Introducing Sample Proportions

Teachers Teaching with Technology™
Professional Development from Texas Instruments

Probability and statistics

Answers & Notes

TI-Nspire | Investigation | Student | 60 min

7 8 9 10 **11 12**

Introduction

A 2010 survey of attitudes to climate change, conducted in Australia by the CSIRO, reported that 40% of respondents thought of climate change in terms of natural temperature variability, rather than in terms of human-induced temperature change.

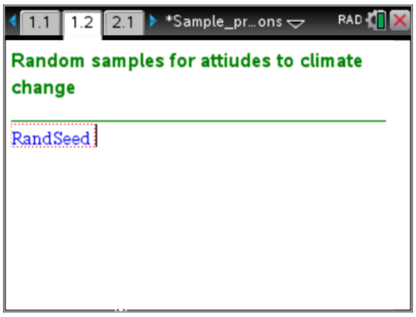
(Reference: https://publications.csiro.au/rpr/download?pid=csiro:EP105359&dsid=DS3).

How reliable is this proportion, which is based on a random sample? Does this statistic reflect the proportion of the entire Australian population with this belief about climate change?

In this activity, you will investigate how much we can expect proportions from random samples to vary from sample to sample.

Simulating random sampling for attitudes to climate change

Open the TI-Nspire document 'Sample_proportions'. You will be using this document to generate random samples from a large population. To ensure that your results are not identical to those of other students, you will 'seed' the random number generator, as follows.
Navigate to Page 1.2. In the Math Box, after the 'RandSeed' command, input a space followed by a number unique to you – such as the last 4 digits of your phone number. Press enter to execute the command.



Random samples for attiudes to climate change

RandSeed

Question 1

Why do you think that you might get identical results to those of other students in the room, if you do not 'seed' the random number generator of your handheld?

The technology uses an algorithm (rule) to generate pseudo-random numbers. The algorithm will generate the same numbers for handhelds with default factory settings. 'Seeding' with a number unique to you will initialise or reset the random number generator.

1.0 Drawing random samples of size 50

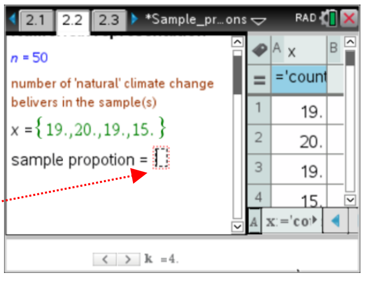
In this section, you will simulate drawing random samples of size 50 from a large population. You will use the sample results to estimate the proportion of the population who believe that climate change is due to natural temperature variability (hereafter referred to as 'natural' climate change).

- Let n denote the sample size; therefore, in this section, $n = 50$.
- Let p denote proportion of the population who believe in 'natural' climate change.
- Let X denote the **number, in random samples,** who believe in 'natural' climate change, and let x denote the values of the variable, X .

Author: Frank Moya

- Let $\hat{P}$ denote the **proportion, in random samples,** who believe in 'natural' climate change, and let $\hat{p}$ denote the values of the variable, $\hat{P}$. The values, $\hat{p}$, are **estimators** of the true population proportion, p.

1.1 Numerical representation of sample proportions: *n* = 50



$n = 50$

number of 'natural' climate change belivers in the sample(s)

x = {19.,20.,19.,15.}

sample propotion =

	A x	B
=	='count	
1	19.	
2	20.	
3	19.	
4	15.	
A	x:='co	

Navigate to Page 2.2. Adjust the slider value to $k = 1$, to simulate drawing a single random sample from a large population, where the value of the population proportion, p, is unknown to you. The number of people, x, in the sample, who believe in 'natural' climate change is displayed. Note that the values of x are displayed in the spreadsheet (in the column named x) and as a list in the left-hand panel.

Question 2

a. In the Math Box indicated, use the observed value of x to calculate a **point estimate** of the proportion of the population who believe in 'natural' climate change.

