0528 Slides

MA 116

May 2025

Review1

Suppose x is a normal variable with mean $\mu = 1$ and standard deviation σ. Given $P(x \leq 0) = 0.1$.

1. What is $P(x \geq 0)$?
2. What is $P(x \geq 2)$?

Review2

This change of variable method is used to calculate the probability of a normal variable x lies in a specific region $x \leq a$ or $x \geq a$ by relizing this event it the same as the event $z \leq \frac{a-\mu}{\sigma}$ or $z \geq \frac{a-\mu}{\sigma}$ for a standard normal variable z.

Change of variable method

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

Change of variable method

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.

Suppose x is a normal variable with a mean $\mu = 4$ and a standard deviation $\sigma = 2$. Calculate $P(x \geq 7)$.

Review3

Suppose a quantitative population can be approximated by the standard normal distribution. Let's associate a variable z to a random individual in this population.

1. What is $P(z \geq 1.23)$?
2. What is $P(-0.91 \leq z \leq 2.01)$?
3. What is $P(z \leq 3)$?

Review4: Calculate z_α

$-z_\alpha = z_{1-\alpha}$

1. Given a value $0 \leq \alpha \leq 1$. What is z_α?
2. Give a brief reasoning to justify the above formula.
3. Find $z_{0.5}$.
4. Find $z_{0.1}$.
5. Find $z_{0.9}$.
6. Find $z_{0.0228}$.

Summary of CH.8

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

Review5

Suppose the daily phone use time among BU students can be approximated by a bell curve with a mean 40 minutes and a standard deviation 36 minutes. (Population size is by assumption very large.) Fix a sample size $n = 16$ and consider the random variable $\overline{x}$, the sample mean of random samples of size 10.

1. What is $\mu_{\overline{x}}$?
2. What is $\sigma_{\overline{x}}$?
3. Is $\overline{x}$ approximately normally disributed?

Review 6

Suppose the daily phone use time among BU students can NOT be approximated by a bell curve. The random variable x associated to this population has a population mean 40 minutes and a population standard deviation 36 minutes. (Population size is by assumption very large.) Fix a sample size $n = 36$ and consider the random variable $\overline{x}$, the sample mean of random samples of size 36.

1. Is $\overline{x}$ approximately normally disributed? Give a brief justification.
2. Given $\mu_{\overline{x}} = 40$ and $\sigma_{\overline{x}} = 6$, use change of variable to calculate the possibility that a random sample of size 36 has a sample mean less than 34.

Review 7

Suppose 70% of US citizens in a state voted in the last Federal election. Fix a sample size $n = 100$ and consider the sample proportion variable $\hat{p}$.

1. Is $\hat{p}$ normally distributed?
2. What is $\mu_{\hat{p}}$?
3. We want to know if we obtain a random sample of 100 US citizens in this state, what's the possibility that less than 50% of the 100 individuals voted. We are looking for (choose one option and provide a brief explanation)
    1. $P(\hat{p} < 50\%)$
    2. $P(p < 50\%)$
    3. $P(z < 50\%)$
    4. $P(p = 70\%)$
4. Given $\sigma_{\hat{p}} = 0.05$. Let $z = \dfrac{\hat{p} - 0.7}{0.05}$. The probability we are looking for is the same as $P(z < c)$. Calculate the number c.

Example.

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.

