Distribution of Sample Proportions

Teachers Teaching with Technology™
Professional Development from Texas Instruments

Probability and statistics

Answers & Teacher Notes

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

TI-Nspire | Investigation | Student | 90 min

Introduction

From previous activity: This activity assumes knowledge of the material covered in the activity ***Introducing sample proportions***, which introduced **statistical inference**. This involves using a **sample statistic**, in this case the proportion of a sample with a particular attribute to **estimate** the corresponding **population parameter**: the true population proportion. In that activity, it was assumed the count X of 'successes' in a sample is a binomial random variable, $X \sim \text{Bi}(n, p)$, where n is the sample size and p is the proportion of 'successes' in the population. The sample proportion is also a random variable, $\hat{P} = \frac{X}{n}$.

It was shown that the mean of $\hat{P}$, $\text{E}(\hat{P}) = p$, and the standard deviation of $\hat{P}$, $\text{SD}(\hat{P}) = \sqrt{\frac{p(1-p)}{n}}$.

Overview of this activity: In this activity you will investigate further the distribution of the sample proportion, $\hat{P}$; in particular, the effect of changing the sample size and the population proportion. A model for the distribution of sample proportions is also explored.

Why simulate repeated random sampling?

In statistical inference we see how trustworthy a procedure is by asking what would happen if we repeated it many times. This leads to the idea of the **sampling distribution** of the statistic.

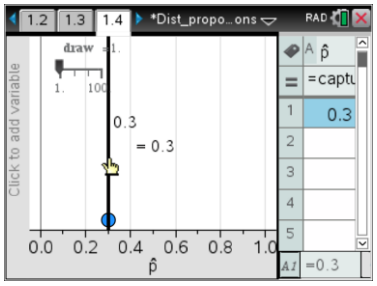
The **sampling distribution** of the sample proportion is the probability distribution of values taken by this statistic in all possible samples of the same size from the same population.

In practice, when conducting opinion polls and the like, it isn't feasible to repeat the sampling many times. However, the use of simulated random sampling, using TI-Nspire, allows us to imagine the results of all the possible random samples that the pollster didn't take - as illustrated in the following exercise.

Mobile Internet Subscriber Simulation

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

Australian Bureau of Statistics (ABS) data shows that in 2015, 50% of Australian internet service subscribers had a mobile wireless connection[1]. The population proportion of 'mobile' subscribers in this population is therefore known to be: $p = 0.5$.

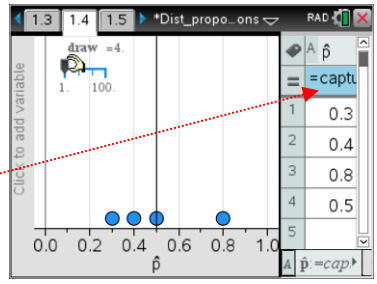


[1] http://www.abs.gov.au/ausstats/abs@.nsf/mf/8153.0

Author: Frank Moya

Navigate to Page 1.4. Assume that the single result shown represents the sample proportion $\hat{p}$ of 'mobile' subscribers from a surveyed random sample of 10 internet subscribers ($n=10$), drawn from the large population for which: $p=0.5$.

We can imitate carrying out the survey many more times using random number simulation. Each slider increment will draw a new random sample, and the sample proportion for the sample will be added to the spreadsheet and to the graph. Add 99 samples, so that the slider value is 100. To **reset** the simulation, set the slider value to 1. Then click on the spreadsheet cell 'A=' and press enter.



Question 1

a. What are the lowest and highest observed values of $\hat{p}$?

Answers will vary

b. What is the modal value of sample proportions?

Answers will vary

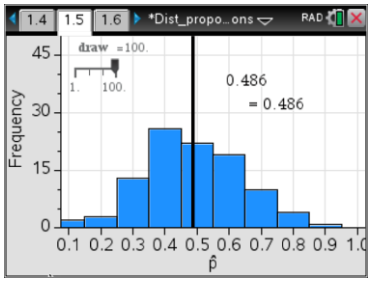
c. What is the mean of the sample proportions (shown by the vertical line) for the 100 samples?

Answers will vary, but should be close to 0.5

d. How close is the mean of the sample proportions to the population proportion?

Answers will vary, but should be close to sample and population proportions are likely to be close.

Reset the simulation and **navigate to Page 1.5**. Use the slider to repeat the simulation of 100 samples. A histogram of the distribution of sample proportions will emerge. **Navigate to Page 1.6**, where the summary statistics for the histogram are calculated.



Question 2

Use the histogram, dot plot (Page 1.4) and summary statistics to describe, as fully as possible, the distribution of sample proportions of $n=10$ and $p=0.5$.