$\frac{x}{50}$, where *x* can be the variable symbol, or its value. The computed value is the point estimate.

b. How likely do you think it is that the point estimate, calculated above, ends up being identical to the true population proportion?

It is possible, but unlikely, that the point estimate is exactly equal to the true sample proportion.

On Page 2.2, press ctrl+tab until the spreadsheet panel of the split screen is selected. Press ctrl+R to simulate drawing a different sample of size 50 from the population.

Question 3

a. Use 3 more observed value of x, obtained after pressing ctrl+R, to calculate 3 more **point estimates** of the proportion of the population who believe in 'natural' climate change.

It is likely that each point estimate has a different value.

b. How much variability is there between point estimates obtained from the different samples?

Answers will vary.

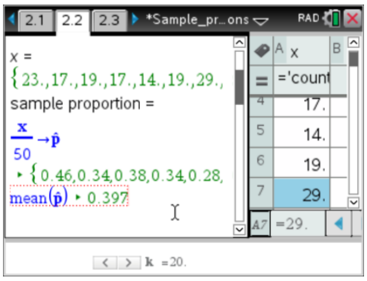
c. Insert a new Mathbox (ctrl+M) in the left-hand panel. In this Math Box, calculate the mean of the four point estimates.

$\frac{\text{mean}(x)}{50}$, or similar, to compute the mean sample proportion.

d. Explain why the mean of the four sample proportions is likely to give a better estimate of the population proportion than the individual point estimates.

It is likely that some point estimates are greater and some less than the true population proportion. The mean should 'balance' the over-estimates and under-estimates

1.2 Simulating multiple samples: *n* = 50



2.1 2.2 2.3 *Sample_pr...ons RAD

x =
$\{23.,17.,19.,17.,14.,19.,29.,$
sample proportion =
$\frac{x}{50}\rightarrow\hat{p}$
$\blacktriangleright \{0.46,0.34,0.38,0.34,0.28,$
$\text{mean}(\hat{p}) \blacktriangleright 0.397$

	A x	B
=	='count	
4	17.	
5	14.	
6	19.	
7	29.	

A7 =29.

k =20.

On Page 2.2, adjust the slider value to $k = 20$, which simulate drawing 20 random samples of size 50. The number of 'natural' climate change believers in each simulated sample is displayed in the list and in the spreadsheet.

In the Math Box from Question 2a, calculate the sample proportions by inputting the expression: $\frac{x}{50}$. Store these proportions as a variable, $\hat{p}$; i.e. press [sto→]([ctrl][+][var]), then [∞β°]([ctrl]+[📖]) and select symbol, $\hat{p}$. Press [enter] to execute the calculation.

Question 4

Compare the largest and smallest number of 'natural' climate change believers in the samples, and their corresponding sample proportions. Calculate the percentage difference between these two proportions (i.e. the magnitude of the difference, divided by the maximum value and multiplied by 100).

Calculation of the form $\frac{100(\hat{p}_2 - \hat{p}_1)}{\hat{p}_2}$, $\hat{p}_2 > \hat{p}_1$.

In the Math Box from Question 3c, calculate the mean value of all the sample proportions, by inputting the expression: $\text{mean}(\hat{p})$, and pressing [enter] to execute the calculation.

Incrementally increase the number of samples drawn by changing the value of k to 40, 60, 80 … 200.

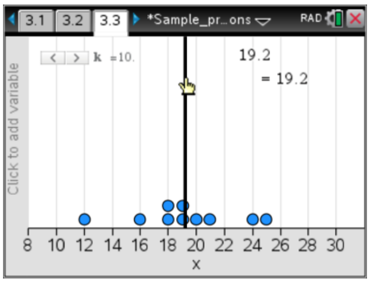
Question 5

Do you notice any pattern in the mean of the sample proportions, as the number of samples, k, increases?