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 $z_{0.025}$

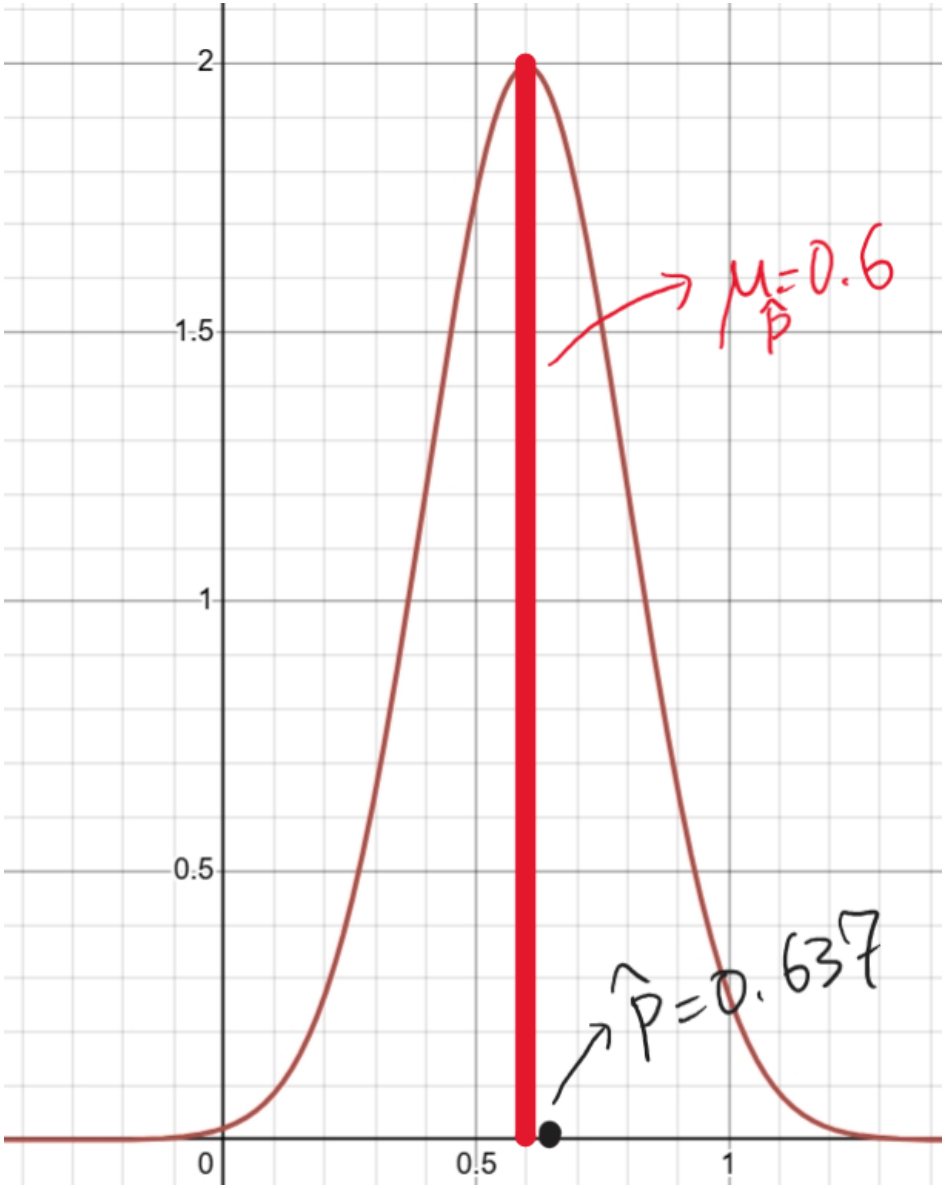
We have $z_{0.025} = 1.96$. From the diagram, the area between -1.96 and 1.96 is 0.95.
Interpretation: Suppose a quantitative population follows the standard normal distribution. If we pick a data point a from the population at random, the possibility that $-1.96 \leq a \leq 1.96$ is 95%.

