Confidence intervals for proportions

Teachers Teaching with Technology™
Professional Development from Texas Instruments

Student Activity

7 8 9 10 **11** **12**


TI-Nspire


Investigation


Student


60 min

Introduction

From previous activity

This activity assumes knowledge of the material covered in the activity ***Distribution of sample proportions***. That activity focused on key features of the sampling distribution of sample proportions.

Through simulated random sampling for various sample sizes, n, and population proportions, p, it was found that:

- The sampling distribution of the sample proportion, $\hat{P}$, is centred at p.
- For a given sample size, the spread and symmetry of the sampling distribution of $\hat{P}$ depends on the population proportion, p. The spread is greatest for $p = 0.5$. The sampling distribution also becomes more symmetric closer to $p = 0.5$.
- For a given value of p, the spread and symmetry of the sampling distribution of $\hat{P}$ depends on the sample size, n. The spread **decreases** as the sample size increases, while symmetry increases with sample size.
- For large samples, the sampling distribution of $\hat{P}$ can be modelled by a normal distribution, $\mathrm{N}\left(\mu_{\hat{P}}, {\sigma_{\hat{P}}}^2\right)$, where $\mu_{\hat{P}} = \mathrm{E}\left(\hat{P}\right) = p$ and $\sigma_{\hat{P}} = \mathrm{SD}\left(\hat{P}\right)\sqrt{\dfrac{p(1-p)}{n}}$.

Overview of this activity

In this activity you will investigate the precision of the estimator, $\hat{p}$, and confidence intervals for the population proportion, p. Through simulation, you will explore variations in confidence intervals between samples, and come to understand the significance of confidence intervals for p.

Interval estimates of the population proportion

If we take a random sample from a large population, such as an opinion poll, we will obtain a single value, $\hat{p}$, that provides us with an estimate of the true population proportion, p. But it is unlikely that $\hat{p}$ will be exactly equal to p, so it is valuable to combine the estimate with information about the **precision** of the estimate.

Simulations for the precision of the estimator $\hat{p}$

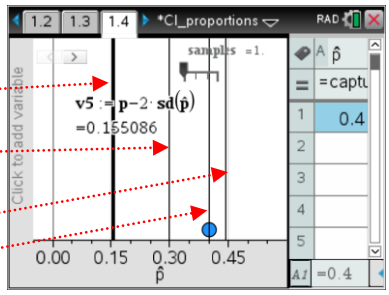
Open the TI-Nspire document 'CI_proportions'. **Navigate to Page 1.2** and follow the instructions to 'seed' the random number generator.

Navigate to Page 1.3. Suppose that in a large population the proportion of the population with a particular attribute is $p = 0.3$. Samples of size $n = 40$ will be drawn from this population. Later, you will change the values of n and p on Page 1.3, and repeat the simulation using the new values.

Simulation of 100 samples with *n* = 40 and *p* = 0.3

Author: Frank Moya

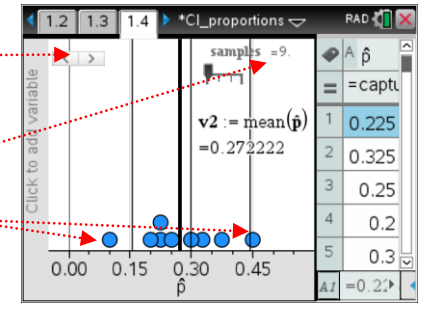
Navigate to Page 1.4.



- The left vertical boundary indicates the value of $p-2\sqrt{\frac{p(1-p)}{n}}$
- The vertical line at 0.30 indicates the value of p
- The right vertical boundary indicates the value of $p+2\sqrt{\frac{p(1-p)}{n}}$
- The variable vertical line indicates the mean of $\hat{P}$

Question 1

What is the significance of the values of $p-2\sqrt{\frac{p(1-p)}{n}}$ and $p+2\sqrt{\frac{p(1-p)}{n}}$?



Use the slider on the top left-hand corner of **Page 1.4** to simulate drawing random samples of size n from a large population.
Each time you click the right or left slider arrow, a new sample is drawn, and the sample proportion for that sample is added to the spreadsheet and to the graph.
The number of samples drawn is shown on the top right-hand side.
Look for sample proportions, $\hat{p}$, with values outside the boundaries $p\pm2\sqrt{\frac{p(1-p)}{n}}$. Stop when the number of samples drawn is 100.

