Instructor: STH, TA: Habiba Afroz
Semester : Summer 2025

§

1 Sampling Distribution of Sample Means

In this chapter we are interested in the **distribution of the sample means in repeated sampling**, which means we are interested in the distribution of $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Here $X_1, X_2, \ldots, X_n$ are random variables which are part of a random sample of size n. You should always keep in mind random sample means random rows....

X_i	$Value$
X_1	=?
X_2	=?
...	...
X_n	=?

One of the important assumptions in this case, whether or we not we have **i.i.d. random sample**. Here i.i.d. random sample means the random variables $X_1, X_2, \ldots, X_n$ are all identically and independently distributed. Let's discuss what this means and what this gives us.

1. *What does Identical Distributions Mean?....* All the random variables in the random sample $X_1, X_2, \ldots, X_n$ have identical or same distribution (again this means no matter what the distribution is they are same for all rows in the sample). Note that this also means they have same population mean and same population variance, so if the population mean is μ and population variance is σ^2, then we have

$$\mathbb{E}[X_1] = \mathbb{E}[X_2] = \ldots = \mathbb{E}[X_n] = \mu$$
$$\mathbb{V}[X_1] = \mathbb{V}[X_2] = \ldots = \mathbb{V}[X_n] = \sigma^2$$

2. *What does Independent Distributions Mean?....* All the random variables in the random sample $X_1, X_2, \ldots, X_n$ are independent. This means knowing the value of one random variable does not give any information about the other random variables. For example if we know $X_1 = 5$, this does not give any information about $X_2, X_3, \ldots, X_n$. Also this means for any i and j, we have

$$\mathbb{C}\text{ov}(X_i, X_j) = 0 \text{ for any } i \neq j$$

What does i.i.d. random sample give us?

Here is what we get if we assume i.i.d. random sample with population mean μ and population variance σ^2, first

$$\begin{aligned}
\mathbb{E}[\overline{X}] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] \\
&= \frac{1}{n}\left(\underbrace{\mathbb{E}[X_1]}_{\mu} + \ldots + \underbrace{\mathbb{E}[X_n]}_{\mu}\right) \quad \text{(since } X_i\text{'s have identical distribution and identical means)} \\
&= \frac{1}{n} \cdot n\mu \\
&= \mu
\end{aligned}$$

$$\begin{aligned}
\mathbb{V}[\overline{X}] &= \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n^2}\mathbb{V}\left[\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n^2}\mathbb{V}\left(X_1 + \ldots + X_n\right) \\
&= \frac{1}{n^2}\left(\mathbb{V}\left(X_1\right) + \ldots + \mathbb{V}\left(X_n\right)\right) \quad \text{(since } X_i\text{'s are independent)} \\
&= \frac{1}{n^2}\left(\underbrace{\mathbb{V}\left(X_1\right)}_{\sigma^2} + \ldots + \underbrace{\mathbb{V}\left(X_n\right)}_{\sigma^2}\right) \quad \text{(since } X_i\text{'s are independent and also have identical variance)} \\
&= \frac{1}{n^2} \cdot n\sigma^2 \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

So we get a very nice result that if the population mean is μ and population variance is σ^2, then the sampling distribution of the sample mean $\overline{X}$ has mean μ and variance $\frac{\sigma^2}{n}$.

What happens when dependent?... Recall

If any two random variables X_1 and X_2 are not independent or dependent then we have

$$\mathbb{V}(X_1 + X_2) = \mathbb{V}(X_1) + \mathbb{V}(X_2) + 2\mathbb{C}\text{ov}(X_1, X_2)$$

So clearly covariance remains when they are no independent. when we have $X_1, \ldots, X_n$, we have

$$\mathbb{V}(X_1 + \ldots + X_n) = \mathbb{V}(X_1) + \ldots + \mathbb{V}(X_n) \underbrace{+2\mathbb{C}\text{ov}(X_1, X_2) + 2\mathbb{C}\text{ov}(X_1, X_3) + \ldots + 2\mathbb{C}\text{ov}(X_{n-1}, X_n)}_{\text{all pairs of Covariances}}$$

When the random variables are independent what happens is that **all these covariances become zero**... so you simply get

$$\mathbb{V}(X_1 + \ldots + X_n) = \mathbb{V}(X_1) + \ldots + \mathbb{V}(X_n)$$

And this leads to the important result that $\mathbb{V}(\overline{X}) = \frac{\sigma^2}{n}$

However when they are not independent, you have to take into account all these covariances. So in that case the formula for the variance of the sample mean does not hold and in particular this can happen when we have a finite population and we are doing sampling without replacement. In that case the random variables are not independent, because if we know one value, it gives us some information about the other values since we are not replacing the values back to the population.

