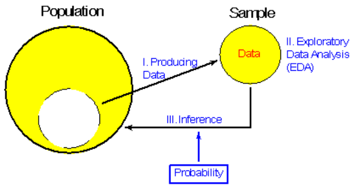
Confidence Intervals

Chapter 7

Inference

- We are in the fourth and final part of the course - statistical inference, where <u>we draw conclusions about the population based on the data obtained from a sample chosen from it</u>.



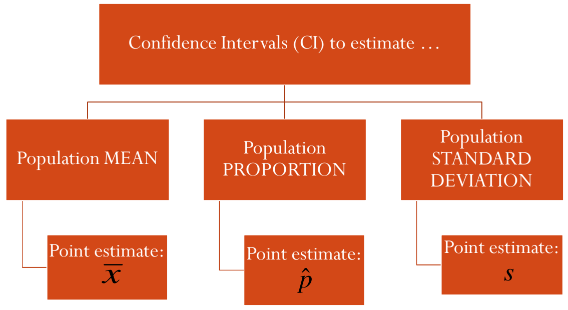
Our Goal in Inference

We wish to know the population parameter

If ALL the populations, whatever we are interested in, would be manageable in size, we would just figure out the population parameter. Then there would be no need for inference.

Confidence Intervals (CI)

- **<u>The goal: to give a range of plausible values for the estimate of the unknown population parameter</u>**
  - **the population mean, μ,**
  - **the population proportion, p**
  - **the population standard deviation, σ**
- **We start with our best guess: the sample statistic**
  - **the sample mean $\bar{x}$,**
  - **the sample proportion $\hat{p}$**
  - **the sample standard deviation, s**
- **Sample statistic = point estimate**



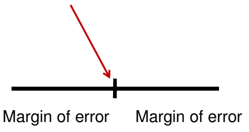
Confidence Intervals (CI) to estimate …

Population MEAN	Population PROPORTION	Population STANDARD DEVIATION
Point estimate: $\bar{x}$	Point estimate: $\hat{p}$	Point estimate: s

---

Confidence Intervals (CI)

CI = point estimate ± margin of error



Margin of error

- **Shows how accurate we believe our estimate is**
- **The smaller the margin of error, the more precise our estimate of the true parameter**
- **Formula:**

$$E = \begin{pmatrix}\text{critical}\\ \text{value}\end{pmatrix} \cdot \begin{pmatrix}\text{standard deviation}\\ \text{of the statistic}\end{pmatrix}$$

Confidence Intervals (CI) for a Mean

- **Suppose a random sample of size n is taken from a normal population of values for a quantitative variable whose mean μ is unknown, when the population's standard deviation σ is known.**
- A confidence interval (CI) for μ is:

CI = point estimate ± margin of error

$$\underset{\textbf{Point estimate}}{\bar{x}} \pm \underbrace{z^* \cdot \frac{\sigma}{\sqrt{n}}}_{\textbf{Margin of error (m or E)}}$$

So what's z*???

- A confidence interval is associated with a confidence level. We will say: **"the 95% confidence interval for the population mean is …"**
- The most common choices for a confidence level are
  - 90% :z* = 1.645
  - 95% : z* = 1.96,
  - 99% : z* = 2.576.

Statement: (memorize!!)

We are _________% confident that the true mean *context* lies within the interval _______ and _______.

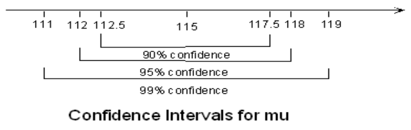
Using the calculator

- **Calculator:** STAT→TESTS →7:ZInterval…
  - Inpt: Data Stats

Use this when you have data in one of your lists

Use this when you know $\bar{x}$ and σ

The Trade-off



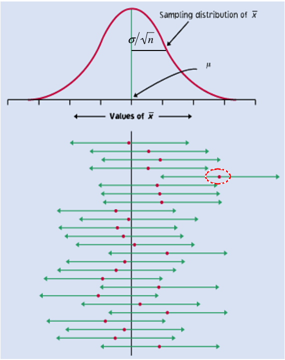
Confidence Intervals for mu

- ***There is a trade-off between the level of confidence and precision in which the parameter is estimated.***
- **higher level of confidence – wider confidence interval**
- **lower level of confidence – narrower confidence interval**

95% confident means:

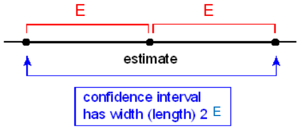
In 95% of all possible samples of this size *n*, μ will indeed fall in our confidence interval.