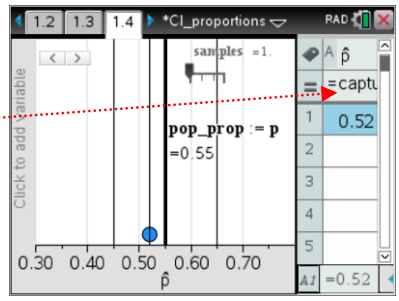
Question 2

a. From the 100 samples drawn, how many of the observed values of $\hat{p}$ were outside the boundaries defined by $p\pm2\sqrt{\frac{p(1-p)}{n}}$?

b. In the long run, what is the approximate percentage of observed values of $\hat{p}$ that you would expect to lie outside of these boundaries?

c. Explain your answer to **part b.** above.

Simulation of 100 samples with different values of *n* and *p*



Navigate back to Page 1.3. Choose different values for n and p by editing the value in the Mathbox then pressing [Enter].
Navigate to Page 1.4. Reset the simulation by selecting the formula cell 'A=' of the spreadsheet, then press [Enter][Enter].
The boundaries for $p\pm2\sqrt{\frac{p(1-p)}{n}}$ will have been automatically recalculated and shown on the graph. Change the window settings for the graph, as appropriate.
(To change window settings, click on an empty part of the graph, then [Ctrl]+[Menu] > Zoom > Window settings.)

Use the slider on the top right-hand corner to simulate drawing 100 random samples, as described earlier.

Question 3

a. From the 100 samples drawn, how many of the observed values of $\hat{p}$ were outside the boundaries in this case? Is the result consistent with your expectations?

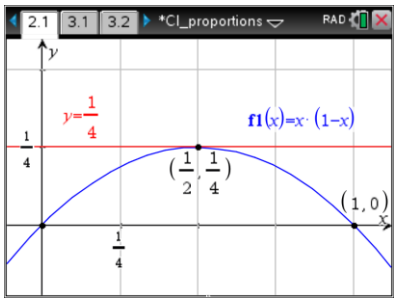
Author: Frank Moya

b. An instruction on Page 1.3 suggests that when choosing values of n and p, the following should be met: $np \geq 10$ and $n(1-p) \geq 10$. Why do you think that these criteria need to be met?

Planning the precision of the estimator

Question 4

a. If $p = 0.3$, find the minimum sample size required to ensure that $\text{SD}(\hat{P}) \leq 0.01$.



b. **Navigate to Page 2.1**. Using this graph, or otherwise, show that $p(1-p) \leq \frac{1}{4}$, where p is a population proportion.

c. Hence show that for any value of p, $\text{SD}(\hat{p}) \leq \frac{1}{2\sqrt{n}}$.

d. Discuss the significance of the result in **part c.** above.

Standard error of $\hat{P}$

In the simulations from Page 1.3, we saw that if a random sample of size n is drawn from a large population, where the population proportion is p, then it is very likely (approximately 95%) that the sample proportion, $\hat{p}$, will be within two standard deviations of p. That is

$$\Pr\left(p - 2\sqrt{\frac{p(1-p)}{n}} \leq \hat{p} \leq p + 2\sqrt{\frac{p(1-p)}{n}}\right) \approx 0.95.$$

However, there is a practical problem with using the standard deviation of $\hat{P}$, $\text{SD}(\hat{P}) = \sqrt{\frac{p(1-p)}{n}}$, as a measure of the precision of the estimator. Namely, $\text{SD}(\hat{P})$ depends on knowing the value of p. But when we are using $\hat{p}$ as an estimate of p, it is because p is unknown to us (otherwise, why would we need to estimate it?).

From **Question 4 c.**, we know that $\text{SD}(\hat{P}) \leq \frac{1}{2\sqrt{n}}$, so this the maximum that $\text{SD}(\hat{p})$ can be.

We can, however, obtain a better **estimate** of $\text{SD}(\hat{p})$ by using $\hat{p}$ as an approximation to p.