So now we write the important theorem, which explains the sampling distribution of the sample mean $\overline{X}$ for both cases, when we have sampling with replacement (or infinite population) and when we have sampling without replacement from a finite population.

Theorem 1.1 (Sampling Distribution of Sample Mean)**.** *Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with mean μ and variance σ^2.*

For sampling with replacement (or infinite population):

$$\mathbb{E}[\overline{X}] = \mu$$
$$\mathbb{V}[\overline{X}] = \frac{\sigma^2}{n}$$

For sampling without replacement from finite population of size N:

$$\mathbb{E}[\overline{X}] = \mu$$
$$\mathbb{V}[\overline{X}] = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

*where $\frac{N-n}{N-1}$ is called the **finite population correction factor**.*

Remark 1.2.

- The finite population correction factor approaches 1 as $N \to \infty$, which means for large populations, the difference between sampling with and without replacement becomes negligible.
- if we do not have independence, we cannot use the formula σ^2/n for the variance of the sample mean. The formula for the expectation holds in both cases. In that case the covariance between the sample means must be taken into account.

Standard Error of the Sample Mean

Here the standard deviation of the sample mean $\overline{X}$ is called the **standard error**, again we need to distinguish between sampling with and without replacement. So we have

For sampling without replacement from finite population of size N

$$\begin{aligned} \mathrm{SE}(\overline{X}) &= \sqrt{\mathbb{V}(\overline{X})} \\ &= \sqrt{\frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} \end{aligned}$$

For sampling with replacement or infinite population

$$\begin{aligned} \mathrm{SE}(\overline{X}) &= \sqrt{\mathbb{V}(\overline{X})} \\ &= \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} \end{aligned}$$

1.1 Basic Problem Related to Sampling Distribution

1. This is similar to the discussion of Chapter 6.1 (page 252) of Newbold et al. (2020).

   Suppose we have following population

$$20, 30, 40, 20, 50$$

   (a) What is the population expectation in this case? (you can do average or expectation, it should be the same)

   (b) Suppose we collected a sample of size 3 from the population, how many samples can we draw? (assume we are doing sampling without replacement)? Write all the samples.

   (c) Calculate all the sample means and write the sampling distribution of sample means $\bar{X}$. This means writing the PMF of $\overline{X}$.

   (d) What is the mean of the sampling distribution, i.e., $\mathbb{E}[\overline{X}]$?

   (e) What is the variance of the sampling distribution, i.e., $\mathbb{V}[\overline{X}]$?

   (f) What is the standard error of the sample mean $\overline{X}$, or what is $\text{SE}(\overline{X})$?

   (g) Do you get $\mathbb{E}[\overline{X}] = \mu$?

   (h) In this case we have a finite population and we are doing sampling without replacement. So we need to use the finite population correction factor. The variance of the sampling distribution is given by:

$$\mathbb{V}[\overline{X}] = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

   Show that the answer you got from question (e) is same as the answer you get from this formula. (Hint: First calculate the population variance σ^2 using the population data above, then use the formula to calculate the variance of the sampling distribution)

   (i) Suppose we don't know population or population variance σ^2, we only have one sample which is $20, 40, 50$. Based on this sample what is the estimate of the standard error? (Hint: Replace σ by s in the formula of standard error)

Solution:

(a) Here the Population mean can be calculated as

$$\mu = \frac{20 + 30 + 40 + 20 + 50}{5} = 32 \tag{1}$$

But we can also calculate Population PMF (or TRUE PMF) and apply the formula for the expectation of a discrete random variable.

x	$f(x)$
20	2/5
30	1/5
40	1/5
50	1/5

In this case we get

$$\begin{aligned}\mu &= \sum x f(x) \\ &= 20 \cdot \frac{2}{5} + 30 \cdot \frac{1}{5} + 40 \cdot \frac{1}{5} + 50 \cdot \frac{1}{5} = 32\end{aligned}$$

The answer should be same, but calculating direct average is easier. Recall Expectation formula gives us a way to calculate the average of a population when we have the PMF but we don't have the whole population data. But in this case we have the whole population data, so we can do average directly.

(b) Number of samples of size $n = 3$, is $\binom{5}{3} = 10$.