In only 5% of samples would miss μ.



The Margin of Error

- The width (or length) of the CI is exactly twice the margin of error (E):



- The margin of error is therefore "in charge" of the width of the confidence interval.

Comment

- The margin of error (E) is

$$\boxed{E = z^* \cdot \frac{\sigma}{\sqrt{n}}}$$

and since *n*, the sample size, appears in the denominator, **increasing *n* will reduce the margin of error** for a fixed *z**.

How can you make the margin of error smaller?

- **z* smaller**
  (lower confidence level)
- **σ smaller**
  (less variation in the population)
- **n larger**
  (to cut the marg t
  be 4 times as big

▪Really cannot change!

Margin of Error and the Sample Size

- In situations where a researcher has some flexibility as to the sample size, the **researcher can calculate in advance what the sample size is that he/she needs in order to be able to report a confidence interval with a certain level of confidence and a certain margin of error.**

Calculating the Sample Size

$$E = z^* \cdot \frac{\sigma}{\sqrt{n}} \longrightarrow \boxed{n = \left(z^* \cdot \frac{\sigma}{E}\right)^2}$$

Clearly, the sample size *n* must be an integer.
Calculation may give us a non-integer result.
In these cases, we should always
round up to the next highest integer.

Example

- IQ scores are known to vary normally with standard deviation 15. How many students should be sampled if we want to estimate population mean IQ at 99% confidence with a margin of error equal to 2?

$$n = \left(z^* \frac{\sigma}{E}\right)^2 = \left(2.576\frac{15}{2}\right)^2 = 373.26 \quad \Rightarrow \quad n = 374$$

They should take a sample of 374 students.

Assumptions for the validity of $\bar{x} \pm z^* \cdot \frac{\sigma}{\sqrt{n}}$

- The sample must be random
- The standard deviation, σ, is known
- and either
  - the sample size must be large ($n \geq 30$) or
  - for smaller sample the variable of interest must be normally distributed in the population.

Steps to follow

1. **Check conditions:** SRS, σ is known, and either n ≥ 30 or the population distribution is normal
2. **Calculate** the CI for the given confidence level
3. **Interpret** the CI

Example 1

- A college admissions director wishes to estimate the mean age of all students currently enrolled. In a random sample of 20 students, the mean age is found to be 22.9 years.

Form past studies, the standard deviation is known to be 1.5 years and the population is normally distributed. Construct a 90% confidence interval of the population mean age.

Step 1: Check conditions

- A college admissions director wishes to estimate the mean age of all students currently enrolled. In a **random sample** of 20 students, the mean age is found to be 22.9 years. Form past studies, the **standard deviation is known** to be 1.5 years and the population is **normally distributed**.

- SRS ✓
- σ is known ✓
- The population is normally distributed ✓

Step 2: Calculate the 90% CI using the formula

$$\bar{x} = 22.9$$
$$\sigma = 1.5$$
$$n = 20$$
$$z^* = 1.645$$

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} = 22.9 \pm 1.645 \frac{1.5}{\sqrt{20}} = 22.9 \pm 0.6 = (22.3, 23.5)$$

Step 2: Calculate the 90% CI using the calculator

- **Calculator:** STAT→TESTS →7:ZInterval…
  - Inpt: Data [✓ Stats]
  - σ = 1.5
  - $\bar{x}$ = 22.9
  - n = 20
  - C-Level: .90
  - Calculate

ZInterval : (22.3, 23.5)

Step 3: Interpretation

- We are 90% confident that the mean age of all students at that college is between 22.3 and 23.5 years.

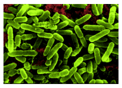
Example 1

- How many students should he ask if he wants the margin of error to be no more than 0.5 years with 99% confidence?

$$n=\left(z^{*}\cdot\frac{\sigma}{E}\right)^{2}=\left(2.576\frac{1.5}{0.5}\right)^{2}=59.72$$

- **Thus, he needs to have at least 60 students in his sample.**

Example 2



A scientist wants to know the density of bacteria in a certain solution. He makes measurements of 10 randomly selected sample:

24, 31, 29, 25, 27, 27, 32, 25, 26, 29 $*10^6$ bacteria/ml.

