Ch5 - Sampling Distribution

(Sampling Dist. of Means and Proportions)

Statistics For Business and Economics - I

Shaikh Tanvir Hossain

East West University, Dhaka

Last Updated May 21, 2025

Outline

Outline

- In this chapter we will see what happens when we do sampling. The important thing to always remember is that our sample itself is a random object. So anything that we can calculate from the sample is also random. Random objects that we calculate from the sample is generally called *Sample Statistic* (it's a function of the random sample). For example, sample mean, sample proportion are all examples of sample statistics.
- When a sample Statistic targets a population parameter we call it an *Estimator*. It's important that in real life we will never know the true population parameter. But we can use a sample an an estimator to estimate the population parameter.
- In *repeated sampling*, the probability distribution of a sample statistic or the probability distribution of an estimator is called *Sampling Distribution*.
- The idea of Sampling Distribution is very important and almost like THE fundamental topic Statistics. It helps us to asses the variability of the sample statistic.
- So let's start... 🚶 🚶 🚶 .

Sample, Population and Statistical Inference

Sample, Population and Statistical Inference

1. Statistical Estimation

Estimation

- Consider the sample again,

	Gender	Monthly Income (tk)	ECO-101 Grade	# Retakes
Student A	Male	3615	B-	3
Student B	Female	49755	A	2
Student C	Male	44758	A	1
Student D	Female	3879	B	0
Student E	Male	22579	A+	2

- Now again, think about following questions,
  - do we know the *population mean of income* from all students, *call it μ?*

    Ans - NO, we don't know μ, (Here μ is just some number which is the population average)
  - Similarly do we know the *population proportion of all female students*, call it p?

    *Ans - NO (here **p** is probability)*

- Now can we *estimate* μ or p? Ans: Yes - we can use our sample to *estimate*.
- For example, *sample mean of income*, denote this with $\bar{x}$ can be used to estimate μ. For example here $\bar{x} = 124586$ (Just simply take the average). This is the *estimate of the population mean μ*.
- Similarly *sample proportion of female students*, call it $\bar{p}$ can be used to estimate population proportion p. Here if we think Female $= 1$, and Male $= 0$, and then sample proportion is $\bar{p} = 0.4$ (Just simply take the average). This is the *estimate of the population proportion **p***.

Estimation

- This is the idea of *Estimation*, that is

  there is a target parameter for example μ or p, which is a population object, but we don't know it, then we will estimate these objects with some numbers using a sample.

- The number that we calculate using a sample is called *an estimate*.
- Now *does our estimate get better if we increase the size of the sample (or sample size n)*?
- The answer is Yes - if we have a good sample and we increase the sample size then maybe we expect that we will *eventually be very very close to the target parameter!*
- What happens if we get a bad sample? Then even if we increase the sample size, our estimate won't improve.
- There is a famous saying in Statistics, *Garbage in Garbage out!* This means if you have a bad sample, then even if you increase the sample size, you will not get a good estimate.

Estimation

- But if we have a good sample, there is a famous law in Statistics, called *Law of Large Numbers*, this says if we increase the sample size, then our estimate will get better and better and we will eventually hit the target parameter! In notation we can write this as

  $$\text{if } n \to \infty \text{ then } \bar{x} \to \mu \text{ and this happens with very high probability}$$

- This is one of the most important results in Statistics, that is with good samples, we can estimate the target parameter with high accuracy if we have a very large sample.
- This idea is also known as *Consistency* of an *estimator*, which we will discuss in coming sections.... Question - *What is an estimator?*

Estimation

- So again to summarize, this process of targeting something from the population and then guessing that with the help of sample is known as *Estimation* or more accurately this is called *Point Estimation*, it's a concept in Inferential Statistics, but there is another method known as hypothesis testing, we won't cover it here you will see it in the next course.
- Both estimation and testing are part of inferential statistics
- Question - *Since we can think the sample is random does this estimate change with different samples*? Ans: obviously YES!
- How do we write this? We need to think about random variables.... here the idea of *Estimator* comes...

Estimation

- Now let's consider another data, suppose we collected a data from 10 students, which are just income of 10 students, and we are interested in the population mean μ.