The distribution is likely to be centred close to 0.5, and likely to be quite spread out, in the interval [0, 1]. The shape might be fairly symmetrical. The mode and mean are likely to be near 0.5.

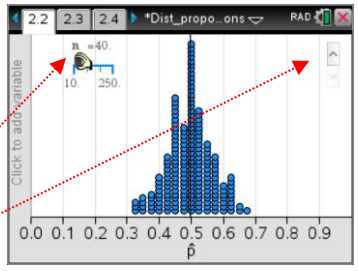
Changing the sample size

Suppose that for the mobile internet subscriber survey, we change the size of the sample that is drawn from the population with $p=0.5$.

Question 3

Predict how the shape of the distribution of the sample proportion will change, as the samples size increases.

Answers will vary



Navigate to Page 2.2. When you select a value of n with the slider, 200 random samples of that sample size are drawn. The spinner allows you to redraw a new set of 200 samples of that size.

Using the slider, progressively change the sample size from $n=10$ to $n=250$.

Author: Frank Moya

Navigate to Page 2.3. Repeat the above on this page.

Question 4

a. As the sample size increases, what aspects of the distribution of sample proportions remain the same?

As *n* increases, the mean and mode of the distribution are likely to remain near 0.5. The distribution is also likely to remain fairly symmetrical

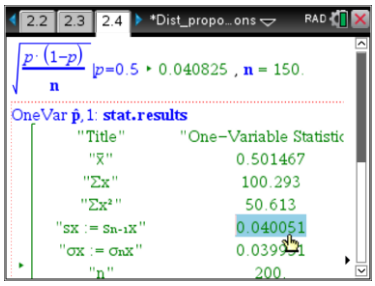
b. As the sample size increases, describe how the distribution changes.

As *n* increases, the most prominent changes are likely to be a decrease in the spread of the distribution, and a decrease, and an increase in possible values that $\hat{p}$ can take (e.g. for *n* = 10, $\hat{p}$ can only change in increments of 0.1, but for *n* = 100, $\hat{p}$ can change in increments of 0.01).

c. How did your prediction in Question 3 compare with what you actually observed?

Answers will vary.

Navigate to Page 2.4, where the theoretical standard deviation of $\hat{P}$, and observed summary statistics for the 200 samples are displayed.



Question 5

a. For sample sizes of $n = 10, 50, 100 \text{ and } 200$, record the theoretical and observed standard deviation for the distribution of sample proportions, correct to 4 decimal places.

	Theoretical SD	Observed SD
$n = 10$	0.1581	Answers will vary, but are likely to be close to the corresponding theoretical values.
$n = 50$	0.0707	
$n = 100$	0.0500	
$n = 200$	0.0354	

b. What aspect of the distribution is measured by the standard deviation?

$\text{SD}\left(\hat{P}\right)$ is a measure of the spread of the values of $\hat{p}$

c. What trend do you observe in the value of the standard deviation as the sample size increases?

As *n* increases, $\text{SD}\left(\hat{P}\right)$: the spread of the values of $\hat{p}$, decreases.

d. How does the formula for the standard deviation of $\hat{P}$ explain this trend?

$SD\left(\hat{P}\right) = \sqrt{\dfrac{p(1-p)}{n}}$, therefore, as $n \to \infty$ (large sample size), $\text{SD}\left(\hat{P}\right) \to 0$ (the spread of the distribution of $\hat{P}$ is small).

e. In terms of using a sample proportion to estimate the true population proportion, explain why a small standard deviation for the distribution of $\hat{P}$ is desirable.

Since the distribution is centred at *p*, a small spread makes it more likely that the sample proportion is close to the true value of the population proportion.

Changing the population proportion

In this section, you will explore the sampling distribution for samples of a fixed size, drawn from populations with different population proportions, within the interval $0.05 \le p \le 0.95$.

Author: Frank Moya

Question 6

Suppose that you draw repeated random samples of size 10, from different populations for which the population proportions are $p = 0.5, 0.6, 0.7, 0.8, 0.9$.

a. Predict how the shape and features of the distribution of sample proportions will change, as the value of p **increases**.

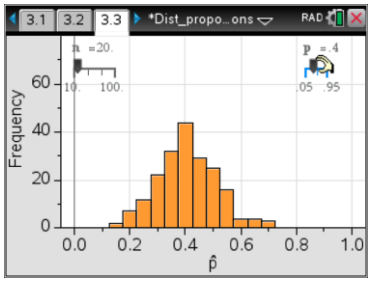
Answers will vary

Suppose that you draw repeated random samples of size 10, from different populations for which the population proportions are $p = 0.5, 0.4, 0.3, 0.2, 0.1$.

b. Predict how the shape and features of the distribution of sample proportions will change, as the value of p **decreases**.

