Section 3.2 – Estimating a Population Proportion

In this section we present methods for using a sample proportion to estimate the value of a population proportion.

- ✓ The sample proportion is the best point estimate of the population proportion.
- ✓ We can use a sample proportion to construct a confidence interval to estimate the true value of a population proportion, and we should know how to interpret such confidence intervals.
- ✓ We should know how to find the sample size necessary to estimate a population proportion.

Definition

A ***point estimate*** is a single value (or point) used to approximate a population parameter.

For ***example***, the point estimate for the population proportion is $\hat{p} = \frac{x}{n}$, where x is the number of individuals in the sample with a specified characteristic and n is the sample size.

The ***sample proportion*** $\hat{p}$ is the best point estimate of the population proportion p.

Example

In a Pew Research Center poll, 70% of 1501 randomly selected adults in the United States believe in global warming, so the sample proportion is $\hat{p} = 0.70$. Find the best point estimate of the proportion of all adults in the United States who believe in global warming.

Solution

Because the sample proportion is the best point estimate of the population proportion, we conclude that the best point estimate of p is 0.70. When using the sample results to estimate the percentage of all adults in the United States who believe in global warming, the best estimate is 70%.

Example

In July of 2008, a Quinnipiac University Poll asked 1783 registered voters nationwide whether they favored or opposed the death penalty for persons convicted of murder. 1123 were in favor.
Obtain a point estimate for the proportion of registered voters nationwide who are in favor of the death penalty for persons convicted of murder.

Solution

Obtain a point estimate for the proportion of registered voters nationwide who are in favor of the death penalty for persons convicted of murder.

$$\hat{p} = \frac{x}{n} = \frac{1123}{1783} \underline{= 0.63}$$

Definitions

A ***confidence interval*** (or ***interval estimate***) for an unknown parameter consists of an interval of numbers based on a point estimate

A ***confidence level*** is the probability $1 - \alpha$ (often expressed as the equivalent percentage value) that the confidence interval actually does contain the population parameter, assuming that the estimation process is repeated a large number of times. (The confidence level is also called ***degree of confidence***, or the ***confidence coefficient***.) is denoted $(1-\alpha)\cdot 100\%$

Most common choices are 90%, 95%, or 99%. (a = 10%), (a = 5%), (a = 1%)

For ***example***, a 95% level of confidence ($\alpha = 0.05$) implies that if 100 different confidence intervals are constructed, each based on a different sample from the same population, we will expect 95 of the intervals to contain the parameter and 5 not to include the parameter.

Confidence interval estimates for the population proportion are of the form

Point estimate ± margin of error.

The margin of error of a confidence interval estimate of a parameter is a measure of how accurate the point estimate is.

Interpreting a Confidence Interval

We must be careful to interpret confidence intervals correctly. There is a correct interpretation and many different and creative incorrect interpretations of the confidence interval $0.677 < p < 0.723$.

Correct: "We are 95% confident that the interval from 0.677 to 0.723 actually does contain the true value of the population proportion p."
This means that if we were to select many different samples of size 1501 and construct the corresponding confidence intervals, 95% of them would actually contain the value of the population proportion p.
(Note that in this correct interpretation, the level of 95% refers to the success rate of the process being used to estimate the proportion.)

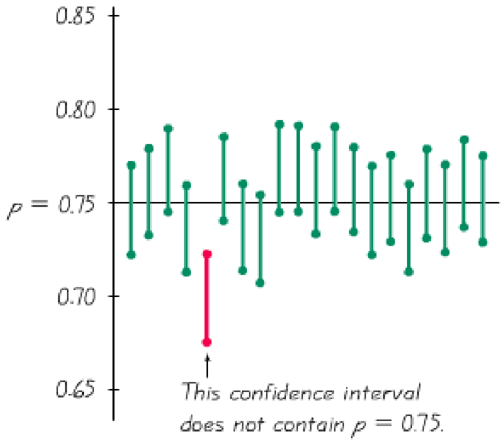
Incorrect: "There is a 95% chance that the true value of p will fall between 0.677 and 0.723." It would also be incorrect to say that 95% of sample proportions fall between 0.677 and 0.723.

The margin of error depends on ***three*** factors:

Level of confidence: As the level of confidence increases, the margin of error also increases.
Sample size: As the size of the random sample increases, the margin of error decreases.
Standard deviation of the population: The more spread there is in the population, the wider our interval will be for a given level of confidence.