From past studies the scientist knows that the distribution of bacteria level is normally distributed and the population standard deviation is $2*10^6$ bacteria/ml.

a. What is the point estimate of μ?

$\bar{x}$ **=27.5 $*10^6$ bacteria/ml.**

Example 2

- b. Find the 95% confidence interval for the mean level of bacteria in the solution.
  - **Step 1:** check conditions: SRS, normal distribution, σ is known. All satisfied.
  - **Step 2:** CI:

$$\bar{x}\pm z^{*}\frac{\sigma}{\sqrt{n}}=27.5\pm1.96\frac{2}{\sqrt{10}}=27.5\pm1.24=(26.26,28.74)$$

  - **Step 3:** Interpret: we are 95% confident that the mean bacteria level in the whole solution is between 26.26 and 28.74 $*10^6$ bacteria/ml.

Example 2

Using the calculator:

- Enter the number into on of the lists, say L1
- STAT → TESTS →7: ZInterval
  - Inpt: Data
  - σ: 2
  - List: L1
  - Freq: 1 (it's always 1)
  - C-Level: .95
  - Calculate
- (26.26, 28.74)

Example 2

- c. What is the margin of error?

  From part b:

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} = 27.5 \pm 1.96 \frac{2}{\sqrt{10}} = 27.5 \pm 1.24 = (26.26, 28.74)$$

Thus, the margin of error is

E=1.24 $*10^6$ bacteria/ml.

Example 2

- d. How many measurements should he make to obtain a margin of error of at most $0.5*10^6$ bacteria/ml with a confidence level of 95%?

$$n = \left(z^* \cdot \frac{\sigma}{E}\right)^2 = \left(1.96 \frac{2 \times 10^6}{0.5 \times 10^6}\right)^2 = 61.4656$$

- Thus, he needs to take 62 measurements.

Assumptions for the validity of

$$\bar{x} \pm z^* \cdot \frac{\sigma}{\sqrt{n}}$$

- **The sample must be random**
- **The standard deviation, σ, is known and <u>either</u>**
  - **The sample size must be large ($n \geq 30$) or**
  - **For smaller sample the variable of interest must be normally distributed in the population.**

- The only situation when we cannot use this confidence interval, then, is when the sample size is small and the variable of interest is not known to have a normal distribution. In that case, other methods called non-parameteric methods need to be used.

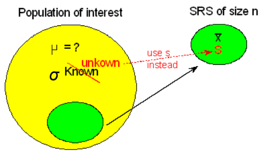
Example 3

In a randomized comparative experiment on the effects of calcium on blood pressure, researchers divided 54 healthy, white males at random into two groups, takes calcium or placebo. The paper reports a mean seated systolic blood pressure of 114.9 with standard deviation of 9.3 for the placebo group. Assume systolic blood pressure is normally distributed.

<u>Can you find a z-interval for this problem? Why or why not?</u>

So what if σ is unknown?

- Well, there is some good news and some bad news!



The good news is that we can easily **replace the population standard deviation, σ, with the *sample* standard deviation *s*.**

And the bad news is…

- that once σ has been replaced by *s*, we lose the Central Limit Theorem together with the normality of $\bar{X}$ and therefore the **confidence multipliers *z** ** for the different levels of confidence are (generally) not accurate any more.
- The new multipliers come from a different distribution called the "***t distribution***" and are therefore denoted by **t*** (instead of z*).

CI for the population mean when σ is unknown

- The confidence interval for the population mean μ when **σ is unknown** is therefore:

$$\boxed{\bar{x} \pm t^{*} \cdot \frac{s}{\sqrt{n}}}$$

z* vs. t*

- There is an important difference between the confidence multipliers we have used so far (z*) and those needed for the case when σ is unknown (t*).
  - z*, depends only on the level of confidence,
  - t* *depend on both the level of confidence and on the sample size* (for example: the t* used in a 95% confidence when n=10 is different from the t* used when n=40).

t-distribution

- There is a different *t* distribution for each sample size. We specify a particular *t* distribution by giving its degrees of freedom. The degrees of freedom for the one-sample *t* statistic come from the sample standard error *s* in the denominator of *t*. Since *s* has *n*-1 degrees of freedom, the *t*-distribution has *n*-1 degrees of freedom.