This gives the **standard error** of the estimator, $\text{SE}(\hat{P})$, where

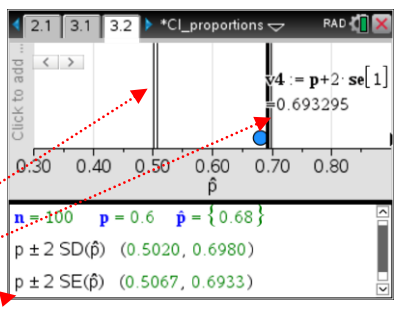
$$\text{SE}(\hat{P}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

The standard error can be used in place of the standard deviation, which it estimates. In particular,

$$\Pr\left(p - 2\text{SE}(\hat{P}) \leq \hat{p} \leq p + 2\text{SE}(\hat{P})\right) \approx 0.95.$$

Simulation to compare $\text{SD}(\hat{P})$ and $\text{SE}(\hat{P})$

In this simulation, you will observe a comparison of the intervals $\left(p - 2\text{SD}(\hat{P}),\ p + 2\text{SD}(\hat{P})\right)$ and $\left(p - 2\text{SE}(\hat{P}),\ p + 2\text{SE}(\hat{P})\right)$. In **Page 3.1**, the default values have been set to $n = 100$, $p = 0.6$.



Navigate to Page 3.2. The graph shows the boundaries for the interval $\left(p - 2\text{SD}(\hat{P}),\ p + 2\text{SD}(\hat{P})\right)$ as approximately $(0.50, 0.70)$. Use the slider arrows to draw a new sample. The interval for $\left(p - 2\text{SE}(\hat{P}),\ p + 2\text{SE}(\hat{P})\right)$ is recalculated for each sample, and shown on the graph and in the Mathbox below the graph.

Author: Frank Moya

The simulation can be repeated with new values of n and p selected on Page 2.1.

Question 5

From the results of the simulation, discuss the reasonableness of using standard error, rather than standard deviation, as an indicator of the precision of the estimator, $\hat{p}$.

Confidence intervals for the population proportion, *p*

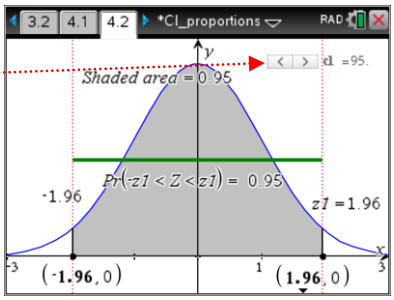
In the previous section we saw that the standard error is an indicator of the precision of the **sample statistic**: the **sample proportion**, $\hat{p}$.

However, what we are much more interested in is the precision to which we can estimate the **population parameter**: the **population proportion**, p. We are looking for is a range of values - that is, an interval – that we are reasonably sure contains the true value of p. This is called a **confidence interval**.

In the simulations that will be used to investigate confidence intervals, we will assume that values of n and p are appropriate for the sampling distribution of $\hat{P}$ to be modelled by a normal distribution, $\mathrm{N}\left(\mu_{\hat{p}}, \sigma_{\hat{p}}^{\,2}\right)$, where $\mu_{\hat{p}} = \mathrm{E}\left(\hat{P}\right) = p$ and $\sigma_{\hat{p}} = \mathrm{SD}\left(\hat{P}\right) = \sqrt{\dfrac{p(1-p)}{n}}$.

Levels of confidence and quantiles for the standard normal distribution

Navigate to Page 4.2. The graph of $Z \sim \mathrm{N}(0,1)$, the standard Normal random variable, is shown. Use the slider to adjust the percentage of the area under the curve to be shaded, between 50% and 99%, symmetrically about the origin.



Question 6

a. What is the relationship between the slider value and the shaded area?
b. What is the significance of the value labelled $z1$?
c. What is the significance of the green line segment?

From Page 4.2 we can see that $\Pr(-1.96 < Z < 1.96) \approx 0.95$, where $z1 = 1.96$ is the quantile of the standard Normal distribution that symmetrically shades approximately 95% of the area under the curve.

Question 7

Use the slider on Page 4.2 to help you complete the following table, stating corresponding value of $z1$ to two decimal places.

