Ch 5 Notes
Concepts and Problems on Sampling Distributions

ECO 104 (Section 7 and 9) @ EWU
Instructor: STH, TA: Habiba Afroz

§

1 Sampling Distribution of Sample Means

Suppose if we have an i.i.d. (identically and independently distributed) random sample, this means all the random variables in the random sample $X_1, X_2, \ldots, X_n$ have the same distribution (again this means no matter what is the distribution they are same). And further this means they have sample means and same variance,

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

Now moreover, if we assume they are independent, this means all pairs of covariances are zero

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

You should ask what happens if they are not independent. For two variables, X_1 and X_2 if they are not independent 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 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 $\mathbb{V}(\overline{X}) = \frac{\sigma^2}{n}$ So now we write the important theorem, that is

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.

Remark 1.3. Important... if we do not have independence, we cannot use the formula for the variance of the sample mean. The formula for the expectation holds. In that case the covariance between the sample means must be taken into account.

Here the standard deviation of the sample mean $\overline{X}$ is called the **standard error**, and we can write

$$\text{SE}(\overline{X}) = \sqrt{\mathbb{V}(\overline{X})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

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:

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

(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)

2 Exact Sampling Distribution Under Normality

If the population is normally distributed, then the sampling distribution of the sample mean $\overline{X}$ is also normally distributed, 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 $\overline{X}$?

3 Approximate Sampling Distribution Under Large Samples

If the population is not normally distributed, the Central Limit Theorem states that the sampling distribution of the sample mean $\overline{X}$ will be approximately normally distributed for sufficiently large sample sizes (usually we think $n \geq 30$).

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:*

$$\overline{X} \xrightarrow{d} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{as } n \to \infty$$

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

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

where $\xrightarrow{d}$ denotes convergence in distribution.

Remark 3.2. The Central Limit Theorem is remarkable because it applies to **any** population distribution (uniform, exponential, skewed, etc.) as long as the population has finite mean and variance. The approximation becomes better as n increases.

Remark 3.3. 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

Suppose we have following dataset of 5 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
4	1	X_4
5	0	X_5

Here if we think about a random sample of 5 students from EWU, the random variables X_1, X_2, X_3, X_4, X_5 all are Bernoulli distributed with a success probability p, where p is the proportion of happy students in the population. So we can write

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

Now in this case the mean is same for everyone which is

$$\mu = \mathbb{E}[X_i] = p \text{ for all } i = 1, 2, \ldots, 5$$

Now the variance is also same for everyone

$$\sigma^2 = \text{Var}(X_i) = p(1-p) \text{ for all } i = 1, 2, \ldots, 5$$

Now we can apply the Central Limit Theorem (CLT) to the sample mean $\overline{X}$. Here is since our population target is the population proportion p, we will use $\overline{p}$, rather than $\overline{X}$, and note that $\overline{p}$ is the sample proportion. Now the CLT says,

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

or standardized:

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

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 that in this case,

$$\mathbb{E}[\overline{p}] = p$$
$$\mathbb{V}(\overline{p}) = \frac{p(1-p)}{n}$$

and 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

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

3.2 Problem Under CLT

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?

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.

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.