Here are the samples

Sample Number	Sample
1	$20, 30, 40$
2	$20, 30, 20$
3	$20, 30, 50$
4	$20, 40, 20$
5	$20, 40, 50$
6	$20, 20, 50$
7	$30, 40, 20$
8	$30, 40, 50$
9	$30, 20, 50$
10	$40, 20, 50$

(c) Recall sampling distribution of $\bar{X}$ is the distribution of all possible sample means. So we need to calculate all the sample means and then write the PMF of $\bar{X}$. We write the sample means in the following table

#	Sample (distinct units)	$\overline{X}$
1	$20, 30, 40$	$\frac{1}{3}(20+30+40) = \frac{90}{3} = 30$
2	$20, 30, 20$	$\frac{1}{3}(20+30+20) = \frac{70}{3}$
3	$20, 30, 50$	$\frac{1}{3}(20+30+50) = \frac{100}{3}$
4	$20, 40, 20$	$\frac{1}{3}(20+40+20) = \frac{80}{3}$
5	$20, 40, 50$	$\frac{1}{3}(20+40+50) = \frac{110}{3}$
6	$20, 20, 50$	$\frac{1}{3}(20+20+50) = \frac{90}{3} = 30$
7	$30, 40, 20$	$\frac{1}{3}(30+40+20) = \frac{90}{3} = 30$
8	$30, 40, 50$	$\frac{1}{3}(30+40+50) = \frac{120}{3} = 40$
9	$30, 20, 50$	$\frac{1}{3}(30+20+50) = \frac{100}{3}$
10	$40, 20, 50$	$\frac{1}{3}(40+20+50) = \frac{110}{3}$

These are all sample means, now we can write the PMF of $\overline{X}$, which is the sampling distribution of $\overline{X}$

$\overline{X}$	$f(\overline{x})$
70/3	1/10
80/3	1/10
30	3/10
100/3	2/10
110/3	2/10
40	1/10

(d) Let's calculate the mean of the sampling distribution of $\bar{X}$, again since we have now the PMF of $\bar{X}$, we can use the formula for the expectation of a discrete random variable. So the mean of the sampling distribution

$$\begin{aligned}\mathbb{E}[\bar{X}] &= \sum \bar{x} f(\bar{x}) \\ &= \left(\frac{70}{3} \times \frac{1}{10}\right) + \left(\frac{80}{3} \cdot \frac{1}{10}\right) + \left(30 \times \frac{3}{10}\right) \\ &+ \left(\frac{100}{3} \times \frac{2}{10}\right) + \left(\frac{110}{3} \times \frac{2}{10}\right) + \left(40 \times \frac{1}{10}\right) \\ &= 32\end{aligned}$$

or we can also calculate the mean of the sampling distribution directly from all the possible sample means... the result will be same, I will skip it here.

(e) Now we calculate the variance of the sampling distribution. We will use the formula $\mathbb{V}(\bar{X}) = \mathbb{E}[\bar{X}^2] - \big(\mathbb{E}[\bar{X}]\big)^2$. We already have $\mathbb{E}[\bar{X}] = 32$, so we need to calculate $\mathbb{E}[\bar{X}^2]$.

$$\begin{aligned}\mathbb{E}\left[\bar{X}^2\right] &= \left(\frac{70}{3}\right)^2 \frac{1}{10} + \left(\frac{80}{3}\right)^2 \frac{1}{10} + 30^2 \frac{3}{10} + \left(\frac{100}{3}\right)^2 \frac{2}{10} + \left(\frac{110}{3}\right)^2 \frac{2}{10} + 40^2 \frac{1}{10} \\ &= \frac{4900}{9} \cdot \frac{1}{10} + \frac{6400}{9} \cdot \frac{1}{10} + 900 \cdot \frac{3}{10} + \frac{10000}{9} \cdot \frac{2}{10} + \frac{12100}{9} \cdot \frac{2}{10} + 1600 \cdot \frac{1}{10} \\ &= \frac{490}{9} + \frac{640}{9} + \frac{270 \cdot 9}{9} + \frac{2000}{9} + \frac{2420}{9} + \frac{160 \cdot 9}{9} \\ &= \frac{9420}{9} = \frac{3140}{3} \approx 1046.667.\end{aligned}$$

So we get

$$\mathbb{V}(\bar{X}) = \mathbb{E}\left[\bar{X}^2\right] - \mathbb{E}[\bar{X}]^2 = \frac{3140}{3} - 32^2 = \frac{3140}{3} - 1024 = \frac{68}{3} \approx 22.667,$$

(f) Standard error is (Recall Standard error is the standard deviation of the sampling distribution)

$$\mathrm{SE}(\bar{X}) = \sqrt{\mathbb{V}(\bar{X})} = \sqrt{\frac{68}{3}} \approx 4.761$$

(g) Yes we did get $\mathbb{E}[\bar{X}] = \mu = 32$, ... this is called **unbiasedness** of the sample mean. This means if we sample many times and take the average of all the sample means, we will get the population mean.