Confidence level (percentage of total area that is shaded)	**Standard Normal quantile** ($z1$)
50%	
75%	
90%	
95%	1.96
99%	

Standard Normal approximation of the distribution of $\hat{P}$

Assume the approximation $\hat{P} \sim \mathrm{N}\left(\mu_{\hat{p}}, \sigma_{\hat{p}}^{\,2}\right)$, where $\mu_{\hat{p}} = p$ and $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$, and let $Z \sim \mathrm{N}(0,1)$.

Standardising gives $Z = \dfrac{\hat{P} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \dfrac{\hat{P} - p}{\sqrt{\dfrac{p(1-p)}{n}}}$.

Author: Frank Moya

95% confidence interval for the true population proportion, *p*

From Page 4.2 we know that:

$\Pr(-1.96 < Z < 1.96) \approx 0.95$, therefore

$$\Pr\left(-1.96 < \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} < 1.96\right) \approx 0.95 .$$

Rearranging the inequalities gives

$$\Pr\left(\hat{P} - 1.96\sqrt{\frac{p(1-p)}{n}} < p < \hat{P} + 1.96\sqrt{\frac{p(1-p)}{n}}\right) \approx 0.95 .$$

However, when using sampling to estimate a population proportion (for example, in an opinion poll), we have a single observed value $\hat{p}$ of the random variable $\hat{P}$.

Furthermore, as discussed earlier, the standard deviation needs to be approximated by the standard error: $\mathrm{SE}(\hat{P}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, as p is unknown.

The **approximate 95% confidence interval** for p is calculated as

$$\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right).$$

Margin of error

For a confidence interval, the range of values above and below the sample proportion, $\hat{p}$, is called the **margin of error**. For a 95% confidence interval, the margin of error can be written $M = 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. Of course, if we knew the value of $\mathrm{SD}(\hat{P})$, we could use $M = 1.96\mathrm{SD}(\hat{P})$, rather than the standard error approximation.

Question 8

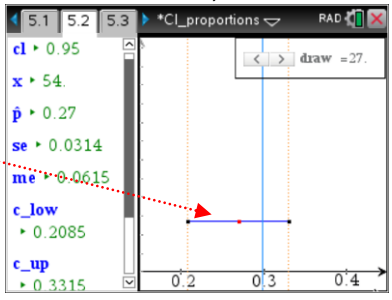
a. Refer back to **Question 4 c**. For a 95% confidence interval with sample size n, write an expression, in terms of n, for the greatest possible value of the margin of error.

b. Opinion polls usually report a margin of error, for a 95% confidence interval, of around 2 or 3 percent. Suppose that a pollster surveyed a random sample of 1600 people and calculated a 95% confidence interval. What is the maximum margin of error that could be obtained? Write the answer as a percentage, correct to two decimal places.

Understanding the significance of a 95% confidence intervals for *p*

Suppose that in a particular city 30% of households are connected to a new internet network, BNB.

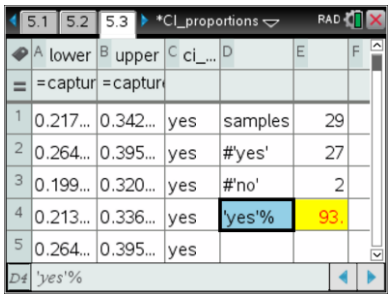
Navigate to Page 5.2. Use the slider arrow to simulate drawing a random sample of size $n = 200$ from this population, where $p = 0.3$. The value of the sample proportion, $\hat{p}$, is used to calculate and graph a 95% confidence intervals for p. Note that in the screen-shot shown, the interval contains the value $p = 0.3$.



For the sample, the numerical statistics are shown on the left-hand panel of the screen, where cl = confidence level, x = number of successes in the sample, se = standard error, me = margin of error and c_low and c_up are the endpoints of the confidence interval.

Author: Frank Moya

Navigate to Page 5.3. As you were drawing samples on Page 5.2, the spreadsheet was being populated with the confidence interval endpoints. Column C records 'yes' if the interval contains p, and 'no' otherwise. Cells E2 and E4 record and number and percentage of 'yes'.



	A lower	B upper	C ci_...	D	E	F
=	=captur	=captur				
1	0.217...	0.342...	yes	samples	29	
2	0.264...	0.395...	yes	#'yes'	27	
3	0.199...	0.320...	yes	#'no'	2	
4	0.213...	0.336...	yes	'yes'%	93.	
5	0.264...	0.395...	yes			

Question 9

a. Using the slider, draw 100 samples. How many of the 100 confidence intervals generated contain the value of p?