Caution:

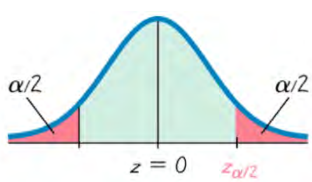
- Know the correct interpretation of a confidence interval.
- Confidence intervals can be used informally to compare different data sets, but the overlapping of confidence intervals should not be used for making formal and final conclusions about equality of proportions.



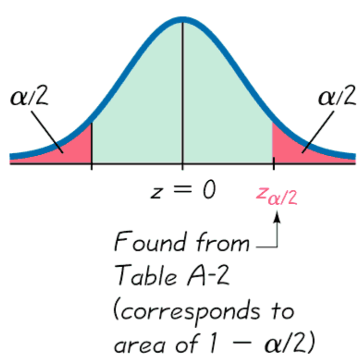
Critical Values

A standard z score can be used to distinguish between sample statistics that are likely to occur and those that are unlikely to occur. Such a z score is called a critical value. Critical values are based on the following observations:

1. Under certain conditions, the sampling distribution of sample proportions can be approximated by a normal distribution.



2. A z score associated with a sample proportion has a probability of $\alpha/2$ of falling in the right tail.



3. The z score separating the right-tail region is commonly denoted by $z_{\alpha/2}$ and is referred to as a ***critical value*** because it is on the borderline separating z scores from sample proportions that are likely to occur from those that are unlikely to occur.

Definition

A ***critical value*** is the number on the borderline separating sample statistics that are likely to occur from those that are unlikely to occur. The number $z_{\alpha/2}$ is a critical value that is a z score with the property that it separates an area of $\alpha/2$ in the right tail of the standard normal distribution.

Notation for Critical Value

The critical value $z_{\alpha/2}$ is the positive z value that is at the vertical boundary separating an area of $\alpha/2$ in the right tail of the standard normal distribution. (The value of $-z_{\alpha/2}$ is at the vertical boundary for the area of $\alpha/2$ in the ***left tail***.) The subscript $\alpha/2$ is simply a reminder that the z score separates an area of $\alpha/2$ in the ***right tail*** of the standard normal distribution.

Definition

When data from a simple random sample are used to estimate a population proportion p, the ***margin of error***, denoted by **E**, is the maximum likely difference (with probability $1-\alpha$, such as 0.95) between the observed proportion $\hat{p}$ and the true value of the population proportion p. The margin of error E is also called the maximum error of the estimate and can be found by multiplying the critical value and the standard deviation of the sample proportions:

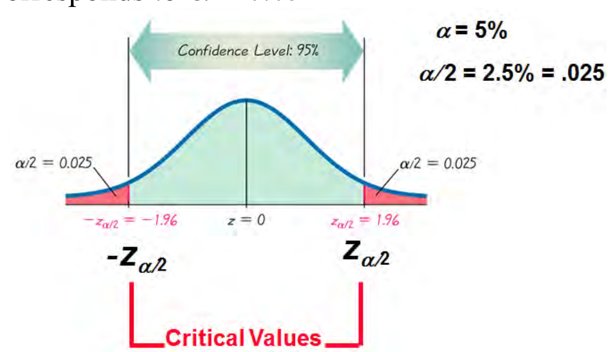
$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

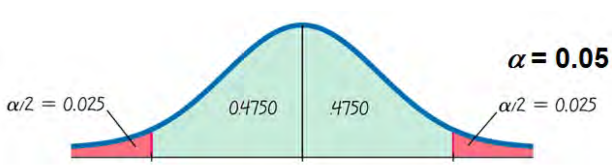
Example

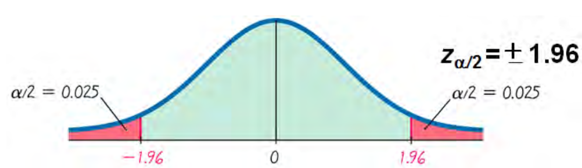
Find the critical value $z_{\alpha/2}$ corresponding to a 95% confidence level.

<u>***Solution***</u>

A 95% confidence level corresponds to $\alpha = 0.05$







The area in each of the red-shaded tails is $\frac{\alpha}{2} = 0.025$. The cumulative area to its left must be $1 - 0.024 = 0.975$. From the Normal Distribution Table, the area of 0.975 corresponds to $z = 1.96$. For a 95% confidence level, the critical value is therefore $z_{\alpha/2} = 1.96$

Confidence Interval for Estimating a Population Proportion *p*

Notation

p = population proportion
$\hat{p}$ = sample proportion
n = number of sample values
E = margin of error
$z_{\alpha/2}$ = z score separating an area of $\alpha/2$ in the right tail of the standard normal distribution

Requirements

1. The sample is a simple random sample.
2. The conditions for the binomial distribution are satisfied: there is a fixed number of trials, the trials are independent, there are two categories of outcomes, and the probabilities remain constant for each trial.
3. There are at least 5 successes and 5 failures.