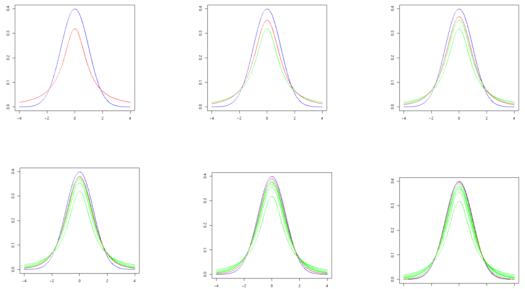
t-distribution

- The *t*-distribution is bell shaped and symmetric about the mean.
- The total area under the *t*-curve is 1
- The mean, median, and mode of the *t*-distribution are equal to zero.
- The tails in the *t*-distribution are “thicker” than those in the standard normal distribution.
- As the df (sample size) increases, the *t*-distribution approaches the normal distribution. After 29 df the *t*-distribution is very close to the standard normal *z*-distribution.

Historical Reference

- William Gosset (1876-1937) developed the t-distribution while employed by the Guinness Brewing Company in Dublin, Ireland. Gosset published his findings using the name “Student”. The t-distribution is, therefore, sometimes referred to as “Student’s t-distribution”.

Density of the *t*-distribution (red and green) for 1, 2, 3, 5, 10, and 30 df compared to normal distribution (blue)



Calculator

- **Calculator:**
- STAT→TESTS →8:TInterval…
  - Inpt: Data Stats

Use this when you have data in one of your lists

Use this when you know $\bar{x}$ and s

Example

- To study the metabolism of insects, researchers fed cockroaches measured amounts of a sugar solution. After 2, 5, and 10 hours, they dissected some of the cockroaches and measured the amount of sugar in various tissues. Five roaches fed the sugar solution and dissected after 10 hours had the following amounts of sugar in their hindguts:

Example

- 55.95, 68.24, 52.73, 21.50, 23.78
- Find the 95% CI for the mean amount of sugar in cockroach hindguts:

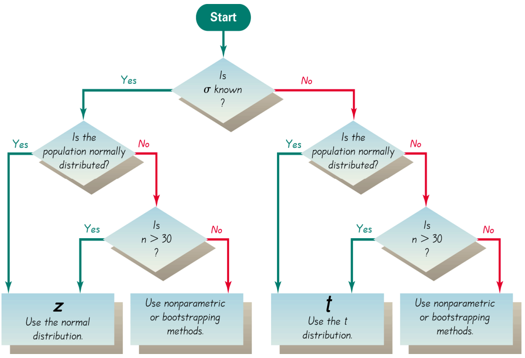
  $\bar{x}$ = 44.44 s = 20.741

  The degrees of freedom, df=n-1=4, and from the table we find that for the 95% confidence, t*=2.776.
- Then

$$\bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}} = 44.44 \pm 2.776 \cdot \frac{20.741}{\sqrt{5}} = (18.69, 70.19)$$

Example

- The large margin of error is due to the small sample size and the rather large variation among the cockroaches.
- **Calculator:**
  - Put the data in L_1.
  - STAT→TESTS →8:TInterval…
    - Inpt: Data Stats
    - List: L_1
    - Freq: 1
    - C-level: .95

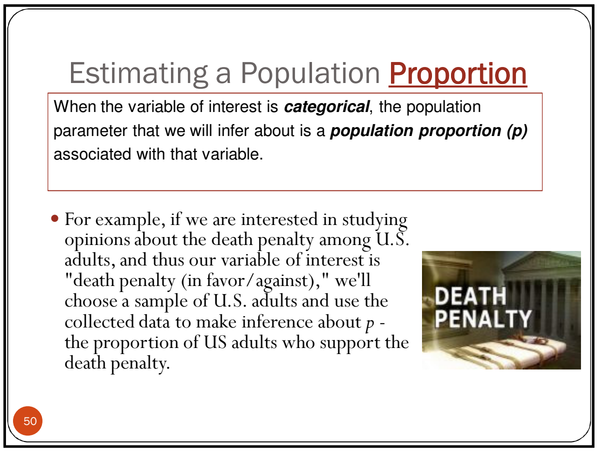


Examples: You take:

- 24 samples, the data are normally distributed, σ is known
  - normal distribution with σ $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$
- 14 samples, the data are normally distributed, σ is unknown
  - $\bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}$ t-distribution with s
- 34 samples, the data are not normally distributed, σ is unknown $\bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}$
  - normal distribution with s
- 12 samples; the data are not normally distributed, σ is unknown
  - cannot use the normal distribution or the t-distribution