Reset the slider value to 1 (click on slider, then [Ctrl]+[Menu] > Settings). Draw a further 100 samples. The spreadsheet will continue to be populated.

b. From your 200 samples, what percentage of confidence intervals contained the value of p?

c. For a confidence level of 0.95, in the long run, what percentage of confidence intervals do you think will contain the true value of the population proportion?

In the previous simulation we saw that a confidence interval is obtained from the random variable $\hat{P}$, so the interval itself can be considered a random interval; the interval varies from one sample to the next, just as the value of the random variable does.

Question 10

Suppose that in a simulation similar to that on Page 5.2 we generate 100 independent 95% confidence intervals for p, where the value of p is unknown to us.

Let Y be the number of intervals that contain the value of p.

a. Explain why Y can be regarded as a binomially distributed random variable.

b. If $Y \sim \text{Bi}(a,b)$, what are the values of a, b and $\text{E}(Y)$?

Navigate to Page 6.1. Carry out the following calculations on the bottom 'Calculator' panels of Pages 6.1 and 6.2.

c. Suppose that 100 new 95% confidence intervals are to be generated.

i. Find the probability, correct to two decimal places, that exactly 5 confidence intervals do not contain the true value of p.

ii. Find the probability, correct to two decimal places, that at least 5 confidence intervals do not contain the value of true value of p.

Changing the level of confidence

An **approximate *C* % confidence interval** for p is calculated as

$$\left(\hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$, where z is the quantile of the standard Normal distribution.

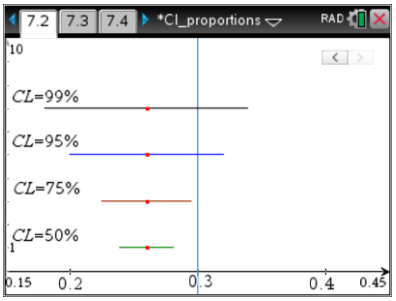
Question 11

Refer to Page 4.2 and Question 7 to complete the following.

Confidence level	Standard Normal quantile (z)	*C* % confidence interval
50%		$\left(\hat{p} - \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$
75%		$\left(\hat{p} - \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$
99%		$\left(\hat{p} - \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$

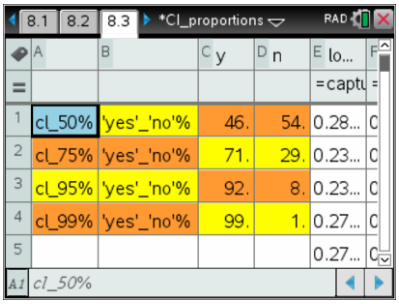
Author: Frank Moya

Comparing *C* % confidence intervals for *p*



Navigate to Page 7.2. Use the slider arrow to simulate drawing a random sample of size $n = 200$ from this population, where $p = 0.3$. The value of the sample proportion, $\hat{p}$, is used to calculate and graph 4 different confidence intervals for p, with confidence levels of 50%, 75%, 95% and 99%. Note that in the screen-shot shown, only two of the four interval contain $p = 0.3$.

Navigate to Page 7.3. As you were drawing samples on Page 7.2, the spreadsheet was being populated with the confidence interval endpoints for the four confidence levels.



The percentage of confidence intervals containing the value of p is recorded in Column C: Cell C1 – 50% confidence to Cell C4 – 99% confidence.

Navigate to Page 8.3. Problem 8 is the same as Problem 7, except that the confidence interval data for 100 samples has already been captured on Page 8.3. Use the slider arrows on Page 8.2 to select an additional 50 samples. The data on Page 8.3 is automatically updated.

Use the graphs and data to answer the following questions.

Question 12

a. For a confidence level of *C* %, in the long run, what percentage of confidence intervals will contain the true value of the population proportion?

b. Consider the following statements, made by two statisticians with regards to confidence intervals:
   - there is a trade-off between the level of confidence and the precision of the interval;
   - there is a trade-off between margin of error and level of confidence.

   Explain what these statements mean.

END OF ACTIVITY

Author: Frank Moya