Procedure for Constructing a Confidence Interval for *p*

1. Verify that the required assumptions are satisfied. (The sample is a simple random sample, the conditions for the binomial distribution are satisfied, and the normal distribution can be used to approximate the distribution of sample proportions because $np \geq 5$, and $nq \geq 5$ are both satisfied.)
2. Refer to Standard Normal Distribution Table and find the critical value $z_{\alpha/2}$ that corresponds to the desired confidence level.
3. Evaluate the margin of error $E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$
4. Using the value of the calculated margin of error, E and the value of the sample proportion, $\hat{p}$, find the values of $\hat{p} - E$ and $\hat{p} + E$. Substitute those values in the general format for the confidence interval:
$$\hat{p} - E < p < \hat{p} + E$$
5. Round the resulting confidence interval limits to three significant digits.

Example

In the Chapter Problem we noted that a Pew Research Center poll of 1501 randomly selected U.S. adults showed that 70% of the respondents believe in global warming. The sample results are n = 1501, and $\hat{p} = 0.70$

a) Find the margin of error E that corresponds to a 95% confidence level.
b) Find the 95% confidence interval estimate of the population proportion p.
c) Based on the results, can we safely conclude that the majority of adults believe in global warming?
d) Assuming that you are a newspaper reporter, write a brief statement that accurately describes the results and includes all of the relevant information.

Solution

Requirement check: simple random sample; fixed number of trials, 1501; trials are independent; two categories of outcomes (believes or does not); probability remains constant. Note: number of successes and failures are both at least 5.

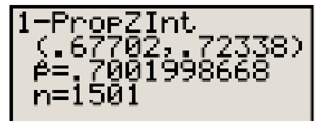
a) Use the formula to find the margin of error.

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$= 1.96\sqrt{\frac{(0.70)(0.30)}{1501}}$$

$$= 0.023183$$

b) The 95% confidence interval:



$$\hat{p} - E < p < \hat{p} + E$$

$$0.70 - 0.023183 < p < 0.70 + 0.023183$$

$$0.677 < p < 0.723$$

***TI*-84:** Go STATS → TESTS → select A: 1-PropZInt

For x values mulptiply .7 (70%) by 1501 (n), however go back and round the number (.7*1501 = 1050.7) therefore the x-value is 1051

c) Based on the confidence interval obtained in part (b), it does appear that the proportion of adults who believe in global warming is greater than 0.5 (or 50%), so we can safely conclude that the majority of adults believe in global warming. Because the limits of **0.677** and **0.723** are likely to contain the true population proportion, it appears that the population proportion is a value greater than 0.5.

d) Here is one statement that summarizes the results: 70% of United States adults believe that the earth is getting warmer. That percentage is based on a Pew Research Center poll of 1501 randomly selected adults in the United States. In theory, in 95% of such polls, the percentage should differ by no more than 2.3 percentage points in either direction from the percentage that would be found by interviewing all adults in the United States.

Analyzing Polls

When analyzing polls consider:

1. The sample should be a simple random sample, not an inappropriate sample (such as a voluntary response sample).
2. The confidence level should be provided. (It is often 95%, but media reports often neglect to identify it.)
3. The sample size should be provided. (It is usually provided by the media, but not always.)
4. Except for relatively rare cases, the quality of the poll results depends on the sampling method and the size of the sample, but the size of the population is usually not a factor.

Caution

Never follow the common misconception that poll results are unreliable if the sample size is a small percentage of the population size. The population size is usually not a factor in determining the reliability of a poll.

Sample Size

Suppose we want to collect sample data in order to estimate some population proportion.
Sample size needed for a specified margin of error, E, and level of confidence $(1 - \alpha)$:

$$n = \hat{p}\hat{q}\left(\frac{z_{\alpha/2}}{E}\right)^2$$

Round-Off Rule for Determining Sample Size

If the computed sample size n is not a whole number, round the value of n up to the next larger whole number.

