0527 Slides

MA 116

May 2025

Random sample is a sample choosen from a population at random.
Simple random sample is a random sample such that each individual in the population has an equal chance of being chosen.

In this course, the two terms are used interchangeably.

Fix a population of size N with population mean μ and population standard deviation σ. Fix a sample size n. Consider the random variable $\overline{x}$ of sample mean of random samples of size n.

Theorem

the random variable $\overline{x}$ has mean

$$\mu_{\overline{x}} = \mu$$

regardless of n or N. If $n < 0.05N$, the variable $\overline{x}$ has a standard deviation

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}.$$

$\sigma_{\overline{x}}$ is also called the **standard error of the mean**. Note that $\overline{x}$ as a random variable depends on n, while its mean $\mu_{\overline{x}}$ is independent of n, its standard deviation $\sigma_{\overline{x}}$ does depend on n.

Describe the probability distribution of $\overline{x}$ as a variable

Theorem.

If a random variable x is (approximately) normally distributed, then the probability distribution of $\overline{x}$ would also be approximately normal. This result is independent of n.**This result does not require that n < 0.05N.**

In the case that the distribution of x is not normal, we have the following theorem.

Central Limit Theorem

If n < 0.05N, as n increases, the distribution of $\overline{x}$ becomes more and more normal. When $n \geq 30$, we claim that the bell curve is a good approximation of the distribution of $\overline{x}$, i.e. the distribution of $\overline{x}$ is approximately normal.

RMK. Theorem 1 holds regardless of the distribution of x (normal or not).

Fix a population of size N and a characteristic. Let p be the population portion of this characteristic in this population.

Let $\mu_{\hat{p}}$ denote the mean of the sample proportion, and let $\sigma_{\hat{p}}$ denote the standard deviation of the sample proportion. Note that $\sigma_{\hat{p}}$ depends on our choice of n, but this dependency is not reflected in its notation.

Theorem. Assume that any sampling is random.

$\mu_{\hat{p}} = p$. This holds regardless of n or N.

Theorem. Assume $n \leq 0.05N$ and that any sampling is random

1. As n increases, the shape of the distribution of the sample proportion becomes approximately normal. When $np(1-p) \geq 10$, we say the bell curve is a good approximation of the distribution of $\hat{p}$.
2. $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$.

Summary of CH.6

Fix a population and a sample size n. Assume all samplings are random.

Theorem		need $n < 0.05N$?
1	$\mu_{\overline{x}} = \mu$	NO
2	$\sigma_{\overline{x}} = \sigma/\sqrt{n}$	YES
3	x being approximately normal implies $\overline{x}$ is approximately normal	NO
4	CLT: $n \geq 30$ implies $\overline{x}$ is approximately normal	YES
5	$\mu_{\hat{p}} = p$	NO
6	$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$	YES
7	$np(1-p) \geq 10$ implies $\hat{p}$ is approximately normal	YES

Why do we want to describe the distribution of $\hat{p}$

Suppose we obtain 1000 simple random samples of size 20 and analysis this data set to estimate the distribution of $\hat{p}$.

If our only purpose is to estimate p, then instead of obtaining **1000 simple random samples of size 20**, we may simply obtain a big simple random sample of a large enough size n. Then calculate $\hat{p} = \frac{b}{n}$ for this big sample, where b is the number of individuals in this sample with that characteristic. We may then claim that $p \cong \hat{p}$.

Being able to describe the distribution of $\hat{p}$ is more powerful than this. A description of the distribution of $\hat{p}$ contains more information than its mean $\mu_{\hat{p}} = p$, so it does more than estimating the population proportion p.

Example 2.2

Question.

According to NHS, 15% of all Americans have hearing trouble. In a random sample of 120 Americans, what is the probability at most 12% have hearing trouble?

Answer. We are looking for $P(\hat{p} \leq 12\%)$, where $\hat{p}$ is sample proportion with sample size 120, viewed as a variable.

Step I. (i). 120 is surely less than $0.05\times$American population;
(ii) $np(1-p) = 120 \cdot 0.15 \cdot 0.85 = 15.3 \geq 10$. Then, we may say the distribution of $\hat{p}$ is approximately normal.

Step II. $\mu_{\hat{p}} = p = 0.15$ holds regardless of n or N. $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$ holds since $n < 0.05N$. Then $\sigma_{\hat{p}} = \sqrt{0.15 \cdot 0.85/120} = 0.03$.