Some Cautions:

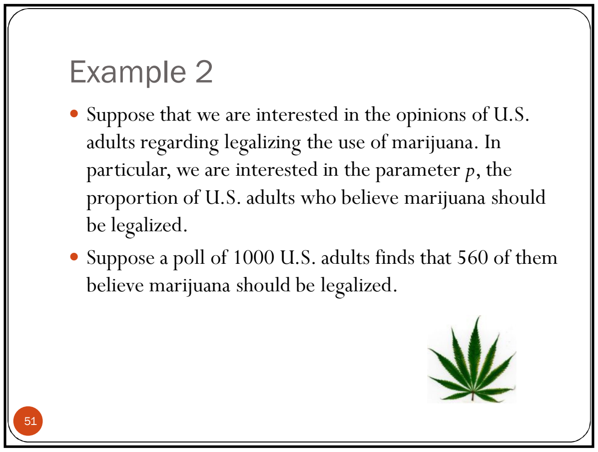
- **The data MUST be a SRS from the population**
- **The formula is not correct for more complex sampling designs, i.e., stratified, etc.**
- **No way to correct for bias in data**
- **Outliers can have a large effect on confidence interval**
- **Must know σ to do a z-interval – which is unrealistic in practice**



Estimating a Population Proportion

When the variable of interest is ***categorical***, the population parameter that we will infer about is a ***population proportion (p)*** associated with that variable.

- For example, if we are interested in studying opinions about the death penalty among U.S. adults, and thus our variable of interest is "death penalty (in favor/against)," we'll choose a sample of U.S. adults and use the collected data to make inference about *p* - the proportion of US adults who support the death penalty.

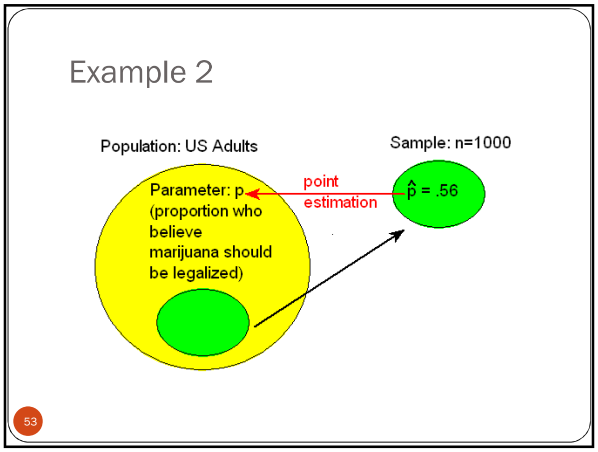


Example 2

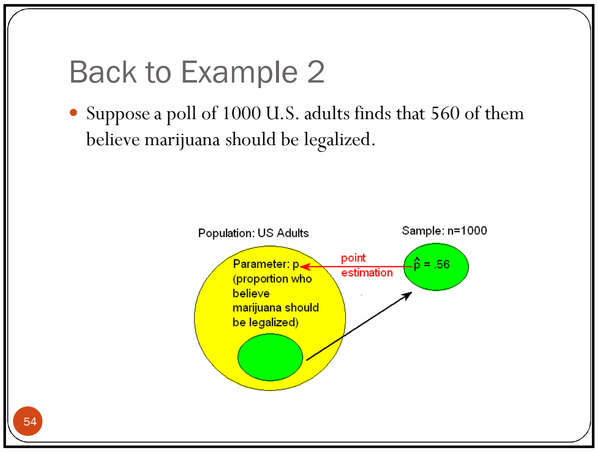
- Suppose that we are interested in the opinions of U.S. adults regarding legalizing the use of marijuana. In particular, we are interested in the parameter *p*, the proportion of U.S. adults who believe marijuana should be legalized.
- Suppose a poll of 1000 U.S. adults finds that 560 of them believe marijuana should be legalized.

Example 2

- If we wanted to estimate *p*, the population proportion by a single number based on the sample, it would make intuitive sense to use the corresponding quantity in the sample, the sample proportion $\hat{p} = 560/1000 = 0.56$. We say in this case that 0.56 is the ***point estimate*** for **p**, and that in general, we'll always use $\hat{p}$ as the ***point estimator*** for *p*.
- Note, again, that when we talk about the *specific value* (.56), we use the term estimate, and when we talk in general about the *statistic* we use the term *estimator.* Here is a visual summary of this example:



Example 2



Back to Example 2

- Suppose a poll of 1000 U.S. adults finds that 560 of them believe marijuana should be legalized.