	Income	Random variable
1.	20	$X_1 = ?$
2.	60	$X_2 = ?$
3.	20	$X_3 = ?$
4.	-20	$X_4 = ?$
5.	-30	$X_5 = ?$
6.	-10	$X_6 = ?$
7.	80	$X_7 = ?$
8.	10	$X_8 = ?$
9.	30	$X_9 = ?$
10.	40	$X_{10} = ?$

Table 1: Income data

- As you probably already know, a data set can be thought in two ways, a fixed data or a random data.
- Note in the left column we have fixed data, but in the right column we have random variables. So X_1 is a random variable, X_2 is a random variable, ..., X_{10} is a random variable. Important is in the case of *realized data / fixed data*, the randomness is gone and we have observed the value.
- Generally when we think about n random variables, $X_1, X_2, X_3, \ldots, X_n$, we will call it a *random sample* (the other one is the fixed sample!)
- Now we will talk about Estimator.

Estimation

- The idea of *Estimator* comes when we think about a random sample. An Estimator is a *function of a random sample*, hence this is also a *random variable*. For example an estimator of μ is

$$\bar{X} = f(X_1, X_2, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Now this is not just ordinary sample mean. This is a random variable, since it's a function of random variables $X_1, X_2, \ldots, X_n$. It's value will change from sample to sample ....
- So now you should always remember the difference between $\bar{x}$ and $\bar{X}$. For a fixed sample $\bar{x}$ is a fixed number, but $\bar{X}$ is a random variable. This is random since it changes from sample to sample. But again, when we calculate it for a fixed sample, then we get a fixed number $\bar{x}$. Here $\bar{x}$ is a constant and it's not random.
- So the random variable $\bar{X}$ is an estimator of μ. And the fixed number $\bar{x}$ is an estimate of μ.

Estimation

- *Pay CLOSE ATTENTION, this is very important*

  Now since $\bar{X}$ is a random variable, question is

  what is the probability distribution of $\bar{X}$? or Expectation of $\mathbb{E}(\bar{X})$? Or $\mathbb{V}\mathrm{ar}(\bar{X})$
- To understand this we need to talk abut the *sampling distribution of $\bar{X}$*, which we will do in the next section.
- In repeated sampling the probability distribution of $\bar{X}$ is called *sampling distribution of $\bar{X}$*.

Sampling Distributions

Sampling Distributions

Mean and Variance of $\bar{X}$

Estimator $\bar{X}$, $\mathbb{E}(\bar{X})$ and $\mathbb{V}\mathrm{ar}(\bar{X})$

- Before we dive with the sampling distribution, let's talk about the estimator $\bar{X}$, and it's mean and variance. When it comes to $\bar{X}$ we are interested in 3 important questions when it comes to estimator,

1. What is the Expectation of the random variable $\bar{X}$, written as $\mathbb{E}(\bar{X})$?
2. What is the variance of the random variable $\bar{X}$, written $\mathbb{V}\mathrm{ar}(\bar{X})$?
3. What is the probability distribution of $\bar{X}$ (this is what we call *sampling distribution of $\bar{X}$!*)

Estimator $\bar{X}$, $\mathbb{E}(\bar{X})$ and $\mathbb{V}\mathrm{ar}(\bar{X})$

- *Question: What does $\mathbb{E}(\bar{X})$ mean?*
- *Question: What does $\mathbb{V}\mathrm{ar}(\bar{X})$ mean?*
- Answer to the first question is - *if we do repeated sampling - then our estimates, in this case example sample means $\bar{x}$ performs on average?*.
- An important question is - *does we have $\mathbb{E}(\bar{X})$ becomes equal to μ ?*

Estimator $\bar{X}$, $\mathbb{E}(\bar{X})$ and $\mathbb{V}\mathrm{ar}(\bar{X})$

- The second question answers *how much variability we have in our estimates if we do repeated sampling*.
- For example if we have $\mathbb{V}\mathrm{ar}(\bar{X})$ is small, then we know that our estimate is always close to μ (this is good!). But if we have $\mathbb{V}\mathrm{ar}(\bar{X})$ is large, then we know that our estimate is not always close to μ (this is a bad!)
- The answer to the third question is what we call *Sampling Distribution of Sample Means*. Note that, this is the distribution of sample means $\bar{x}$, that we get from repeated sampling!
- Definitely if we know the answer of 3, we know the answers of 1 and 2.
- Let's try to understand with the following picture. Suppose we do sampling many times and calculate $\bar{x}$ many times, here are four situations that can happen, at the center we have μ and the black dots are the estimates $\bar{x}$ for different samples.