Step III. Let $z = \dfrac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \dfrac{\hat{p} - 0.15}{0.03}$.

Step IV. We realize that $P(\hat{p} \leq 12\%) = P(z \leq -1)$ via the formula

Suppose x is a normal variable with mean μ and standard deviation σ. Let $z = \dfrac{x - \mu}{\sigma}$. Then:

$$P(x \leq a) = P(z \leq \frac{a - \mu}{\sigma})$$

$$P(x \geq a) = P(z \geq \frac{a - \mu}{\sigma})$$

for any number a.

by taking $a = 0.12$, normal variable to be $\hat{p}$.

Step V. Recall properties of a standard normal distribution (or look up the SND Table), we realize that $P(z \leq -1) = 16\%$.

Step VI. Conclusion: In a random sample of 120 Americans, the probability at most 12% have hearing trouble is about 16%.

More examples

Example. Note population is again not a quantitative data set.

Suppose the true fraction of all US citizens who trust the president is $p = 0.46$. Can you describe the sampling distribution of $\hat{p}$ with a sample size $n = 100$?

Step I. (i). 100 is surely less than $0.05 \times$American population;
(ii) $np(1-p) = 100 \cdot 0.46 \cdot 0.54 = 24.84 \geq 10$. Then, we may say the distribution of $\hat{p}$ is approximately normal.
Step II. $\mu_{\hat{p}} = p = 0.46$ as always.
Step III. Since $n < 0.05N$ holds, we have that

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.46 \cdot 0.54}{100}} = 0.05.$$

Step IV. Conclusion: The sampling distribution of $\hat{p}$ with a sample size $n = 100$ is approximately normal with a mean at 0.46 and a standard deviation about 0.05.

Bell curve approximation. Any sample proportion is $0 \leq \hat{p} \leq 1$. When saying the distribution of some $\hat{p}$ can be approximated by a bell curve $f(\hat{p})$, this model, as a bell curve, will always be nonzero in the region $\hat{p} > 1$ and $\hat{p} < 0$. This is an example of discrepancy between model and reality.

Example. Calculating $z_{0.025}$.

Step I. Definition of $z_{0.025}$?
Step II. Draw a standard normal curve and label what 0.025 and $z_{0.025}$ mean in the diagram.
Step III. Look at the Standard Normal Distribution Table and find this $z_{0.025}$.

Interpret this $z_{0.025}$

We have $z_{0.025} = 1.96$. From the diagram, the area between -1.96 and 1.96 is 0.95.
Interpretation: Suppose a quantitative population follows the standard normal distribution. If we pick a data point a from the population at random, the possibility that $-1.96 \leq a \leq 1.96$ is 95%.

	Population Parameter	Sample Statistic
Mean	μ	$\overline{x}$
Median	η	M
Variance	σ^2	s^2
Standard Deviation	σ	s
Proportion	p	$\hat{p}$

Oftentimes we want to estimate those population parameters. Let's make **estimation** more precise.

Plan of Part II of class

We are going to skip Section 9.3 (estimating σ), Section 10.4 (hypothesis test for σ), and Section 11.4 (inference about two σ), and focus on μ and p. After that, we will study Ch.14 Regression (Part III of class).

Suppose we have a population, and we want to estimate a population parameter of this population.

The simplest type of statistic used to make inferences about a population parameter is a **point estimator**.

Definition. (point estimator.)

A point estimator of a population parameter is a formula that tells us how to use a sample data set to calculate a sample statistic that can be used as an estimate of the population parameter.

Example. A population consists of ages of 10000 people. To estimate μ, we obtain a random sample of size 10 and calculate $\overline{x} = \dfrac{a_1 + \ldots + a_{10}}{10}$ where my sample $\{a_1, \ldots, a_{10}\}$ are ages of 10 randomly picked individuals. We say $\overline{x}$ is a point estimator of μ.

Fix a population, population parameter, pick a sample, then the point estimator obtained from this sample is a single number.

Example. A population consists of all residents in Boston and whether they have a full time job or not. Let p be the population proportion of having a full-time job. To estimate p, we obtain a random sample of size 3 and calculate $\hat{p} = \frac{2}{3}$ where my sample {Yes, No, Yes} consists of the full-time work status of 3 randomly picked residents of Boston. We say $\hat{p} = \frac{2}{3}$ is a point estimator pf p, the proportion of people having a full-time job.