Population proportion

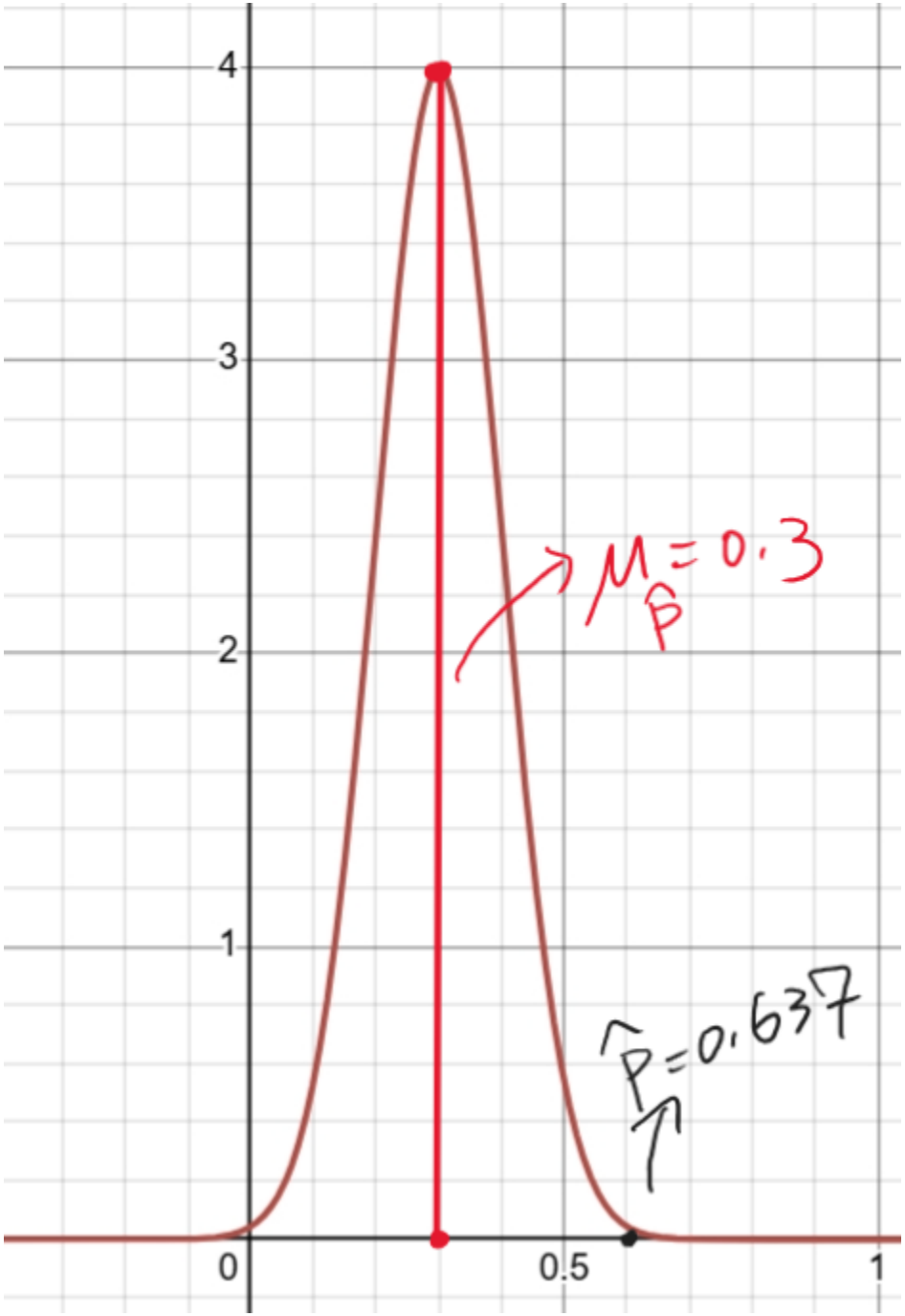
Suppose 1000 people are randomly chosed from all US citizens and 637 answer that they trust the president. How would you estimate the true fraction of all US citizens who trust the president?

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.

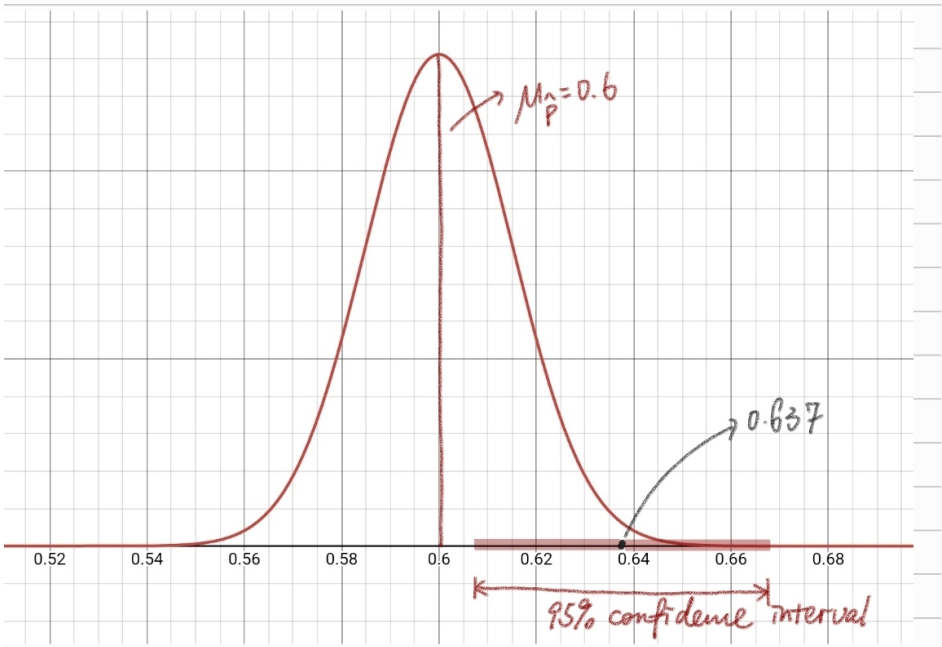
Let's first figure out what our α is. $(1-\alpha)100\% = 95\%$ implies that $\alpha = 0.05$, so $\alpha/2 = 0.025$. SND Table tells us that $z_\alpha = z_{0.025} = 1.96$.