Answers will vary. Pattern unlikely, with the mean likely to fluctuate

1.3 Graphical representation of sample proportions: *n* = 50

Navigate to Page 3.1. In Problem 3 you will simulate drawing samples of size 50 from a large population where the population proportion of 'natural' climate change believers is known to be $p = 0.4$. Of course, in practice we would not carry out sampling if we already knew the value of the population proportion. In this activity, the purpose of sampling is to understand how sampling behaves.



On Page 3.2 you can view numerical representations of the simulations, similar to that of Problem 2. However, in the pages that follow, you will see the graphical representation of these results.

Navigate to Page 3.3 and adjust the slider value to $k = 1$, to simulate drawing a single random sample of size 50, from a population where $p = 0.4$. The number of people, x, in the sample, who believe in 'natural' climate change is displayed graphically. Adjust the slider value to simulate drawing 2, 3, … 10 samples from this population. The vertical line indicates the mean value.

Author: Frank Moya

Question 6

a. For your 10 samples ($k = 10$), what are the observed maximum, minimum and mean number of 'natural' climate change believers?

Answers will vary.

b. In a sample of 50 people, how many 'natural' climate change believers would you **expect,** if the population proportion is 0.4? Why isn't this expected number obtained each time you draw a sample?

Expected number = $50 \times 0.4 = 20$. There is variability between samples - 0.4 is the expected long run sample proportion.

c. Adjust the slider value to $k = 200$. In how many of the 200 samples was the observed value of x equal to the expected value?

Answers will vary.

1.4 The sample count as a random variable

In the preceding activities, you have seen that the value, x , counts the number of 'natural' climate change believers in a sample. The count can therefore be considered as random variable X , whose values, x , vary from sample to sample. Furthermore, X is **binomially distributed**, with **parameters** n (sample size) and p (population proportion); that is $X \sim \mathrm{Bi}(n, p)$.

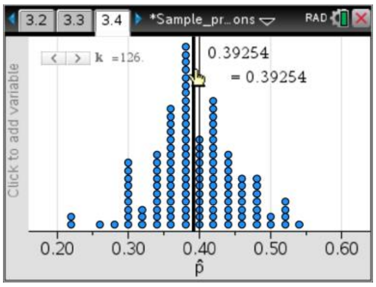
Question 7

Assume that counting a 'natural' climate change believer denotes 'success'. Explain why X is a *binomial* random variable, with the population proportion being the probability of 'success' in a single trial.

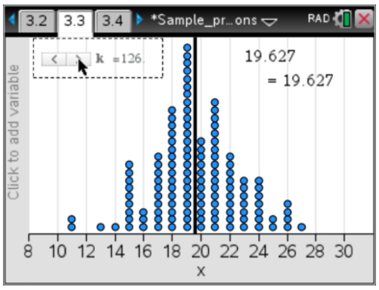
A single trial consists of determining whether a randomly chosen individual believes in 'natural' climate change. X is binomial because the individual either believes in 'natural' climate change ('success'), or they do not ('failure')

1.5 The sample proportion as a random variable

Navigate to Page 3.4, and incrementally increase the number of samples observed, by adjusting the value of k , up to $k = 200$. The observed sample proportion for each of the k samples is displayed graphically. The vertical line shows the mean of the sample proportions.



If you navigate back to Page 3.3, you will observe an analogous graphical pattern for the 'success' count, for the samples of size 50.



Question 8

a. Explain why the graphs in Pages 3.3 and 3.4 have identical shapes.

The two graphs have identical shapes because the two quantities, *x* and $\hat{p}$ are related by a constant ratio

b. For a particular sample, what is the relationship between the sample count of 'successes', x , and the sample proportion of 'successes', $\hat{p}$.

