ. Then, if n is large enough (rule of thumb: $n \geq 30$)

$$\bar{x} \dot{\sim} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

NOTE: CLT holds for **any** distribution (even a discrete distribution)!

Oct 5 Lecture Stopped Here

Example 7.5: Normal Distribution Case

A machine fills empty glass bottles with soda. Let X = amount of soda poured into a bottle and assume X is normally distributed with a mean of 298mL and standard deviation of 3mL, i.e.

$$X \sim N(298, 3)$$

1. The label on the bottle says that it contains 295 mL of soda. Find the probability that a randomly selected bottle contains less soda than advertised.

Example 7.5: Normal Distribution Case

A machine fills empty glass bottles with soda. Let X = amount of soda poured into a bottle and assume X is normally distributed with a mean of 298mL and standard deviation of 3mL, i.e.

$$X \sim N(298, 3)$$

1. The label on the bottle says that it contains 295 mL of soda. Find the probability that a randomly selected bottle contains less soda than advertised.

$$\begin{aligned} P(X < 295) &= P\left(\frac{X - 298}{3} < \frac{295 - 298}{3}\right) \\ &= P(Z < -1) = pnorm(-1) \\ &= 0.1587 \end{aligned}$$

Example 7.5: Normal Distribution Case

2. Now, find the probability that the average amount of soda per bottle in a randomly chosen 6-pack is less than 295mL. (Hint: Think about the distribution of the sample mean.)

Example 7.5: Normal Distribution Case

2. Now, find the probability that the average amount of soda per bottle in a randomly chosen 6-pack is less than 295mL. (Hint: Think about the distribution of the sample mean.)

$X \sim N(298, 3) \Rightarrow \bar{X} \sim N\left(298, \frac{3}{\sqrt{6}}\right)$ since the mean is taken over 6 bottles.

$$\begin{aligned} P(\bar{X} < 295) &= P\left(\frac{\bar{X} - 298}{3/\sqrt{6}} < \frac{295 - 298}{3/\sqrt{6}}\right) \\ &= P(Z < -2.45) = pnorm(-2.45) \\ &= 0.0071 \end{aligned}$$

Interpretation: Only about 0.7 % of any sampled 6-pack will have an average amount less than 295mL/bottle, whereas almost 16% of individual bottles are underfilled.

Example 7.6: General Case (Applying the CLT)

Suppose the mean number of children per household in the U.S. is 1.90 with a standard deviation of 1.68. Let X = number of children in a household.

1. Does X have a normal distribution?

Example 7.6: General Case (Applying the CLT)

Suppose the mean number of children per household in the U.S. is 1.90 with a standard deviation of 1.68. Let X = number of children in a household.

1. Does X have a normal distribution?
   No, X is a discrete random variable, not continuous.

2. Let $\bar{X}$ = average number of children in a random sample of 50 households. Is $\bar{X}$ discrete or continuous? What is the sampling distribution of $\bar{x}$?

Example 7.6: General Case (Applying the CLT)

Suppose the mean number of children per household in the U.S. is 1.90 with a standard deviation of 1.68. Let X = number of children in a household.

1. Does X have a normal distribution?
   No, X is a discrete random variable, not continuous.

2. Let $\bar{X}$ = average number of children in a random sample of 50 households. Is $\bar{X}$ discrete or continuous? What is the sampling distribution of $\bar{x}$?

   $\bar{X}$ is a continuous random variable! Since $n > 30$, by the CLT:

$$\bar{X} \dot{\sim} N\left(1.90, \frac{1.68}{\sqrt{50}}\right) = N(1.90, .2376)$$

Example 7.6: General Case (Applying the CLT)

3. Sample 50 households. Find the probability that the average number of children per household in this sample is more than 5.

Example 7.6: General Case (Applying the CLT)

3. Sample 50 households. Find the probability that the average number of children per household in this sample is more than 5.

$$\begin{aligned} P(\bar{X} > 5) &= P\left(\frac{\bar{X} - 1.90}{.2376} > \frac{5 - 1.90}{.2376}\right) \\ &= P(Z > 13.05) \\ &\approx 0 \end{aligned}$$

It is highly unlikely that any sample of 50 households will have an average number of children that is bigger than 5.

The Sampling Distribution of a Sample Proportion

Notation:

$$p = \text{population proportion of ``success''}$$
$$\hat{p} = \text{sample proportion of ``success''}$$

Statistical Inference: use $\hat{p}$ to estimate p.

Example 7.7

Suppose we take a random sample of 60 U.S. households and 7 say they have been burglarized/robbed. Use this information to estimate p, the true proportion of U.S. households that have been burglarized.

Example 7.7

Suppose we take a random sample of 60 U.S. households and 7 say they have been burglarized/robbed. Use this information to estimate p, the true proportion of U.S. households that have been burglarized.

$$\hat{p} = 7/60 = .117.$$

Goal: $\hat{p}$ varies from sample to sample. Therefore, just as we did for $\bar{x}$, we want to describe the sampling distribution of $\hat{p}$.

Law of Large Numbers for p

The Law of Large Numbers guarantees that as sample size n increases

$$\hat{p} \to p$$

That is, the larger the sample size, the better $\hat{p}$ estimates p

Mean and Standard Deviation

Suppose $\hat{p}$ is the sample proportion for a sample of size n from a population with proportion p. Then the mean and standard deviation of the sampling distribution of $\hat{p}$ are:

mean $= \mu_{\hat{p}} = p$

standard error $= \sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$

Interpretation: If we take lots of samples and calculate $\hat{p}$ for each, the average of the $\hat{p}$ will be close to p.

Central Limit Theorem for $\hat{p}$

Let $\hat{p}$ be the sample proportion based on a sample of size n from a population with probability of success equal to p.

Assume the *expected number of successes* (np) and the *expected number of failures* ($n(1-p)$) are **both at least** 15. Then by the Central Limit Theorem, the sampling distribution of $\hat{p}$ is:

$$\hat{p} \dot{\sim} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

Example 7.8

Suppose that 29% of University of Minnesota students feel that our campus is becoming more dangerous. Take a sample of 100 university students and let $\hat{p}$ = the proportion of the sample that feel campus is becoming less safe.

(a) Find the mean and standard deviation of the sampling distribution of $\hat{p}$.

(b) What is the approximate sampling distribution of $\hat{p}$?

(c) Select one person at random from the sample of students. What is the probaility that his student feels campus is becoming more dangerous?

(d) What is the probability that fewer than 20% of those surveyed feel campus is becoming more dangerous?

Example 7.8

Suppose that 29% of University of Minnesota students feel that our campus is becoming more dangerous. Take a sample of 100 university students and let $\hat{p}$ = the proportion of the sample that feel campus is becoming less safe.

(a) Find the mean and standard deviation of the sampling distribution of $\hat{p}$. $\mu_{\hat{p}} = p = 0.29$ and $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{0.29(0.71)/100} = 0.045$.

(b) What is the approximate sampling distribution of $\hat{p}$?
$\hat{p} \sim N(0.29, 0.045)$

(c) Select one person at random from the sample of students. What is the probability that his student feels campus is becoming more dangerous? 0.29

(d) What is the probability that fewer than 20% of those surveyed feel campus is becoming more dangerous?
$P(\hat{p} < 0.2) = P(\frac{\hat{p}-0.29}{0.045} < \frac{0.2-0.29}{0.045}) = P(Z < -2) = 0.0228$.