Example. A population consists of ages of all residents in Boston. To estimate μ, we obtain a random sample of size 2 and calculate $\overline{x} = \frac{3+4}{2} = 3.5$ where my sample {3, 4} are ages of 2 randomly picked individuals. We say $\overline{x} = 3.5$ is a point estimator of μ.

Examples show that a point estimator of a population parameter can sometimes differ significantly from the true value of the population parameter.

Let's explore ways to determine how large the difference between a point estimate and the true value of the parameter is likely to be–**use sampling distribution of a sample statistics**.

Unbiased and Biased Estimators

If the sample distribution of a sample statistics has a mean equal to the population parameter which the statistics is intended to estimate, the statistics is said to be an unbiased estimator of the parameter. Otherwise, the statistics is said to be a biased estimator of the parameter.

For the two statistics $\overline{x}$ and $\hat{p}$ we care about, both are unbiased estimator of the parameter they intended to estimate:

$$\mu_{\overline{x}} = \mu,\ \mu_{\hat{p}} = p,$$

as long as our sampling is random.

Example of a biased estimator

This topic will not be tested.

Sample variance

Let $\{x_1, ..., x_n\}$ be a sample from a population with population mean μ and population variance σ^2. Recall the formula of sample variance is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

We may check that $E[s^2] = \sigma^2$, so s^2 defined this way is an unbiased estimator of the population variance σ^2.

This topic will not be tested.

However, if one attempts to define Sample Variance to be

$$r^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2,$$

they would end up with $E[r^2] = \frac{n-1}{n}\sigma^2 \neq \sigma^2$. This formula would **underestimate** the true population variance σ^2 because $\frac{n-1}{n}$ for any positive integer n.

We say r^2 is a biased estimator of the population variance σ^2.

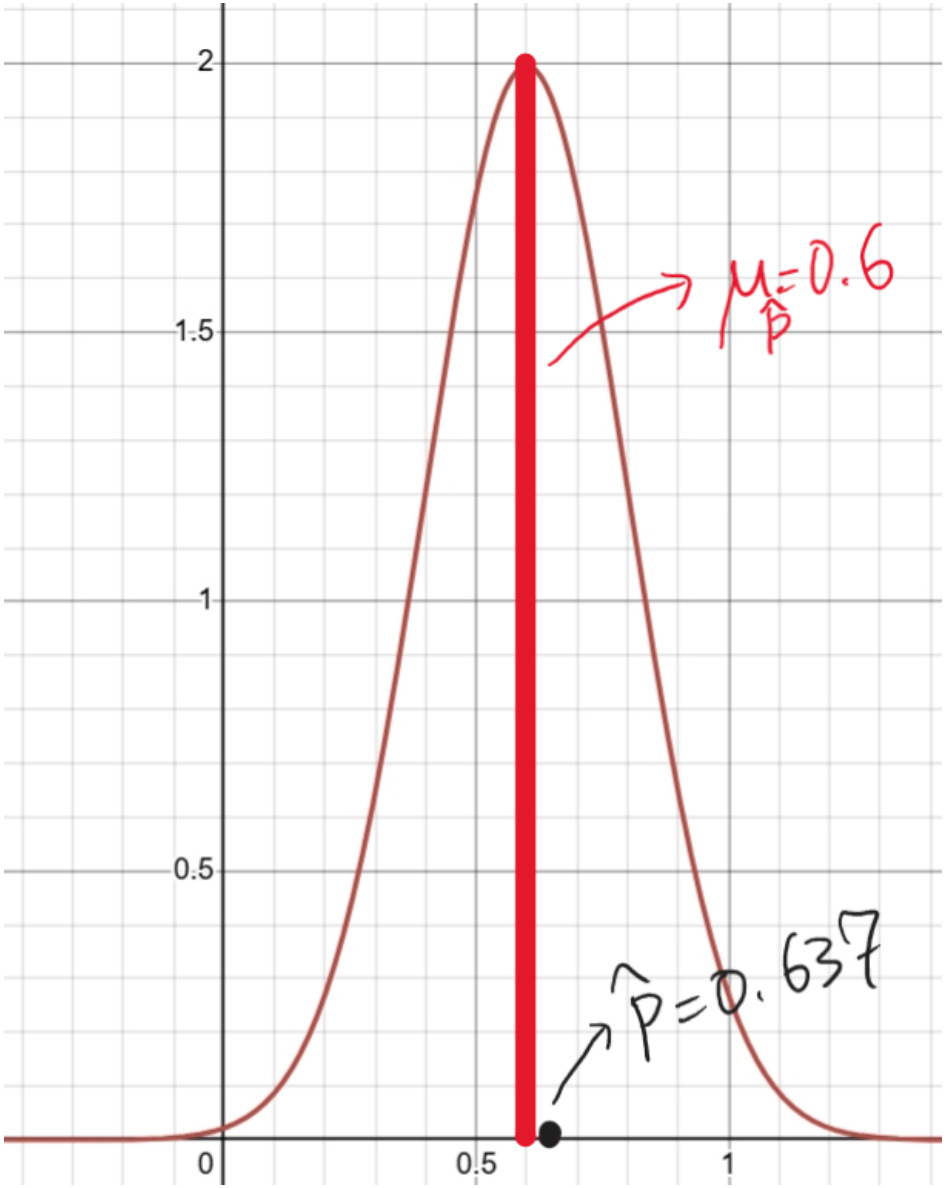
However, we can still say that r^2 is a point estimator of the population variance σ^2.