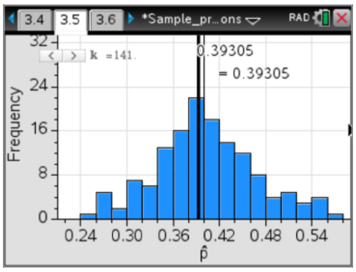
$$\hat{p} = \frac{x}{n}$$

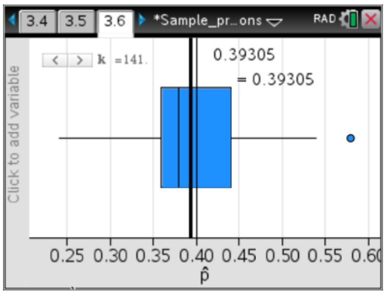
You will have observed that each time you take a random sample from a large population, the sample count of 'successes', x, and the sample proportion of 'successes', $\hat{p}$, vary from sample to sample; furthermore, for a particular sample of size n, $\hat{p}=\dfrac{x}{n}$.

Just as the count can be considered a random variable X, whose values, x, vary from sample to sample, the sample proportion can, likewise, be considered a random variable $\hat{P}$, whose values, $\hat{p}$, vary in the same fashion as x. Therefore, considering the random variables, $\hat{P}=\dfrac{X}{n}$.

2. Expectation and standard deviation of the sample proportion

Navigate to Page 3.5, and incrementally increase the number of samples observed, by adjusting the value of k, up to $k=200$. The observed frequencies of sample proportions for the k samples is displayed as a histogram. Navigate to Page 3.6 to view the same data, displayed as a boxplot.





From these, we observe that the sample proportion has a **distribution**, for which a mean and standard deviation can be computed.

Below we will consider the theoretical expectation (mean) and standard deviation of the sample proportion.

Question 9

Recall that $\hat{P}=\dfrac{X}{n}$, where $X\sim \mathrm{Bi}(n,p)$. Also, for a binomial random variable, $\mathrm{E}(X)=\mu=np$.

Further, if $Y=aX$, where a is a constant, then $\mathrm{E}(Y)=a\mathrm{E}(X)$

a. Using the above information, show that $\mathrm{E}(\hat{P})=p$.

$$\mathrm{E}(\hat{P})=E\left(\frac{X}{n}\right)$$

$$=\frac{1}{n}E(X)$$

$$=\frac{1}{n}\times np$$

$$\mathrm{E}(\hat{P})=p$$

b. Explain the significance of the result obtained in **part a.** above.

The mean of the distribution of sample proportions gives the population proportion. As the distribution is centred at *p*, in the long run, on average, the sample proportion will neither over-estimate nor under-estimate the value of *p*.

Question 10

For a binomial random variable, $\operatorname{var}(X) = np(1-p)$.

Further, if $Y = aX$, where a is a constant, then $\operatorname{var}(Y) = a^2\operatorname{var}(X)$.

a. Using the above information, and Question 9, show that the standard deviation of the sample proportion, $\operatorname{SD}(\hat{P}) = \sqrt{\dfrac{p(1-p)}{n}}$.

$$\begin{aligned}\operatorname{var}(\hat{P}) &= \operatorname{var}\left(\frac{X}{n}\right)\\ &= \frac{1}{n^2}\operatorname{var}(X)\\ &= \frac{1}{n^2}\times np(1-p)\\ &= \frac{p(1-p)}{n}\end{aligned}$$

b. The variance and standard deviation of $\hat{P}$ have n in the denominator. Explain the implications of this, in terms of the spread of the distribution of $\hat{P}$.

A key implication is that as $n \to \infty$ (large sample size), $\operatorname{SD}(\hat{P}) \to 0$ (the spread of the distribution of $\hat{P}$ is small). Since the distribution is centred at *p* (see Q. 10b), a small spread makes it more likely that the sample proportion is close to the true value of the population proportion.

The concepts introduced in this activity are explored further in the activity titled ***Distribution of sample proportions***.

Author: Frank Moya