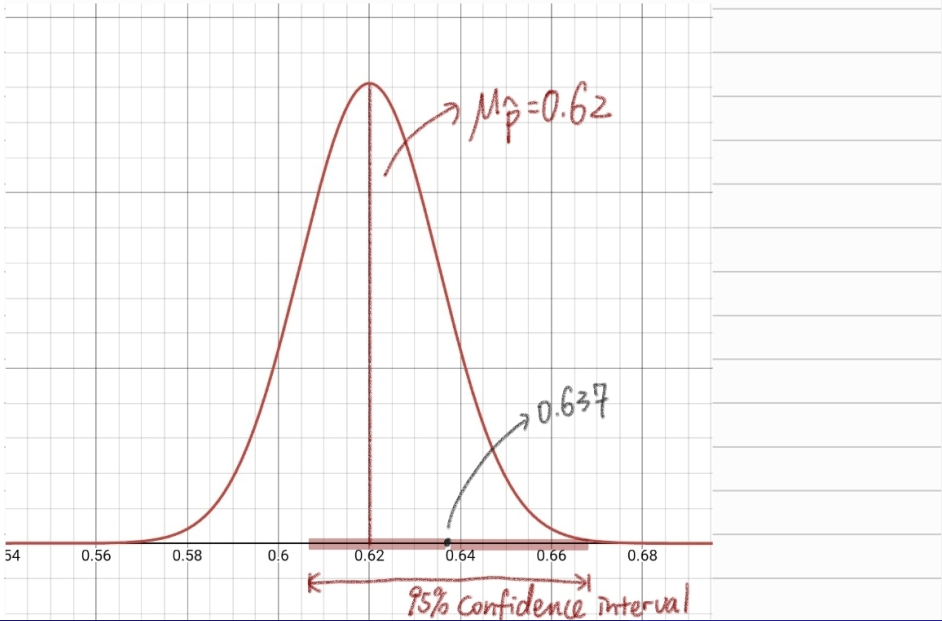
Then $\sigma_{\hat{p}} \cong \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \cong 0.0152$, so our 95% confidence interval of p is

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

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.

Example.

Consider the population proportion p associated to the percentage of US citizens who voted in the last Federal election. Suppose we have an inverval estimator $[0.3, 0.5]$ of this parameter obtained from a particular sample of size 200.

1. What is the point estimator of this interval estimator?
2. What is the margin of error?
3. What is the number of people who voted in this particular sample?

Estimating a population mean

Idea is similar to estimating a population proportion.
To estimate a population mean, we may obtain a particular sample and calculate its sample mean $\overline{x}$. This value serves as a point estimator of population proportion.

Like before, we want to define an interval estimator of the form $\overline{x} \pm E$, where E is the margin of error.

Fix a level of confidence α. Our guess of definition of E to a level of confidence α would be

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

where σ is the population standard deviation and n is the size of a particular sample.

Issue of not knowing σ

Like in the case of estimating p, we do not know σ. In the case of estimating p we solve this issue by claiming

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

Does this work for estimating $\overline{x}$?

If we tentatively define $E = z_{\alpha/2}\frac{s}{\sqrt{n}}$ where s is the standard deviation of a particular sample, we'll encounter a big issue that the true parameter μ does not lie in our interval $\overline{x} \pm E$, $E = z_{\alpha/2}\frac{s}{\sqrt{n}}$ at a rate that's acceptable. (Gosset, p.464.)

To solve this, let's introduce a new random variable t. We define

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}.$$

t is a new variable constructed from the variable $\overline{x}$ just like how we construct z from a change of variable.

We use this variable t to give a better definition of the margin of error E of μ.

Properties of the variable t and its distribution

If the original population is normally distributed, the variable t follows the **Student's t-distribution** with $n-1$ degrees of freedom.

1. The Student's t-distribution is centered about 0.
2. This is a good probability density function.
3. Approaching 0 at two ends.
4. Very similar to the standard normal curve! except at the two tails.

t_α

Sample size $=n$. Define $E = t_{\alpha/2}\dfrac{s}{\sqrt{n}}$ where $t_{\alpha/2}$ is with $n-1$ degrees of freedom.

When $n \geq 30$, we say t converges in distribution to the standard normal distribution.

Why?

Law of large number implies that s approaches σ.