(h) Now we check with finite-population correction formula for the variance. First we need to calculate the population variance, recall the population from question is,

$$20, 30, 40, 20, 50$$

We already calculated the population mean $\mu = 32$. Now we calculate the population variance σ^2 using the formula for the population variance and using PMF, the PMF from the population is

x	$f(x)$
20	2/5
30	1/5
40	1/5
50	1/5

$$\begin{aligned}\sigma^2 &= \mathbb{E}(X^2) - \mu^2 \\ &= \left(20^2 \cdot \frac{2}{5} + 30^2 \cdot \frac{1}{5} + 40^2 \cdot \frac{1}{5} + 50^2 \cdot \frac{1}{5}\right) - 32^2 \\ &= (160 + 180 + 320 + 500) - 1024 \\ &= 1160 - 1024 = 136\end{aligned}$$

Since we have the population, we can also calculate the variance using average of the squared deviations in the population (with denominator N since it's for the population not sample). So we can also calculate the population variance as follows

$$\begin{aligned}\sigma^2 &= \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \\ &= \frac{1}{5}\left[(20-32)^2 + (30-32)^2 + (40-32)^2 + (20-32)^2 + (50-32)^2\right] \\ &= \frac{1}{5}\left[144 + 4 + 64 + 144 + 324\right] \\ &= \frac{680}{5} = 136.\end{aligned}$$

Now we can check

$$\frac{\sigma^2}{n} \cdot \frac{N-n}{N-1} = \frac{136}{3} \cdot \frac{2}{4} = \frac{68}{3}$$

So this is exactly what we got for the variance of the sampling distribution above. So this confirms the finite population correction formula for the variance of the sample mean.

Additional Remarks:

If we know the population standard deviation σ, we can use this correction factor to also calculate the standard error as follows,

$$\begin{aligned}\text{SE}(\bar{X}) &= \sqrt{\mathbb{V}(\bar{X})} \\ &= \sqrt{\frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}} = \sqrt{\frac{136}{3} \cdot \frac{2}{4}} = \sqrt{\frac{68}{3}} \approx 4.761\end{aligned}$$

(i) If we have the sample $20, 40, 50$ and don't know the population or σ, we cannot calculate the standard error directly, but we can definitely estimate the standard error using the sample standard deviation s. The idea is to replace σ by s in the formula of standard error. So we have

$$\widehat{\text{SE}}(\bar{X}) = \frac{s}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} \quad \text{(with finite population correction)}$$
$$\widehat{\text{SE}}(\bar{X}) = \frac{s}{\sqrt{n}} \quad \text{(without finite population correction)}$$

For this sample the sample mean is $\bar{x} = \frac{20+40+50}{3} = 36.67$. Now we can calculate the sample variance s^2 using the formula for the sample variance (with denominator $n-1$ since it's for the sample not population... careful). So we can calculate the sample variance as follows

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \frac{1}{2}\left[(20-36.67)^2 + (40-36.67)^2 + (50-36.67)^2\right] \\ &= \frac{1}{2}\left[277.78 + 11.11 + 177.78\right] \\ &= \frac{1}{2}\cdot 466.67 = 233.33 \end{aligned}$$

You did this previously using the box, I just did with sum, here is the box just to remind you (same thing)

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
20	-16.67	277.78
40	3.33	11.11
50	13.33	177.78

now we sum the last column to get the sample variance

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \frac{1}{2}\left[277.78 + 11.11 + 177.78\right] \\ &= \frac{466.67}{2} = 233.33 \end{aligned}$$

The sample standard deviation is $s = \sqrt{233.33} \approx 15.27$. Now the estimate of the standard error using the corection factor since we have a finite population and we are doing sampling without replacement is given by

$$\widehat{\text{SE}}(\bar{X}) = \frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{\sqrt{700/3}}{\sqrt{3}}\sqrt{\frac{2}{4}} = \sqrt{\frac{350}{9}} \approx 6.24.$$

2 Exact Sampling Distribution Under Normality

If the population is normally distributed, then you might guess that the sampling distribution of the sample mean $\overline{X}$ is also normally distributed. And this is actually true, although we won't prove it here but we give the following theorem. Important is this happens regardless of the sample size, we can write this result in the following theorem

