# CONFIDENCE INTERVALS: CONFIDENCE INTERVAL FOR A POPULATION PROPORTION<sup>\*</sup>

## Susan Dean

# Barbara Illowsky, Ph.D.

This work is produced by OpenStax-CNX and licensed under the Creative Commons Attribution License  $3.0^{\dagger}$ 

#### Abstract

Confidence Interval for a Population Proportion is part of the collection coll0555 written by Barbara Illowsky and Susan Dean with contributions from Roberta Bloom.

During an election year, we see articles in the newspaper that state **confidence intervals** in terms of proportions or percentages. For example, a poll for a particular candidate running for president might show that the candidate has 40% of the vote within 3 percentage points. Often, election polls are calculated with 95% confidence. So, the pollsters would be 95% confident that the true proportion of voters who favored the candidate would be between 0.37 and 0.43: (0.40 - 0.03, 0.40 + 0.03).

Investors in the stock market are interested in the true proportion of stocks that go up and down each week. Businesses that sell personal computers are interested in the proportion of households in the United States that own personal computers. Confidence intervals can be calculated for the true proportion of stocks that go up or down each week and for the true proportion of households in the United States that own personal computers.

The procedure to find the confidence interval, the sample size, the **error bound**, and the **confidence level** for a proportion is similar to that for the population mean. The formulas are different.

How do you know you are dealing with a proportion problem? First, the underlying distribution is binomial. (There is no mention of a mean or average.) If X is a binomial random variable, then  $X \sim B(n, p)$  where n = the number of trials and p = the probability of a success. To form a proportion, take X, the random variable for the number of successes and divide it by n, the number of trials (or the sample size). The random variable P' (read "P prime") is that proportion,

$$P' = \frac{X}{n}$$

(Sometimes the random variable is denoted as  $\hat{P}$ , read "P hat".)

When n is large and p is not close to 0 or 1, we can use the **normal distribution** to approximate the binomial.

 $X \sim N\left(n \cdot p, \sqrt{n \cdot p \cdot q}\right)$ 

If we divide the random variable by n, the mean by n, and the standard deviation by n, we get a normal distribution of proportions with P', called the estimated proportion, as the random variable. (Recall that a proportion = the number of successes divided by n.)

<sup>\*</sup>Version 1.20: Jun 9, 2012 7:26 am -0500

<sup>&</sup>lt;sup>†</sup>http://creativecommons.org/licenses/by/3.0/

OpenStax-CNX module: m16963

$$\frac{X}{n} = P' \sim N\left(\frac{n \cdot p}{n}, \frac{\sqrt{n \cdot p \cdot q}}{n}\right)$$
  
Using algebra to simplify :  $\frac{\sqrt{n \cdot p \cdot q}}{n} = \sqrt{\frac{p \cdot q}{n}}$ 

P' follows a normal distribution for proportions:  $P' \sim N\left(p, \sqrt{\frac{p \cdot q}{n}}\right)$ 

The confidence interval has the form  $(p' - \mathbf{EBP}, p' + \mathbf{EBP})$ .

 $p' = \frac{x}{n}$ 

p' = the estimated proportion of successes (p' is a point estimate for p, the true proportion)

x =the **number** of successes.

n =the size of the sample

The error bound for a proportion is

 $\text{EBP} = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p' \cdot q'}{n}} \qquad where q' = 1 - p'$ 

This formula is similar to the error bound formula for a mean, except that the "appropriate standard deviation" is different. For a mean, when the population standard deviation is known, the appropriate standard deviation that we use is  $\frac{\sigma}{\sqrt{n}}$ . For a proportion, the appropriate standard deviation is  $\sqrt{\frac{p \cdot q}{n}}$ .

However, in the error bound formula, we use  $\sqrt{\frac{p'\cdot q'}{n}}$  as the standard deviation, instead of  $\sqrt{\frac{p\cdot q}{n}}$ 

However, in the error bound formula, the standard deviation is  $\sqrt{\frac{p'\cdot q'}{n}}$ 

In the error bound formula, the sample proportions p' and q' are estimates of the unknown population proportions p and q. The estimated proportions p' and q' are used because p and q are not known. p' and q' are calculated from the data. p' is the estimated proportion of successes. q' is the estimated proportion of failures.

The confidence interval can only be used if the number of successes np' and the number of failures nq' are both larger than 5.

NOTE: For the normal distribution of proportions, the z-score formula is as follows.