Answers will vary

Navigate to Page 3.2. Set the sample size to $n = 10$ and use the other slider to systematically increase the population proportion from $p = 0.05$ to $p = 0.95$. **Navigate to Page 3.3** and observe the same data in a histogram.



Question 7

a. For what value(s) of p is the distribution least symmetrical?

Least symmetrical for at or near $p = 0.05$ and $p = 0.95$.

b. For what value(s) of p is the distribution most symmetrical?

Most symmetrical for at or near $p = 0.5$

c. Apart from symmetry, what other aspects of the distribution change as the value of p changes?

- The skew of the distribution – positively skewed for p < 0.5 and negatively skewed for p > 0.5. The amount of skew increases as the value of p moves away from 0.5.
- The spread of the distribution is greatest for p = 0.5, and the spread decreases as the value of p moves away from 0.5.

Set the sample size to $n = 20$ and systematically increase the population proportion from $p = 0.05$ to $p = 0.95$. Repeat for $n = 50$ and $n = 100$. As you systematically change values, **navigate to Page 3.4**, where the theoretical standard deviation of $\hat{P}$ is calculated.

Question 8

a. Use the results from Page 3.4 to complete the table below, correct to 4 decimal places.

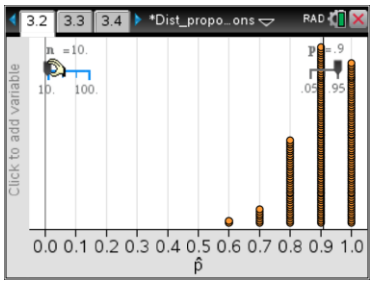
	$p = 0.05$	$p = 0.25$	$p = 0.5$	$p = 0.75$	$p = 0.95$
$n = 10$	0.0689	0.1369	0.1581	0.1369	0.0689
$n = 100$	0.0218	0.0433	0.0500	0.0433	0.0218

b. What trends do you notice in the table of values? How do these trends accord with the observed changes to the shape of the distribution of $\hat{P}$?

- For a given sample size, the $\mathrm{SD}(\hat{P})$ is greatest for $p = 0.5$
- For a given sample size, the value of $\mathrm{SD}(\hat{P})$ is symmetrical about $p = 0.5$
- The above dot points accord with the observed spread of the distribution of $\hat{P}$.

Author: Frank Moya

Focus on the population proportions *p* = 0.1 and *p* = 0.9



You will now examine more closely the distribution of $\hat{P}$ for $p = 0.1$. **Navigate to Page 3.2.** Set the population proportion to $p = 0.1$ and use the other slider to systematically increase the sample size from $n = 10$ to $n = 100$. **Navigate to Page 3.3** and observe the same data in a histogram. Repeat for $p = 0.9$

Question 9

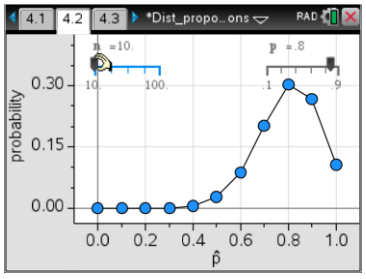
a. For what value(s) of sample size n is the distribution least symmetrical?

For *p* = 0.1 or *p* = 0.9, the distribution of $\hat{P}$ is least symmetrical for *n* = 10

b. For what value(s) of n is the distribution most symmetrical?

For *p* = 0.1 or *p* = 0.9, the distribution of $\hat{P}$ is most symmetrical for *n* = 100.

Theoretical distribution of $\hat{P}$ from Bi(*n*,*p*)



Navigate to Page 4.3. Recall that $\hat{P} = \frac{X}{n}$, where $X \sim \text{Bi}(n, p)$. In the spreadsheet, the probabilities of 0, 1, … n successes are calculated for the theoretical distribution of $\text{Bi}(n, p)$. **Navigate to Page 4.2**. The sliders allow the values of the parameters for $\text{Bi}(n, p)$ to be varied, and the resultant proportion of successes is plotted against their probabilities.

On Page 4.2, set the value of the population proportion to $p = 0.1$. Systematically increase the sample size from $n = 10$ to $n = 100$.

Question 10

What changes do you observe in the plot as the sample size increases?

The number of values that $\hat{p}$ can take increases, and the 'business' end of the graph (i.e. the end where probabilities are not near zero) becomes less skewed; more symmetrical.

Repeat the previous procedure for population proportion of $p = 0.2$ to $p = 0.9$.

Question 11

What are some similarities and differences, for corresponding values of n and p, between the plot on Page 4.2, and the graph on Page 3.2?

The basic shapes and other features of the two representations are essentially the same, for corresponding combinations of *n* and *p* values.