Theorem 2.1 (Normal Distribution of Sample Mean)**.** *Let $X_1, X_2, \ldots, X_n$ be a random sample from a* ***normally distributed population*** *with mean μ and variance σ^2, this means $X_i \sim \mathcal{N}(\mu, \sigma^2)$ for all i. Then the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is also normally distributed, in symbol we write,*

$$\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Standardized form:

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

Remark 2.2. This theorem shows that if the population is normally distributed, then the sample mean is **exactly** normally distributed for any sample size n. This is different from the Central Limit Theorem, which requires large sample sizes for non-normal populations, which we will see next

2.1 Problem Under Normality Assumption

2. Suppose we know that the population of ECO 104 marks in EWU is normally distributed with mean $\mu = 75$ and standard deviation $\sigma = 10$.
   (a) What is the distribution of the sample mean $\overline{X}$ for a sample size of $n = 30$?
   (b) What is the standard error of the sample mean $\overline{X}$?
   (c) Calculate the probability that the sample mean $\overline{X}$ is greater than 78. What is the Frequency interpretation of this probability?
   (d) Calculate the probability that the sample mean $\overline{X}$ is between 78 and 82. What is the Frequency interpretation of this probability?
   (e) Calculate the probability that the sample mean $\overline{X}$ is less than 70. What is the Frequency interpretation of this probability?
   (f) If we know that the sample standard deviation is $s = 12$, what is the **estimate of the standard error** of the sample mean $\overline{X}$?

Solution:

(a) What is the distribution of the sample mean $\overline{X}$ for a sample size of $n = 30$?

Ans:

Since the population is normally distributed with mean $\mu = 75$ and standard deviation $\sigma = 10$, the sampling distribution of the sample mean $\overline{X}$ for a sample size of $n = 30$ is given by

$$\overline{X} \sim \mathcal{N}\left(75, \frac{10^2}{30}\right)$$

(b) What is the standard error of the sample mean $\overline{X}$?

Ans:

The variance of the sampling distribution of the sample mean $\overline{X}$ is $\frac{\sigma^2}{n} = \frac{10^2}{30} = \frac{100}{30} = \frac{10}{3}$. So the standard error is

$$\text{SE}(\overline{X}) = \sqrt{\frac{10^2}{30}} = \sqrt{\frac{100}{30}} = \frac{10}{\sqrt{30}} \approx 1.83$$

(c) Calculate the probability that the sample mean $\overline{X}$ is greater than 78. What is the Frequency interpretation of this probability?

Ans:

This is easy once you standardize the variable. We have

$$\begin{aligned}\mathbb{P}(\overline{X} > 78) &= \mathbb{P}\left(Z > \frac{78-75}{10/\sqrt{30}}\right) \\ &= \mathbb{P}\left(Z > \frac{3}{10/\sqrt{30}}\right) \\ &= \mathbb{P}(Z > 1.64) \\ &= 1 - \mathbb{P}(Z \leq 1.64) \\ &= 1 - 0.9495 = 0.0505\end{aligned}$$

(d) Calculate the probability that the sample mean $\overline{X}$ is between 78 and 82. What is the Frequency interpretation of this probability?

Ans:

You should be able to do this now!

(e) Calculate the probability that the sample mean $\overline{X}$ is less than 70. What is the Frequency interpretation of this probability?

Ans:

You should be able to do this now!

(f) If we know that the sample standard deviation is $s = 12$, what is the **estimate of the standard error** of the sample mean $\overline{X}$?

Ans:

This is easy, just replace σ by s in the formula of standard error

$$\widehat{\mathrm{SE}}(\overline{X}) = \frac{s}{\sqrt{n}} = \frac{12}{\sqrt{30}} \approx 2.19$$

Note we used here " $\widehat{\cdot}$ " symbol to denote estimate.

3 Approximate Sampling Distribution Under Large Samples

If the population is not normally distributed, the only savior for us is the well known **Central Limit Theorem (CLT)**. CLT states that the sampling distribution of the sample mean $\overline{X}$ will be approximately normally distributed for sufficiently large sample sizes (usually we think $n \geq 30$), no matter what the shape of the population distribution is. This is a very powerful result because it allows us to make inferences about the population mean using the normal distribution, even when the population itself is not normal. We state the CLT in the following theorem

Theorem 3.1 (Central Limit Theorem (CLT))**.** *Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with mean μ and finite variance σ^2 (the population can have any distribution). As the sample size n becomes large the sampling distribution of the sample mean $\overline{X}$ approaches a normal distribution:*

$$\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \quad \textit{approximately as } n \to \infty$$

Equivalently, the standardized sample mean converges to a standard normal distribution:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1) \quad \textit{approximately as } n \to \infty$$

Remark 3.2.