Example

The Internet is affecting us all in many different ways, so there are many reasons for estimating the proportion of adults who use it. Assume that a manager for E-Bay wants to determine the current percentage of U.S. adults who now use the Internet. How many adults must be surveyed in order to be 95% confident that the sample percentage is in error by no more than three percentage points?

a) In 2006, 73% of adults used the Internet.
b) No known possible value of the proportion.

Solution

a) ***Given***: $\hat{p} = 0.73 \quad so \quad \hat{q} = 1 - 0.73 = 0.27$

With a 95% confidence level, we have $\alpha = 0.05$, so $z_{\alpha/2} = 1.96$. Also the margin of error is $E = 0.03$ (the decimal equivalent of "3 percentage points")

$$n = \frac{\left[z_{\alpha/2}\right]^2 \hat{p}\hat{q}}{E^2} = \frac{(1.96)^2 (0.73)(0.27)}{0.03^2}$$

≈ 842

We must obtain a simple random sample that includes at least 842 adults.

b) ***Given***: $z_{\alpha/2} = 1.96$ and $E = 0.03$

But with no prior knowledge of $\hat{p}$ *or* $\hat{q}$

$$n = \frac{\left[z_{\alpha/2}\right]^2 \hat{p}\hat{q}}{E^2} = \frac{(1.96)^2 (0.25)}{0.03^2}$$

≈ 1068

To be 95% confident that our sample percentage is within three percentage points of the true percentage for all adults, we should obtain a simple random sample of 1068 adults.

Finding the Point Estimate and E from a Confidence Interval

Point estimate of p: $\widehat{p} = \dfrac{(\textit{upper confidence limit}) + (\textit{lower confidence limit})}{2}$

Margin Error: $E = \dfrac{(\textit{upper confidence limit}) - (\textit{lower confidence limit})}{2}$

Suppose that a simple random sample of size n is taken from a population. A $(1 - \alpha)\cdot 100\%$ confidence interval for p is given by the following quantities

Lower bound: $\hat{p} - E \quad \to \quad \hat{p} - z_{\alpha/2}\sqrt{\dfrac{\hat{p}\hat{q}}{n}}$

Upper bound: $\hat{p} + E \quad \to \quad \hat{p} + z_{\alpha/2}\sqrt{\dfrac{\hat{p}\hat{q}}{n}}$

Note: It must be the case that $n\hat{p}\hat{q}$ and $n \le 0.05N$ to construct this interval.

Example

The article "High-Dose Nicotine Patch Therapy," by Dale Hurt, includes this statement: "off the 71 subjects, 70% were abstinent from smoking at 8 weeks (95% confidence interval, 58% to 81%)," Use that statement to find the point estimate $\hat{p}$ and the margin error E.

Solution

$$\widehat{p} = \frac{(\textit{upper confidence limit}) + (\textit{lower confidence limit})}{2}$$

$$= \frac{0.81 + 0.58}{2}$$

$$\underline{= 0.695}$$

$$E = \frac{(\textit{upper confidence limit}) - (\textit{lower confidence limit})}{2}$$

$$= \frac{0.81 - 0.58}{2}$$

$$\underline{= 0.115}$$

Example

In July of 2008, a Quinnipiac University Poll asked 1783 registered voters nationwide whether they favored or opposed the death penalty for persons convicted of murder. 1123 were in favor.
Obtain a 90% confidence interval for the proportion of registered voters nationwide who are in favor of the death penalty for persons convicted of murder.

Solution

$$\hat{p} = 0.63$$

$$n\hat{p}\hat{q} = 1783(0.63)(1 - 0.63) = 415.6 \ge 0$$

The sample side is definitely less than 5% of the population size

$\alpha = 0.10 \quad \Rightarrow \quad z_{\alpha/2} = z_{0.05} = 1.645$

Lower bound: $\quad \hat{p} - z_{\alpha/2}\sqrt{\dfrac{\hat{p}\hat{q}}{n}} = 0.63 - 1.645\sqrt{\dfrac{0.63(1-0.63)}{1783}} \approx 0.61$

Upper bound: $\quad \hat{p} + z_{\alpha/2}\sqrt{\dfrac{\hat{p}\hat{q}}{n}} = 0.63 + 1.645\sqrt{\dfrac{0.63(1-0.63)}{1783}} \approx 0.65$

We are 90% confident that the proportion of registered voters who are in favor of the death penalty for those convicted of murder is between **0.61**and **0.65**.