Population proportion

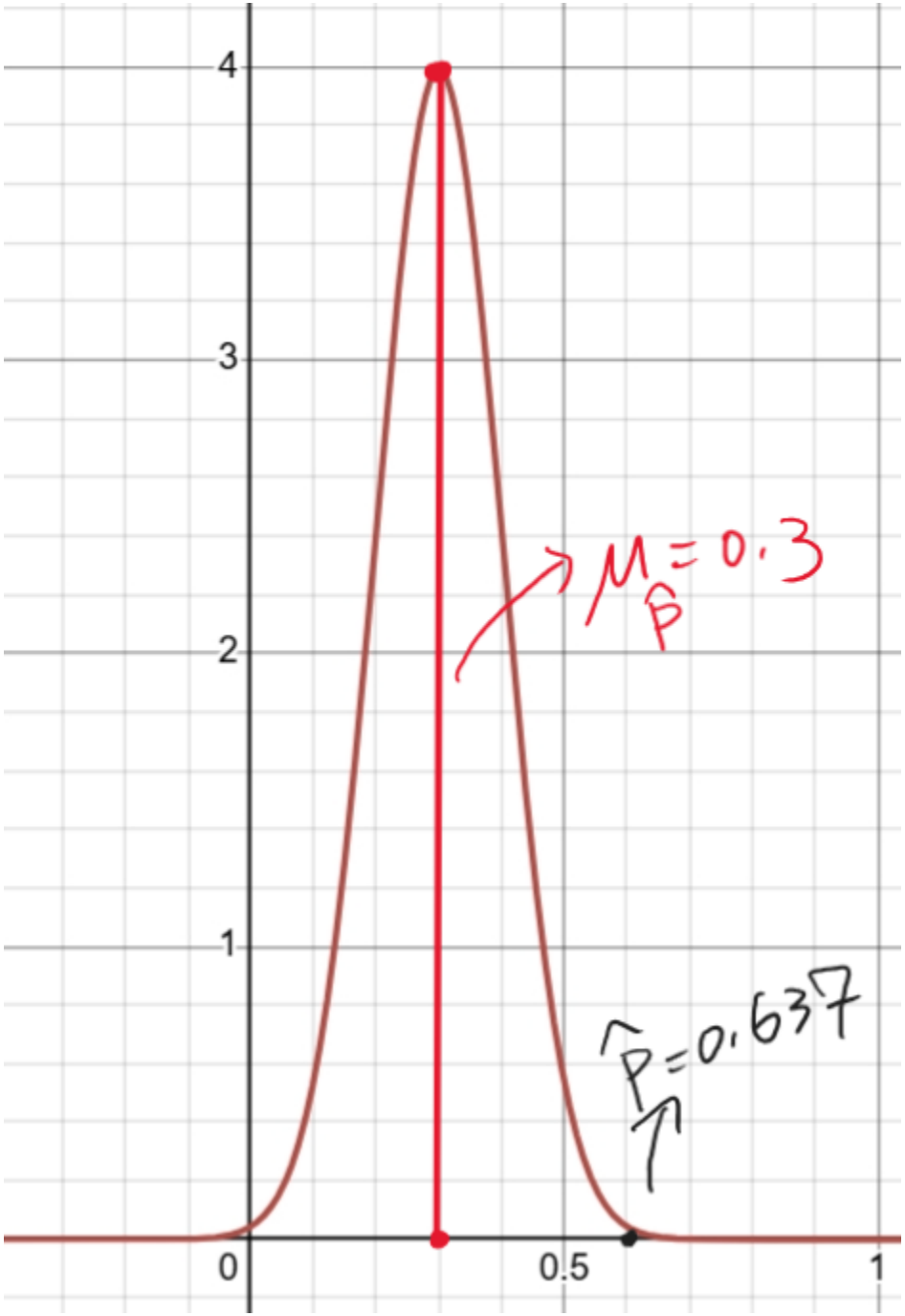
Suppose 1000 people are randomly chosed from all US citizens and 637 answer that they trust the president. How would you estimate the true fraction of all US citizens who trust the president?