- The Central Limit Theorem (CLT) states that the sampling distribution of the sample mean will be approximately normally distributed for sufficiently large sample sizes, regardless of the population's distribution. This is remarkable because it applies to **any** population distribution (uniform, exponential, skewed, etc.) as long as the population has some mean and variance. The approximation becomes better as n increases.
- **Rule of thumb:** For most practical purposes, $n \geq 30$ is considered "large enough" for the CLT to provide a good approximation, though for highly skewed distributions, larger sample sizes may be needed.

We will see an application of CLT for Bernoulli Random Variables.

3.1 CLT for Binomials (or Bernoullis)

Suppose we have following dataset of 30 students from EWU, which says whether they are happy or not.

Student ID	Happy (1) / Not Happy (0)	RV
1	1	X_1
2	0	X_2
3	1	X_3
	$\vdots$	$\vdots$
30	0	X_{30}

Here if we think about sample mean

$$\overline{X} = \frac{1}{30}\sum_{i=1}^{30} X_i = \frac{X_1 + X_2 + \ldots + X_{30}}{30} = \frac{1 + 0 + \ldots + 0}{30}$$

This is actually the **sample proportion of happy students** in the sample. since this is a proportion, **we will use NOT USE $\overline{X}$ but we will use $\overline{p}$**, so here

$$\overline{p} = \frac{1}{30}\sum_{i=1}^{30} X_i$$

Our goal is to get the sampling distribution of $\overline{p}$, or in other words we want to know

$$\overline{p} \sim ???$$

To answer this question we need to know how $X_1, X_2, X_3, \ldots, X_{30}$ are distributed. Actually you already know the answer. Since these are all 0/1 variables, they all are Bernoulli distributed with a **some success probability** p. So we write,

$$X_i \sim \text{Bern}(p) \text{ for all } i = 1, 2, \ldots, 5$$

So now we can write the mean and variance of each X_i as follows,

$$\mu = \mathbb{E}[X_i] = p \text{ for all } i = 1, 2, \ldots, 30$$

and

$$\sigma^2 = \text{Var}(X_i) = p(1-p) \text{ for all } i = 1, 2, \ldots, 30$$

Now the question is what is p, actually here p is **population proportion of happy students** in the population. Now we can calculate the mean and variance of $\overline{p}$ as follows

$$\mathbb{E}[\overline{p}] = \mu = p$$

and

$$\mathbb{V}(\overline{p}) = \frac{\sigma^2}{n} = \frac{p(1-p)}{n}$$

Now we can apply the Central Limit Theorem (CLT) to the sample mean or in this case which is the sample proportion $\overline{p}$. So by CLT we have

$$\overline{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right) \quad \text{approximately as } n \to \infty$$

or standardized:

$$Z = \frac{\overline{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{N}(0,1) \quad \text{approximately as } n \to \infty$$

So what this says is, the sample proportion is approximately normally distributed for large samples with mean p and variance $\frac{p(1-p)}{n}$, and the standardized version Z is also approximately normally distributed with mean 0 and variance 1.

Note in this case the standard error is

$$\text{SE}(\overline{p}) = \sqrt{\mathbb{V}(\overline{p})} = \sqrt{\frac{p(1-p)}{n}}$$

Now one last thing remains, we usually never know the true population proportion p. However, we can use the sample proportion $\overline{p}$ as an estimate for p. And there is a result which says, even if we replace $\overline{p}$ with the sample proportion, the CLT still holds for Z, so this means

$$Z = \frac{\overline{p} - p}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}} \sim \mathcal{N}(0,1) \quad \text{approximately as } n \to \infty$$

However, if you know the population proportion p, then you should use p in the formula of Z instead of $\overline{p}$.

3.2 Problem Under Approximate Normality - Applying CLT for Sample Proportions

3. This is taken from Anderson et al. (2020), Chapter 7.6 Problem 36

   The Wall Street Journal reported that the age at first startup for 55% of entrepreneurs was 29 years of age or less and the age at first startup for 45% of entrepreneurs was 30 years of age or more.

   (a) Suppose a sample of 200 entrepreneurs will be taken to learn about the most important qualities of entrepreneurs. Show the sampling distribution of $\bar{p}$ where $\bar{p}$ is the sample proportion of entrepreneurs whose first startup was at 29 years of age or less.