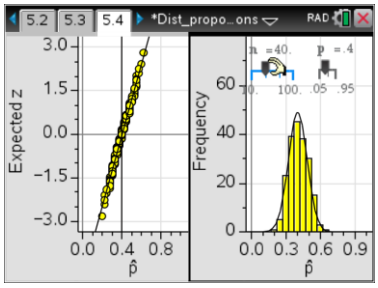
Modelling how sample proportions vary from sample to sample

On Pages 3.2 you observed that as the sample size increases, there are more possible values of the sample proportion. Consequently, for large sample sizes, the dot plot starts to resemble a continuous distribution. You also observed that for larger values of n, the distribution of $\hat{P}$ was fairly symmetrical; not just for population proportions close to 0.5, but also for, say, $p = 0.1$.

Author: Frank Moya

In this section, you will explore the viability of modelling the discrete sampling distribution of $\hat{P}$ with a continuous 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 $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$.

Navigate to Page 5.2. A normal pdf (probability density function) curve is superimposed on the histogram of the distribution of $\hat{P}$. Use the sliders to adjust the sample size, n, and population proportion, p.



Question 12

Based on a visual comparison of the histogram and the corresponding normal pdf curve, for what values of n and p does the normal distribution appear to be a good fit?

For any given value of *p*, the best fit is likely to occur for larger sample sizes: *n* = 100.

For any given value of *n*, the best fit is likely to occur for values of *p* that are close to 0.5.

Navigate to Page 5.3, which shows a normal probability plot. You can adjust the values of n and p.

Question 13

What do you think this normal probability plot is showing?

Answers will vary. Hopefully, students will appreciate that it is a plot of actual values of $\hat{p}$ versus expected values, for a normally distributed random variable.

Navigate to Page 5.4, which combines Pages 5.2 and 5.3 into a single split page. In the left-hand panel, the normal probability plot is displayed, with the expected z on the vertical axis, where z is the standard normal random variable.

Question 14

a. What is the significance of the regression line shown in the normal probability plot?

The regression line shows what the plot will look like if $\hat{p}$ is normally distributed, i.e. it is the line for which actual values correspond to predicted values.

b. How does the normal probability plot tell you whether $\mathrm{N}\left(\mu_{\hat{p}}, {\sigma_{\hat{p}}}^2\right)$ is a good fit for the distribution of $\hat{P}$?

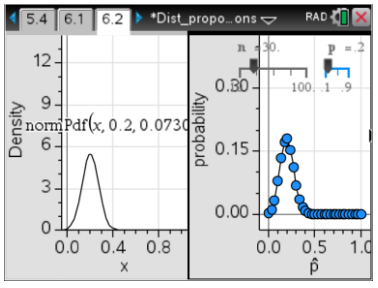
The closer the scatterplot is to the line, the better the fit, and the closer $\hat{P}$ is to normality.

c. Based on your observations of the normal probability plot, for what values of n and p is the normal distribution model most appropriate?

As with Question 12, for any given value of *p*, the best fit is likely to occur for larger sample sizes: *n* = 100. For any given value of *n*, the best fit is likely to occur for values of *p* that are close to 0.5.

Normal approximation to the theoretical distribution of $\hat{P}$ from Bi(*n*,*p*)

Navigate to Page 6.2. In split page, the left hand panel shows the same information as previously observed in Page 4.2: the theoretical distribution of $\hat{P}$, based on observations from the $\mathrm{Bi}(n, p)$ distribution. The right-hand panel shows the normal distribution with mean and standard deviation corresponding to those of $\mathrm{Bi}(n, p)$ - that is,



Author: Frank Moya

$N\left(\mu_{\hat{P}}, \sigma_{\hat{P}}^2\right)$, where $\mu_{\hat{P}} = E\left(\hat{P}\right) = p$ and $\sigma_{\hat{P}}^2 = \frac{p(1-p)}{n}$.

Adjust the values of n and p.

Question 14

a. As the values of n and p are varied, what are some similarities and differences between the normal and binomial distributions, displayed on Page 6.2.

For p > 0.5 or p < 0.5, the binomial distribution graph resembles the normal graph for larger values of *n*.

b. From the graphs on Page 6.2, for what values of n and p is the normal distribution model of the binomial distribution most appropriate? Does this accord with your observations on Page 5.4?

The behaviour of these graphs is likely to be analogous to the graphs on Page 5.4.

Question 15

Write a brief summary in point form, of what you have learnt about the distribution of sample proportions, as a result of doing this activity.

Answers will vary.

Follow up on this activity

In this activity we considered the use of proportions from samples to estimate population proportions. The follow-up activity, ***Confidence intervals for proportions***, explores obtaining intervals within which we are reasonably sure (with a given level of confidence) that the value of the population proportion will lie.

END OF ACTIVITY

Author: Frank Moya