Estimator $\bar{X}$, $\mathbb{E}(\bar{X})$ and $\mathbb{V}\mathrm{ar}(\bar{X})$

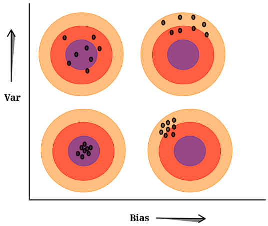

Figure 1: bias variance situations, true value μ is at the center, and the black dots are estimates or $\bar{x}$. Here in all four situations think about repeated sampling, i.e., we are calculating $\bar{x}$ in repeated sampling.

- 1. *top-left:* Here sometimes the estimates are hitting the target, but their accuracy overall is really bad. You can say on average they are performing well, but there is a lot of variability. This is what we call *low-bias & high-variance* situation.
- 2. *bottom-left:* This is better than the last one (in fact this is the best one) here estimates are always very close to the truth and also the variability is very low. This is what is called *low-bias & low-variance* situation. This is ideally what we want.
- 3. *bottom-right:* In this case the variability is not high, but the estimates are more or less always very off from the target. This is called *high-bias & low-variance* situation. This is not good, even if we have low variance.
- 4. *top-right:* This is the worst case, here the estimates are always very off from the target and also the variability is very high. This is called *high-bias & high-variance* situation.

Estimator $\bar{X}$, $\mathbb{E}(\bar{X})$ and $\mathbb{V}\text{ar}(\bar{X})$

- Definitely we want
$$\mathbb{E}(\bar{X}) = \mu$$
- If this happens we call $\bar{X}$ an "unbiased" estimator of μ. It means the sample average is a "good" estimator for the population mean μ. So you can think - if we calculate, sample means many times, *on average* we are not doing a bad job, even if our sample size n is not that big.
- Compare this with consistency!
- Recall consistency means
$$\text{if we increase } n \to \infty \text{ then } \bar{X} \text{ will become close to } \mu \text{ with high probability}$$
- Unbiasedness is for any sample, but consistency comes only when we increase the sample size.
- *An unbiased estimator is always consistent*. But a *consistent estimator may not be unbiased*.
- Now let's talk about the sampling distribution of $\bar{X}$. This is the distribution of $\bar{X}$ when we do repeated sampling.

Sampling Distributions

Sampling distribution of $\bar{X}$

Estimator $\bar{X}$, $\mathbb{E}(\bar{X})$ and $\mathbb{V}\mathrm{ar}(\bar{X})$

- We will state three important results, related to the sampling distribution of $\bar{X}$.

Theorem 5.1: (Mean and Variance of $\bar{X}$ with only i.i.d assumption)

Suppose the population distribution have mean μ and variance σ^2 and sample points are independent, then

$$\mathbb{E}\left(\bar{X}\right) = \mu \text{ and } \mathbb{V}\mathrm{ar}\left(\bar{X}\right) = \frac{\sigma^2}{n} \tag{1}$$

- How do we get this? The proof is very simple, just uses the rule for Expectation and Variance. First let's understand the result and then we will also see how we got this,