   (b) What is the probability that the sample proportion in part (a) will be within $\pm.05$ of its population proportion?

   (c) Suppose a sample of 200 entrepreneurs will be taken to learn about the most important qualities of entrepreneurs. Show the sampling distribution of $\bar{p}$ where $\bar{p}$ is now the sample proportion of entrepreneurs whose first startup was at 30 years of age or more.

   (d) What is the probability that the sample proportion in part (c) will be within $\pm.05$ of its population proportion?

   (e) Is the probability different in parts (b) and (d)? Why?

   (f) Answer part (b) for a sample of size 400 . Is the probability smaller? Why?

Solution:

(a) Suppose a sample of 200 entrepreneurs will be taken to learn about the most important qualities of entrepreneurs. Show the sampling distribution of $\bar{p}$ where $\bar{p}$ is the sample proportion of entrepreneurs whose first startup was at 29 years of age or less.

Ans:

Here the population proportion $p = 0.55$, and the sample size $n = 200$. So by CLT we have

$$\bar{p} \sim \mathcal{N}\left(0.55, \frac{0.55 \times (1 - 0.55)}{200}\right) = \mathcal{N}(0.55, 0.0012375)$$

(b) What is the probability that the sample proportion in part (a) will be within $\pm.05$ of its population proportion?

Ans:

We want to calculate

$$\begin{aligned}\mathbb{P}(p-0.05\leq\bar{p}\leq p+0.05) &= \mathbb{P}(0.50\leq\bar{p}\leq 0.60)\\ &= \mathbb{P}\left(\frac{0.50-0.55}{\sqrt{0.0012375}}\leq Z\leq\frac{0.60-0.55}{\sqrt{0.0012375}}\right)\\ &= \mathbb{P}(-1.42\leq Z\leq 1.42)\\ &= \mathbb{P}(Z\leq 1.42)-\mathbb{P}(Z\leq -1.42)\\ &= 0.9222-0.0778=0.8444\end{aligned}$$

(c) Suppose a sample of 200 entrepreneurs will be taken to learn about the most important qualities of entrepreneurs. Show the sampling distribution of $\bar{p}$ where $\bar{p}$ is now the sample proportion of entrepreneurs whose first startup was at 30 years of age or more.

Ans:

Here the population proportion $p=0.45$, and the sample size $n=200$. So by CLT we have

$$\bar{p}\sim\mathcal{N}\left(0.45,\frac{0.45\times(1-0.45)}{200}\right)=\mathcal{N}(0.45,0.0012375)$$

(d) What is the probability that the sample proportion in part (c) will be within $\pm.05$ of its population proportion?

Ans: We want to calculate

$$\begin{aligned}\mathbb{P}(p-0.05\leq\bar{p}\leq p+0.05) &= \mathbb{P}(0.40\leq\bar{p}\leq 0.50)\\ &= \mathbb{P}\left(\frac{0.40-0.45}{\sqrt{0.0012375}}\leq Z\leq\frac{0.50-0.45}{\sqrt{0.0012375}}\right)\\ &= \mathbb{P}(-1.42\leq Z\leq 1.42)\\ &= \mathbb{P}(Z\leq 1.42)-\mathbb{P}(Z\leq -1.42)\\ &= 0.9222-0.0778=0.8444\end{aligned}$$

(e) Is the probability different in parts (b) and (d)? Why?

Ans:

No, the probability is the same in parts (b) and (d) because the standard deviation of the sampling distribution is the same in both cases.

(f) Answer part (b) for a sample of size 400 . Is the probability smaller? Why?

Ans:

The probability will be smaller for a sample size of 400 because the standard deviation of the sampling distribution decreases as the sample size increases. Specifically, the standard deviation is given by $\sqrt{\frac{p(1-p)}{n}}$, so when n increases, the standard deviation decreases, leading to a narrower distribution and thus a smaller probability of being within a fixed range around the population proportion.

Remarks: You should look at Chapter 7.5 and 7.6 in Anderson et al. (2020) and also Chapter 6.1,6.2 and 6.3 in Newbold et al. (2020) for more problems.

Technical Appendix (NOT FOR THE EXAMS)

Note: Please DO NOT LOSE HOPE when you see this ....this is OPTIONAL, so NO NEED TO READ if you don't want to! If you are curious, then you can ... and of course I can explain this in office hours and you will understand better, God willing!