The CI for *p*

- Thus, the **confidence interval for *p*** is

$$\hat{p} \pm E = \hat{p} \pm z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- For a 95% CI use z*=1.96
- For a 90% CI use z*=1.645
- For a 99% CI use z*=2.576

Calculator:

- STAT→TESTS →A:1-PropZInt…
- x is the number of successes: $x = n\hat{p}$

Conditions

- The CI is reasonably accurate when three conditions are met:
  - The sample was a simple random sample (SRS) from a binomial population
  - Both $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$
  - The size of the population is at least 10 times the size of the sample

Example

- Suppose you have a random sample of 40 buses from a large city and find that 24 buses have a safety violation. Find the 90% CI for the proportion of all buses that have a safety violation.
- Conditions:
  - SRS ✓
  - both $n\hat{p} = 40(\frac{24}{40}) = 24 \geq 10$ and ✓

    $n(1-\hat{p}) = 40(1-\frac{24}{40}) = 16 \geq 10$
  - The size of the population (all the buses) is at least 10 times the size of the sample (40) ✓

90% CI

$$\hat{p} = \frac{24}{40} = 0.6$$

- For 90% CI z*=1.645

$$\hat{p} \pm E = \hat{p} \pm z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.6 \pm 1.645 \cdot \sqrt{\frac{0.6(1-0.6)}{40}} = 0.6 \pm 0.13 = (0.47, 0.73)$$

Interpretation

- **1.** ***What is it that you are 90% sure is in the confidence interval?***
  The proportion of all of the buses in this population that have safety violations if we could check them all.
- **2.** ***What is the meaning (or interpretation) of the confidence interval of 0.47 to 0.73?***
  We are 90% confident that if we could check all of the buses in this population, between 47% and 73% of them would have safety violations.
- **3.** ***What is the meaning of 90% confidence?***
  If we took 100 random samples of buses from this population and computed the 90% confidence interval from each sample, then we would expect that 90 of these intervals would contain the proportion of all buses in this population that have safety violations. In other words, we are using a method that captures the true population proportion 90% of the time.

Margin of Error and Sample Size

- When we have some level of flexibility in determining the sample size, we can set a desired margin of error for estimating the population proportion and find the sample size that will achieve that.
- For example, a final poll on the day before an election would want the margin of error to be quite small (with a high level of confidence) in order to be able to predict the election results with the most precision. This is particularly relevant when it is a close race between the candidates. The polling company needs to figure out how many eligible voters it needs to include in their sample in order to achieve that.
- Let's see how we do that.

Margin of Error and Sample Size

- The confidence interval for p is

$$\hat{p} \pm E = \hat{p} \pm z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- Thus, the margin of error is

$$E = z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Using some algebra we have

$$\boxed{n = \left(\frac{z^*}{E}\right)^2 \hat{p}(1-\hat{p})}$$

$$n = \left(\frac{z^*}{E}\right)^2 \hat{p}(1-\hat{p})$$

- If you have a good estimate $\hat{p}$ of p, use it in this formula, otherwise take the conservative approach by setting $\hat{p} = \frac{1}{2}$.
- You have to decide on a level of confidence so you know what value of z* to use (most common one is the 95% level).
- Also, obviously, you have to set the margin of error (the most common one is 3%).

What sample size should we use for a survey if we want a margin of error to be at most 3%?

Let's use the 95% confidence here, so *z**=1.96.
Also, since we don't have an estimate of *p*, we will use $\hat{p} = 0.5$ .Then

$$n = \left(\frac{z^*}{E}\right)^2 \hat{p}(1-\hat{p}) = \left(\frac{1.96}{0.03}\right)^2 (0.5)(1-0.5) = 1067.111$$

Because you must have a sample size of at least 1067.111, **round up** to 1068. **So *n* should be at least 1068.**

Summary: CI for …

a population proportion	$\hat{p} \pm z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
a population mean, σ is known and normally distributed population or n≥30	$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$ $\bar{x} \pm z^* \frac{s}{\sqrt{n}}$
a population mean, σ is unknown and normally distributed population or n≥30	$\bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}$