Estimator $\bar{X}$, $\mathbb{E}(\bar{X})$ and $\mathbb{V}\mathrm{ar}(\bar{X})$

- Note that, in this result *we don't have any distributional assumption* (i.e., we are not assuming they are normal or binomial or anything else...), we are assuming that the population mean and variance exists and we have i.i.d sample points.
- If we think about $X_1, X_2, \ldots, X_n$ random variable this means we have independent and identically distributed random variables (i.i.d) with the same mean μ and same variance σ^2. Again to explain further, this means
  1. $X_1, X_2, \ldots, X_n$ random variables are independent,
  2. $X_1, X_2, \ldots, X_n$ have same probability distribution where we have $\mathbb{E}(X_1) = \mathbb{E}(X_2) = \mathbb{E}(X_3) = \ldots, \mathbb{E}(X_n) = \mu$, and also $\mathbb{V}\mathrm{ar}(X_1) = \mathbb{V}\mathrm{ar}(X_2) = \mathbb{V}\mathrm{ar}(X_3) = \ldots, \mathbb{V}\mathrm{ar}(X_n) = \sigma^2$
- Finally you should always keep in mind that $\bar{X}$ is a random variable, since it's a function of random variables $X_1, X_2, \ldots, X_n$.
- Now let's see the details....

Estimator $\bar{X}$, $\mathbb{E}(\bar{X})$ and $\mathbb{V}\mathrm{ar}(\bar{X})$

- We know that $\mathbb{E}(X_i) = \mu$ and $\mathbb{V}\mathrm{ar}(X_i) = \sigma^2$, then we can apply the rules for expectation and variance.

$$\mathbb{E}(\bar{X}) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{n \times \mu}{n} = \mu$$

$$\mathbb{V}\mathrm{ar}(\bar{X}) = \mathbb{V}\mathrm{ar}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \overset{*}{=} \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\mathrm{ar}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{n \times \sigma^2}{n^2} = \frac{\sigma^2}{n}$$

- For expectation this is simply the rule for expectation - Expectation of sum of independent random variables is the sum of expectation of each random variable.
- For the variance we used the fact that X_i's are independent, so we can apply the rule for variance for the sum of independent random variables. In particular we used the fact that

$$\mathbb{V}\mathrm{ar}(X_1 + X_2) = \mathbb{V}\mathrm{ar}(X_1) + \mathbb{V}\mathrm{ar}(X_2) \text{ if } X_1 \text{ and } X_2 \text{ are independent}$$

- Finally we also used the fact that $\mathbb{V}\mathrm{ar}(aX) = a^2\mathbb{V}\mathrm{ar}(X)$ for any constant a.

Estimator $\bar{X}$, $\mathbb{E}(\bar{X})$ and $\mathbb{V}\mathrm{ar}(\bar{X})$

- Careful if they are not independent then this is not true! we will have

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

- When we have independence, covariance is zero, so we get

$$\begin{aligned} \mathbb{V}\mathrm{ar}(X_1 + X_2) &= \mathbb{V}\mathrm{ar}(X_1) + \mathbb{V}\mathrm{ar}(X_2) + 2\underbrace{\mathbb{C}\mathrm{ov}(X_1, X_2)}_{=0} \\ &= \mathbb{V}\mathrm{ar}(X_1) + \mathbb{V}\mathrm{ar}(X_2) \end{aligned}$$

Estimator $\bar{X}$, $\mathbb{E}(\bar{X})$ and $\mathbb{V}\mathrm{ar}(\bar{X})$

Theorem 5.2: (Distribution of $\bar{X}$ with Normality and i.i.d assumption)

If the *population distribution is $\mathcal{N}(\mu, \sigma^2)$* (this means population is normally distributed with mean μ and variance σ^2) and the sample points are independent, then

i) $\mathbb{E}(\bar{X}) = \mu$ and $\mathbb{V}\mathrm{ar}(\bar{X}) = \dfrac{\sigma^2}{n}$ [this is same as the last one]

ii) $\bar{X} \sim \mathcal{N}(\mu, \dfrac{\sigma^2}{n})$ from transformation $Z \sim \mathcal{N}(0,1)$ where $Z = \dfrac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$

iii) $t \sim t_{n-1}$ where $t = \dfrac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$ and $S^2 = \dfrac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$

- Careful Here $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ is sample variance but we are thinking it as a random variable (so its value changes in repeated sampling). This makes t a random variable which is a function of random variables $\bar{X}$ and S.
- Number *iii*) says if we replace σ with S (this means replacing population standard deviation with sample standard deviation), we get a new random variable t, which follows t distribution with $n-1$ degrees of freedom. This is called *Student's t-distribution*.
- Here $n-1$ is the parameter of the t distribution. This parameter has a special name, it is called *degrees of freedom*.