Exercises *Section* 3.2 – Estimating a Population Proportion

1. Find the critical value $z_{\alpha/2}$ that corresponds to a 99% confidence level.

2. Find the critical value $z_{\alpha/2}$ that corresponds to a 99.5% confidence level.

3. Find the critical value $z_{\alpha/2}$ that corresponds to a 98% confidence level.

4. Find $z_{\alpha/2}$ for $\alpha = 0.10$.

5. Find $z_{\alpha/2}$ for $\alpha = 0.02$.

6. Express the confidence interval $0.200 < p < 0.500$ in the form $\hat{p} \pm E$

7. Express the confidence interval $0.42 < p < 0.54$ in the form $\hat{p} \pm E$

8. Express the confidence interval 0.222 ± 0.044 in the form $\hat{p} - E < p < \hat{p} + E$

9. Find the point estimate $\hat{p}$ and the margin of error *E* of $(0.320,\ 0.420)$

10. Find the margin of error *E* of $0.542 < p < 0.576$

11. Find the point estimate $\hat{p}$ of $0.824 < p < 0.868$

12. Find the point estimate $\hat{p}$ and the margin of error *E* of $0.772 < p < 0.776$

13. Find the point estimate $\hat{p}$ and the margin of error *E* of $0.433 < p < 0.527$

14. Assume that a sample is used to estimate a population proportion *p*. Find the margin of error *E* that corresponds to the given $n = 1000, \quad x = 400, \quad 95\%\ \textit{confidence}$

15. Assume that a sample is used to estimate a population proportion *p*. Find the margin of error *E* that corresponds to the given $n = 500, \quad x = 220, \quad 99\%\ \textit{confidence}$

16. Assume that a sample is used to estimate a population proportion *p*. Find the margin of error *E* that corresponds to the given $n = 390, \quad x = 130, \quad 90\%\ \textit{confidence}$

17. Assume that a sample is used to estimate a population proportion *p*. Find the margin of error *E* that corresponds to the given 98% confidence; the sample size is 1230, of which 40% are successes.

18. Construct the confidence interval estimate of the population proportion *p* that corresponds to the given $n = 200, \quad x = 40, \quad 95\%\ \textit{confidence}$

19. Construct the confidence interval estimate of the population proportion *p* that corresponds to the given $n = 1236, \quad x = 109, \quad 99\%\ \textit{confidence}$

20. Construct the confidence interval estimate of the population proportion *p* that corresponds to the given $n = 5200, \quad x = 4821, \quad 99\%\ \textit{confidence}$

21. Find the minimum sample size requires to estimate a population proportion or percentage:
Margin of error: 0.045; confidence level: 95%: $\hat{p}$ *and* $\hat{q}$ *unknown*

22. Find the minimum sample size requires to estimate a population proportion or percentage:
Margin of error: 2% points; confidence level: 99%: from prior study, $\hat{p}$ is estimate by the decimal equivalent of 14%

23. The Genetics and IVF Institute conducted a clinical trial of the XSORT method designed to increase the probability of conceiving a girl. As of this writing, 574 babies were born to parents using the XSORT method, and 525 of them were girls.

a) What is the best point estimate of the population proportion of girls born to parents using the XSORT method?

b) Use the sample data to construct a 95% confidence interval estimate of the percentage of girls born to parents using the XSORT method.

c) Based on the results, does the XSORT method appear to be effective? Why or why not?

24. An important issue facing Americans is the large number of medical malpractice lawsuits and the expenses that they generate. In a study of 1228 randomly selected medical malpractice lawsuits, it is found that 856 of them were later dropped or dismissed.

a) What is the best point estimate of the proportion of medical malpractice lawsuits that are dropped or dismissed?

b) Construct a 99% confidence interval estimate of the proportion of medical malpractice lawsuits that are dropped or dismissed.

c) Does it appear that the majority of such suits are dropped or dismissed?

25. A study of 420,095 Danish cell phone users found that 135 of them developed cancer was found to be 0.0340% for those not using cell phones.

a) Use the sample data to construct a 95% confidence interval estimate of the percentage of cell phone users who develop cancer of the brain or nervous system.

b) Do cell phone users appear to have a rate of cancer of the brain or nervous system that is different from the rate of such cancer among those not using cells phones? Why or why not?

26. In an Account survey of 150 senior executives, 47% said that the most common job interview mistake is to have little or no knowledge of the company. Construct a 99% confidence interval estimate of the proportion of all senior executives who have that same opinion. Is it possible that exactly half of all senior executives believe that the most common job interview mistake is to have little or no knowledge of the company? Why or why not?