We first calculate a point estimator $\hat{p} = 0.637$ of the population proportion. How reliable is this point estimator?

To answer this question, let's make use of the sampling distribution of the sample statistics $\hat{p}$: If we have a discription of the sampling distribution of the sample statistics $\hat{p}$ (for example, a discription might be that it's approximately normally distributed with some mean $\mu_{\hat{p}}$ and standard deviation $\sigma_{\hat{p}}$), we would know where this specific sample proportion 0.637 lies in the distribution.



If we know the sampling distribution of $\hat{p}$ can be approximated by this curve, then the particular sample proportion $\hat{p} = 0.637$ we get seems to be a good estimation of p.



If we know the sampling distribution of $\hat{p}$ can be approximated by this curve, then the particular sample proportion $\hat{p} = 0.637$ we get seems to be a bad estimation of p.

Let's recall theorems that may help us describe a sampling distribution of $\hat{p}$. Suppose all samplings are random.

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

When $n < 0.05N$, the formula

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

holds.

When $n < 0.05N$ and $np(1-p) \geq 10$, the distribution of $\hat{p}$ is approximately normal.

Confidence interval of a point estimator

Fix a population parameter p. We obtain a particular sample of size n and calculate its sample proportion $\hat{p} = \frac{b}{n}$. Suppose $n < 0.05N$ and $n\hat{p}(1-\hat{p}) \geq 10$. Fix some α that lies between 0 and 1. Consider the interval

$$\left[\hat{p} - z_{\alpha/2}\sigma_{\hat{p}},\ \hat{p} + z_{\alpha/2}\sigma_{\hat{p}}\right].$$

This is called a $(1-\alpha)100\%$ confidence interval of p, or a confidence interval of p with a level of confidence $(1-\alpha)100\%$.

By our conditions we know that $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$. From mathematical experience, this formula implies that $\sigma_{\hat{p}}$ is **insensitive** to changes of p. Thus, in practice, we use

$$\sigma_{\hat{p}} \cong \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

We say a $(1-\alpha)100\%$ confidence interval of p (about $\hat{p}$) is

$$\left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Example. Suppose we would like to get a 95% confidence interval for p in the setting of

1000 people are randomly chosed from all US citizens and 637 answer that they trust the president.