Estimator $\bar{X}$, $\mathbb{E}(\bar{X})$ and $\mathbb{V}\mathrm{ar}(\bar{X})$

- Also note here $\bar{X}$ is a statistic, Z is a statistic and also $t = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ is a statistic (Recall a statistic is a function of a random sample).
- In practical cases we don't know the population standard deviation σ, so we use the sample standard deviation S to estimate σ. Since $\sqrt{\mathbb{V}\mathrm{ar}(\bar{X})} = \frac{\sigma}{\sqrt{n}}$, is the *standard error* of the sample mean $\bar{X}$. $S/\sqrt{n}$ is called the *estimate of the standard error of the sample mean*.
- Note that this result assumes stronger assumption than the last one, we are assuming population is normal.
- The next theorem will relax this assumption but we will need large n.

Estimator $\bar{X}$, $\mathbb{E}(\bar{X})$ and $\mathbb{V}\mathrm{ar}(\bar{X})$

Theorem 5.3: (Central Limit Theorem (CLT) and related results)

Let $X_1, X_2, \ldots, X_n$ be i.i.d random variables that follow *any distribution* with population mean μ and variance σ^2. Then for *large n (technically we need $n \to \infty$)*, we get following results:

$$i) \quad Z \overset{approx}{\sim} \mathcal{N}(0,1) \text{ where } Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \qquad \text{[CLT]} \tag{2}$$

$$ii) \quad \bar{X} \overset{approx}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

$$iii) \quad t \overset{approx}{\sim} \mathcal{N}(0,1) \text{ where } t = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

- This is one of the most important results in Statistics. This is called *Central Limit Theorem (CLT)*. This says if we have a large sample size, then the distribution of $\bar{X}$ will be approximately normal.
- i) says without any distributional assumption the Z statistic will be approximately normal in large samples.
- ii) says the sample mean $\bar{X}$ will be approximately normal in large samples.

Estimator $\bar{X}$, $\mathbb{E}(\bar{X})$ and $\mathbb{V}\mathrm{ar}(\bar{X})$

- iii) says the T statistic will be approximately normal in large samples. You should compare this T with t in Theorem 5.3 iii). Both are same, so in calculation there is no difference but in assumptions there is a big difference. In Theorem 5.3 iii) we are assuming the population is normal, but here we are not assuming anything about the population. In that case t statistic follows t distribution, but here the sane statistic T follows normal distribution in large samples.

Important Remarks on Sampling Distribution of $\bar{X}$

- So we understood that *the idea of the sampling distribution is a repeated sampling idea*. In real life you can only have one sample, so you can never calculate this using a sample data.
- The last three results tell us that, we can only know the sampling distribution of means under certain assumptions (in particular we need either normality or large sample size)
- If we assume normality (this means our data is normally distributed), then the distribution of the sample means is also normal and this result is for any sample size! This is called the *exact distribution*!
- If we don't assume normality for the population, then usually we have no hope, except for large n.
- The standard deviation of sampling distribution is called *standard error*! This is standard deviation, but this name is special for sampling distribution.
- In general *any function* of the random sample $X_1, X_2, \ldots, X_n$ is called a "*Statistic*", so an estimator is also a *Statistic*. The difference is Estimator is a type of Statistic where we are estimating some target! A statistic might not have any goal, it's just a function of random variables $X_1, X_2, X_3, \ldots, X_n$! The distribution of statistic is called *sampling distribution*.
- For example, $\bar{X}$, Z in Theorem 5.3 are both examples statistics but $\bar{X}$ is an estimator for μ, Z is just a statistic.
- Another example is S^2, where $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$. This is a statistic since it's a function of the random sample. And this is also an estimator for σ^2, because it is targeting population variance σ^2. Note that S^2 is just a sample variance.

Sampling Distributions

Sampling Distribution of Sample Proportions, i.e., Sampling distribution of $\bar{p}$

Sampling Distribution of $\bar{p}$

- Sampling distribution of sample proportion is just a special case of sampling distribution of sample means, except now we are considering *sample mean of Bernoulli random variables.*
- Let's first think *what is a population proportion?* Suppose we have a large population and there are certain proportions of females in this population, let's call this number p. This is the population proportion. Now we can think about taking a sample of size 10 from this population such that all the rows are independent.