Let's see how do we **derive the finite population correction factor**. Suppose we have a finite population of size N with mean μ and variance σ^2. So this means we have a population

$$x_1, x_2, \ldots, x_N$$

Recall we are interested to get the covariance term $\text{Cov}(X_i, X_j)$ for $i \neq j$ where X_i and X_j are two distinct draws from the population without replacement. In the i.i.d. setup or in the sampling with replacement setup, this covariance term is zero. But here since we are doing sampling without replacement, this covariance term is not zero. So we need to calculate this covariance term.
Okay... so suppose we want to calculate the covariance between X_1, X_2 when sampling without replacement. If we know the formula for X_1 and X_2, for other pairs it will be similar. Recall the formula for the covariance is

$$\mathbb{C}\text{ov}(X_1, X_2) = \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1]\mathbb{E}[X_2]$$

Here we know $\mathbb{E}[X_1] = \mathbb{E}[X_2] = \mu$, so we only need to calculate $\mathbb{E}[X_1 X_2]$. Now note, the ordered pair (X_1, X_2) takes (x_i, x_k) with probability $\frac{1}{N(N-1)}$ for any $i \neq k$. Hence

$$\mathbb{E}\left[X_1 X_2\right] = \sum_{i=1}^{N} \sum_{\substack{k=1 \\ k \neq i}}^{N} \frac{x_i x_k}{N(N-1)} = \frac{1}{N(N-1)} \sum_{\substack{i,k=1 \\ i \neq k}}^{N} x_i x_k$$

You might be wondering, why the denominator is $N(N-1)$, You choose the first index in N ways and the second in $N-1$ ways (without replacement) ordered pairs. We can also use the following identity to simplify the summation

$$\sum_{\substack{i,k=1 \\ i \neq k}}^{N} x_i x_k = \left(\sum_{i=1}^{N} x_i \right)^2 - \sum_{i=1}^{N} x_i^2$$

With $\mu = \frac{1}{N} \sum_i x_i$ and $\sigma^2 = \frac{1}{N} \sum_i x_i^2 - \mu^2$, we have

$$\sum_i x_i = N\mu, \quad \sum_i x_i^2 = N\left(\sigma^2 + \mu^2\right)$$

Therefore

$$\mathbb{E}\left[X_1 X_2\right] = \frac{N^2\mu^2 - N\left(\sigma^2 + \mu^2\right)}{N(N-1)} = \frac{N^2\mu^2 - N\sigma^2 - N\mu^2}{N(N-1)} = \frac{(N-1)N\mu^2 - N\sigma^2}{N(N-1)} = \mu^2 - \frac{\sigma^2}{N-1}$$

so we get,

$$\begin{aligned}\mathbb{C}\text{ov}\left(X_1, X_2\right) &= \mathbb{E}\left[X_1 X_2\right] - \mu^2 \\ &= -\frac{\sigma^2}{N-1}.\end{aligned}$$

This means in general for any $i \neq j$, we have

$$\mathbb{C}\text{ov}\left(X_i, X_j\right) = -\frac{\sigma^2}{N-1}.$$

Now we can recall

$$\begin{aligned}\mathbb{V}(\bar{X}) = \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) &= \frac{1}{n^2}\,\mathbb{V}\left(\sum_{i=1}^{n}\left(X_i\right)\right) \\ &= \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathbb{V}\left(X_i\right) + \sum_{i=1}^{n}\sum_{\substack{k=1 \\ k\neq i}}^{n}\mathbb{C}\text{ov}\left(X_i, X_k\right)\right)\end{aligned}$$

Now we can use the results we have

$$\sum_{i=1}^{n}\mathbb{V}\left(X_i\right) = n\sigma^2, \quad \sum_{i=1}^{n}\sum_{\substack{k=1 \\ k\neq i}}^{n}\mathbb{C}\text{ov}\left(X_i, X_k\right) = n(n-1)\left(-\frac{\sigma^2}{N-1}\right)$$

combining everything we get,

$$\mathbb{V}(\bar{X}) = \frac{1}{n^2}\left(n\sigma^2 - \frac{n(n-1)\sigma^2}{N-1}\right) = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right) = \frac{\sigma^2}{n}\cdot\frac{N-n}{N-1}.$$

References

Anderson, D. R., Sweeney, D. J., Williams, T. A., Camm, J. D., Cochran, J. J., Fry, M. J. & Ohlmann, J. W. (2020), *Statistics for Business & Economics*, 14th edn, Cengage, Boston, MA.

Newbold, P., Carlson, W. L. & Thorne, B. M. (2020), *Statistics for Business and Economics*, 9th, global edn, Pearson, Harlow, England.