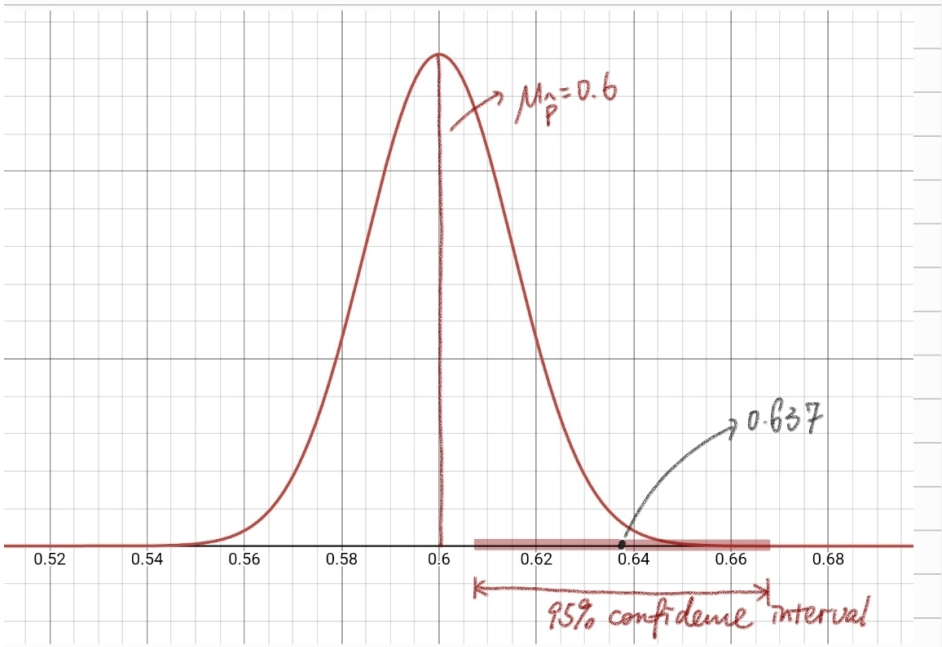
Let's first figure out what our α is. $(1-\alpha)100\% = 95\%$ implies that $\alpha = 0.05$, so $\alpha/2 = 0.025$. SND Table tells us that $z_\alpha = z_{0.025} = 1.96$.

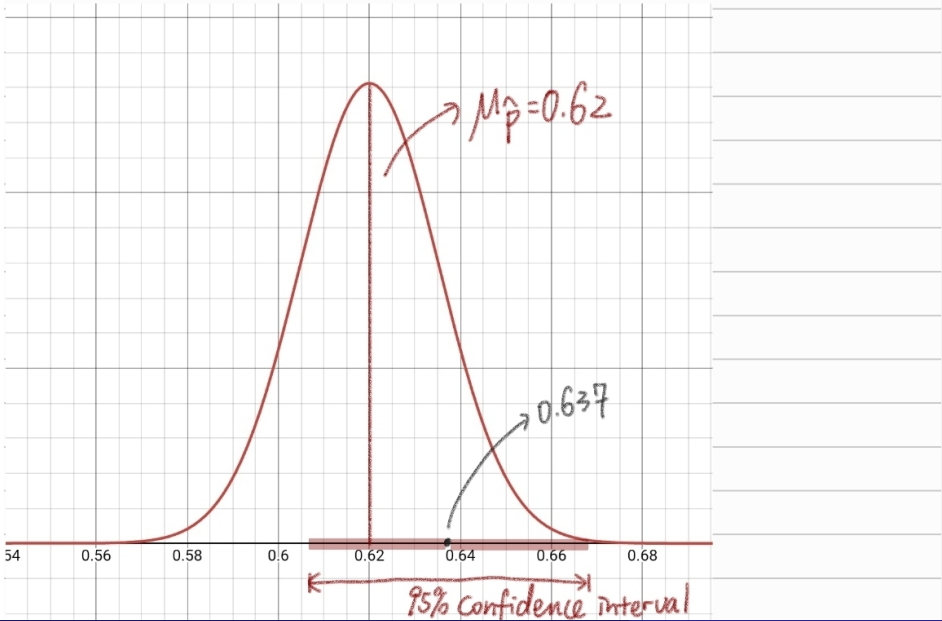
Then $\sigma_{\hat{p}} \cong \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \cong 0.0152$, so our 95% confidence interval of p is

$$[0.637 - 1.96 \cdot 0.0152,\ 0.637 + 1.96 \cdot 0.0152]$$

which is about [0.607, 0.667].

Interpret this interval





A 95% confidence interval indicates that 95% of all random samples of size n from the population, whose parameter p we want to estimate, will results in an interval $\left[\hat{p} - z_{0.05/2}\sigma_{\hat{p}},\ \hat{p} + z_{0.05/2}\sigma_{\hat{p}}\right]$ that contains the parameter p.

Interpret this interval

Strictly speaking, a 95% confidence interval

$$\left[\hat{p} - z_{0.05/2}\sigma_{\hat{p}},\ \hat{p} + z_{0.05/2}\sigma_{\hat{p}}\right]$$

obtained from a particular sample does not imply that there is a 95% probability that p lies in this interval. This is because p is assumed to be a fixed value (intrinsic to our population and does not depend on specific sample chosen), not a random value. So saying "there is a 95% probability that p lies in this interval" makes no sense.