	Income	Random variable
1.	1	$X_1 = ?$
2.	1	$X_2 = ?$
3.	0	$X_3 = ?$
4.	0	$X_4 = ?$
5.	1	$X_5 = ?$
6.	0	$X_6 = ?$
7.	1	$X_7 = ?$
8.	0	$X_8 = ?$
9.	1	$X_9 = ?$
10.	1	$X_{10} = ?$

Table 2: Income data

Sampling Distribution of $\bar{p}$

- Here 1 means Female and 0 means male. Again like before the left column is the observed/realized sample and in the right column we are thinking in terms of random variable $X_1, X_2, X_3, \ldots, X_n$
- In this case we can think the random variables $X_1, X_2, X_3, \ldots, X_n$ are all distributed as Bernoulli distribution with parameter p, in other words we have

$$X_i \sim \mathrm{Ber}(p) \text{ for all } i = 1, 2, 3, \ldots, n$$

- Now we can think about an *estimator for population proportion p*

$$\bar{p} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- This is an estimator because for a fixed sample, $\bar{p}$ is a fixed number, but for a random sample $\bar{p}$ is a random variable.
- Now let's apply Theorem 5.2 and 5.3. Since by assumption $X_1, X_2, X_3, \ldots, X_n$ are all independent and they all follow Bernoulli distribution with parameter p, we have

$$\mathbb{E}(\bar{p}) = p \text{ and } \mathbb{V}\mathrm{ar}(\bar{p}) = \frac{p(1-p)}{n}$$

Sampling Distribution of $\bar{p}$

- How do we get this? Same as before, note that here

$$\mathbb{E}(X_i) = p \text{ and } \mathbb{V}\mathrm{ar}(X_i) = p(1-p) \text{ for } i = 1, 2, \ldots, n$$

  then applying the rules for expectation and variance, we get

$$\mathbb{E}(\bar{p}) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n} p = p$$

$$\mathbb{V}\mathrm{ar}(\bar{p}) = \mathbb{V}\mathrm{ar}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \overset{*}{=} \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\mathrm{ar}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} p(1-p) = \frac{p(1-p)}{n}$$

- So we know that

$$\mathbb{E}(\bar{p}) = p \text{ and } \mathbb{V}\mathrm{ar}(\bar{p}) = \frac{p(1-p)}{n}$$

- Now let's talk about the sampling distribution of $\bar{p}$. In this case we can apply Theorem 5.4 i), which is a large sample result without any distributional assumption, so we get

$$\bar{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right) \text{ also using transformation } Z \sim \mathcal{N}(0,1), \text{ where } Z = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

- Here $\sqrt{\frac{p(1-p)}{n}}$ is the *standard error* of $\bar{p}$, or you can say this is the standard deviation of the sampling distribution of $\bar{p}$.

Sampling Distribution of $\bar{p}$

- Now the issue with the Z statistic is that we don't know the population proportion p. So we can't use Z statistic in practice, solution? use the sample proportion $\bar{p}$ to estimate p, and then construct the t statistic.
- In his case we can apply Theorem 5.4 iii), which is a large sample result without any distributional assumption and also without assuming we know p, so we can form the t statistic

$$T = \frac{\bar{p} - p}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}} \sim \mathcal{N}(0, 1)$$

- You should compare this with the t statistic in Theorem 5.4 (iii), this is exactly same, now we are using $\bar{p}$ instead of $\bar{X}$, and we are using $\sqrt{\bar{p}(1-\bar{p})}$ instead of S.
- In ECO204 you will see t statistic many times, and all of them will follow the same structure,

$$t = \frac{\text{estimator} - \mathbb{E}\left(\text{estimator}\right)}{\left(\text{estimate of the standard error}\right)}$$

- And often in large samples (at least the cases that you will encounter), the distribution of the t statistic will become standard normal. Hence we can use the standard normal distribution to calculate the probability of the t statistic.
- More on this on ECO204...

References