If 
$$P' \sim N\left(p, \sqrt{\frac{p \cdot q}{n}}\right)$$
 then the z-score formula is  $z = \frac{p' - p}{\sqrt{\frac{p \cdot q}{n}}}$ 

## Example 1

Suppose that a market research firm is hired to estimate the percent of adults living in a large city who have cell phones. 500 randomly selected adult residents in this city are surveyed to determine whether they have cell phones. Of the 500 people surveyed, 421 responded yes - they own cell phones. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of adults residents of this city who have cell phones.

#### Solution

- You can use technology to directly calculate the confidence interval.
- The first solution is step-by-step (Solution A).
- The second solution uses a function of the TI-83, 83+ or 84 calculators (Solution B).

#### Solution A

Let X = the number of people in the sample who have cell phones. X is binomial.  $X \sim B(500, \frac{421}{500})$ . To calculate the confidence interval, you must find p', q', and EBP.

n = 500 x = the number of successes = 421  $p' = \frac{x}{n} = \frac{421}{500} = 0.842$ 

p' = 0.842 is the sample proportion; this is the point estimate of the population proportion. q' = 1 - p' = 1 - 0.842 = 0.158

Since CL = 0.95, then  $\alpha = 1 - \text{CL} = 1 - 0.95 = 0.05$   $\frac{\alpha}{2} = 0.025$ . Then  $z_{\frac{\alpha}{2}} = z_{.025} = 1.96$  Use the TI-83, 83+ or 84+ calculator command invNorm(0.975,0,1) to find  $z_{.025}$ . Remember that the area to the right of  $z_{.025}$  is 0.025 and the area to the left of  $z_{0.025}$  is 0.975. This can also be found using appropriate commands on other calculators, using a computer, or using a Standard Normal probability table.

$$\begin{split} \text{EBP} &= z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p' \cdot q'}{n}} = 1.96 \cdot \sqrt{\frac{(0.842) \cdot (0.158)}{500}} = 0.032\\ p' - \text{EBP} &= 0.842 - 0.032 = 0.81\\ p' + \text{EBP} &= 0.842 + 0.032 = 0.874 \end{split}$$

The confidence interval for the true binomial population proportion is (p' - EBP, p' + EBP) = (0.810, 0.874). Interpretation

We estimate with 95% confidence that between 81% and 87.4% of all adult residents of this city have cell phones.

#### Explanation of 95% Confidence Level

95% of the confidence intervals constructed in this way would contain the true value for the population proportion of all adult residents of this city who have cell phones.

## Solution B Using a function of the TI-83, 83+ or 84 calculators:

Press STAT and arrow over to TESTS.

Arrow down to A:1-PropZint. Press ENTER.

Arrow down to x and enter 421.

Arrow down to n and enter 500.

Arrow down to C-Level and enter .95.

Arrow down to Calculate and press ENTER.

The confidence interval is (0.81003, 0.87397).

#### Example 2

For a class project, a political science student at a large university wants to estimate the percent of students that are registered voters. He surveys 500 students and finds that 300 are registered voters. Compute a 90% confidence interval for the true percent of students that are registered voters and interpret the confidence interval.

#### Solution

- You can use technology to directly calculate the confidence interval.
- The first solution is step-by-step (Solution A).
- The second solution uses a function of the TI-83, 83+ or 84 calculators (Solution B).

## Solution A

 $\begin{aligned} x &= 300 \text{ and } n = 500. \\ p' &= \frac{x}{n} = \frac{300}{500} = 0.600 \\ q' &= 1 - p' = 1 - 0.600 = 0.400 \\ \text{Since CL} &= 0.90, \text{ then } \alpha = 1 - \text{CL} = 1 - 0.90 = 0.10 \\ z \stackrel{\alpha}{=} &= z_{.05} = 1.645 \\ z \stackrel{\alpha}{=} &= z_{.05} = 1.645 \end{aligned}$ 

Use the TI-83, 83+ or 84+ calculator command invNorm(0.95,0,1) to find  $z_{.05}$ . Remember that the area to the right of  $z_{.05}$  is 0.05 and the area to the left of  $z_{.05}$  is 0.95. This can also be found using appropriate commands on other calculators, using a computer, or using a Standard Normal probability table.

$$\begin{split} \text{EBP} &= z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p' \cdot q'}{n}} = 1.645 \cdot \sqrt{\frac{(0.60) \cdot (0.40)}{500}} = 0.036\\ p' - \text{EBP} &= 0.60 - 0.036 = 0.564 \end{split}$$

p' + EBP = 0.60 + 0.036 = 0.636

The confidence interval for the true binomial population proportion is (p' - EBP, p' + EBP) = (0.564, 0.636).