From the illustrations, we see that given a particular sample with some $\hat{p}$, we do not know if p lies in the 95% confidence interval about this $\hat{p}$.

We do not know if the random sample we obtained is one of the 95% samples whose interval contain p, or not.

Margin of error

Definition. (Margin of error.)

The margin of error, E, in a $(1-\alpha)100\%$ confidence interval for a population proportion is given by

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

I.e. the $(1-\alpha)100\%$ confidence interval of a particular sample with some $\hat{p}$ is $[\hat{p}-E,\ \hat{p}+E]$.

Confidence interval is also called an **interval estimator**.

Suppose now we want to conduct a survey to estimate a population proportion p by an interval estimator, but how do we pick our sample size?

Sample size too small

Example. A population consists of all residents in Boston and whether they have a full time job or not. Let p be the population proportion of having a full-time job. To estimate p, we obtain a random sample of size 3 and calculate $\hat{p} = \frac{2}{3}$ where my sample {Yes, No, Yes} consists of the full-time work status of 3 randomly picked residents of Boston. A 90% confidence interval of p obtained from this sample would be

$$\left[2/3 - 1.645 \cdot \sqrt{\frac{2}{27}},\ 2/3 + 1.645 \cdot \sqrt{\frac{2}{27}}\right], \text{ about } [0.219, 1.115].$$

Issue. A small sample size n would lead to a big margin of error E, in which case the resulting 90% confidence interval tells us little information.

In the previous example we get a 90% confidence interval of p of $[0.219, 1.115]$, which is awful.

Law of Large Numbers (roughly says)

Large sample sizes produce more precise estimates.

Suppose now we want to conduct a survey to estimate a population proportion p by an interval estimator, and we want to ensure that our margin of error E is up to a specific value 0.03. How do we determine our sample size?

Suppose now we want to conduct a survey to estimate a population proportion p by an interval estimator, and we want to ensure that our margin of error E is up to a specific value $E' = 0.03$. How do we determine our sample size?

Method 1.

Suppose we have knowledge of a **prior point estimator** of p, denoted $\tilde{p}$. We may claim that the sample size required to obtain a $(1-\alpha)100\%$ confidence interval for p with a margin of error up to E' is given by

$$n = \tilde{p}(1-\tilde{p})\left(\frac{z_{\alpha/2}}{E'}\right)^2$$

rounded up to the next integer.

Example. We want to estimate the true fraction of all US citizens who trust the president on 05/27, but we have a point estimator 0.457 from a survey conducted on 05/26. We can take $\tilde{p} = 0.457$.

Method 2.

We take the mathematical maximum of $y(1-y)$ over $0 \leq y \leq 1$, which is 0.25. We may claim that the sample size required to obtain a $(1-\alpha)100\%$ confidence interval for p with a margin of error up to E' is given by

$$n = 0.25 \left(\frac{z_{\alpha/2}}{E'} \right)^2$$

rounded up to the next integer.

This method may lead to a larger sample size than is necessary. In practice, a prior point estimator is often available, in which case we prefer Method 1. over Method 2.

Example. (Method 1.)

We want to estimate the true fraction of all US citizens who trust the president on 05/27, but we have a point estimator 0.46 from a survey conducted on 05/26. We can take $\tilde{p} = 0.46$. What is a sample size required to obtain a 95% condfidence interval for p with a margin error up to $E' = 0.03$?

Step I. Calculate $z_{0.05/2}$.
Step II. Calculate

$$n = \tilde{p}(1 - \tilde{p}) \left(\frac{z_{\alpha/2}}{E'} \right)^2 .$$

Step III. Round the resulting n **up** to the next integer!

Example. (Method 2.)

We want to estimate the true fraction of all US citizens who trust the president on 05/27, but we have no prior estimator. What is a sample size required to obtain a 95% condfidence interval for p with a margin error up to $E' = 0.03$?

Step I. Calculate $z_{0.05/2}$.
Step II. Calculate

$$n = 0.25 \left(\frac{z_{\alpha/2}}{E'} \right)^2 .$$

Step III. Round the resulting n **up** to the next integer!

The resulting size we get from method 2 is always larger than or equal to the resulting size we get from method 1.

Reference. The example

> 1000 people are randomly chosed from all US citizens and 637 answer that they trust the president.

is taken from *Statistics*, 13th edition, by McClave and Sincich, page 337-338.