## Interpretation:

- We estimate with 90% confidence that the true percent of all students that are registered voters is between 56.4% and 63.6%.
- Alternate Wording: We estimate with 90% confidence that between 56.4% and 63.6% of ALL students are registered voters.

## Explanation of 90% Confidence Level

90% of all confidence intervals constructed in this way contain the true value for the population percent of students that are registered voters.

## Solution B

Using a function of the TI-83, 83+ or 84 calculators:

Press STAT and arrow over to TESTS. Arrow down to A:1-PropZint. Press ENTER. Arrow down to x and enter 300. Arrow down to n and enter 500. Arrow down to C-Level and enter .90. Arrow down to Calculate and press ENTER. The confidence interval is (0.564, 0.636).

## 1 Calculating the Sample Size n

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size.

The error bound formula for a population proportion is

• EBP = 
$$z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p'q'}{n}}$$

• Solving for *n* gives you an equation for the sample size.

• 
$$n = \frac{z_{\frac{\alpha}{2}} \cdot \mathbf{p}'\mathbf{q}'}{\mathrm{EBP}^2}$$

#### Example 3

Suppose a mobile phone company wants to determine the current percentage of customers aged 50+ that use text messaging on their cell phone. How many customers aged 50+ should the company survey in order to be 90% confident that the estimated (sample) proportion is within 3 percentage points of the true population proportion of customers aged 50+ that use text messaging on their cell phone.

## Solution

From the problem, we know that EBP=0.03 (3%=0.03) and

 $z_{\frac{\alpha}{2}} = z_{.05} = 1.645$  because the confidence level is 90%

However, in order to find n, we need to know the estimated (sample) proportion p'. Remember that q'=1-p'. But, we do not know p' yet. Since we multiply p' and q' together, we make them both equal to 0.5 because p'q' = (.5)(.5) = .25 results in the largest possible product. (Try other products: (.6)(.4)=.24; (.3)(.7)=.21; (.2)(.8)=.16 and so on). The largest possible product gives us the largest n. This gives us a large enough sample so that we can be 90% confident that we are within 3 percentage points of the true population proportion. To calculate the sample size n, use the formula and make the substitutions.  $n = \frac{z^2 p' q'}{EBP^2}$  gives  $n = \frac{1.645^2(.5)(.5)}{.03^2} = 751.7$ Round the answer to the next higher value. The sample size should be 752 cell phone customers

aged 50+ in order to be 90% confident that the estimated (sample) proportion is within 3 percentage points of the true population proportion of all customers aged 50+ that use text messaging on their cell phone.

\*\*With contributions from Roberta Bloom.

# Glossary

#### **Definition 1: Binomial Distribution**

A discrete random variable (RV) which arises from Bernoulli trials. There are a fixed number,  $n_{i}$ of independent trials. "Independent" means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in n trials. The notation is:  $X \sim B(n,p)$ . The mean is  $\mu = np$  and the standard deviation is  $\sigma = \sqrt{npq}$ . The probability of exactly x successes in n trials is  $P(X = x) = \binom{n}{x} p^x q^{n-x}$ .

### **Definition 2: Confidence Interval (CI)**

An interval estimate for an unknown population parameter. This depends on:

- The desired confidence level.
- Information that is known about the distribution (for example, known standard deviation).
- The sample and its size.

#### Definition 3: Confidence Level (CL)

The percent expression for the probability that the confidence interval contains the true population parameter. For example, if the CL = 90%, then in 90 out of 100 samples the interval estimate will enclose the true population parameter.

#### Definition 4: Error Bound for a Population Proportion(EBP)

The margin of error. Depends on the confidence level, sample size, and the estimated (from the sample) proportion of successes.

## **Definition 5: Normal Distribution**

A continuous random variable (RV) with pdf  $f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}$ , where  $\mu$  is the mean of the distribution and  $\sigma$  is the standard deviation. Notation:  $X \sim N(\mu, \sigma)$ . If  $\mu = 0$  and  $\sigma = 1$ , the RV is called the standard normal distribution.