

# What Percent of the Continental US is Within One Mile of a Road?



Sara Stoudt  
Department of Mathematics  
and Statistics, Smith College

Yue Cao  
Department of Mathematics  
and Statistics, Smith College

Dana Udwin  
Department of Mathematics  
and Statistics, Smith College

Nicholas J. Horton  
Department of Mathematics  
and Statistics, Amherst College  
[nhorton@amherst.edu](mailto:nhorton@amherst.edu)

**Published: March 2014**

## Overview of Lesson

This lesson asks students to sample random latitude and longitude coordinates from the contiguous United States, display these locations on a map, and then determine whether or not the point is within one mile of a road. Students utilize this sampling method to estimate an unknown parameter (the proportion of the continental United States within one mile of a road). The question of how much of the United States lies within one mile of a road provides an interesting context for statistics education because roads have an ecological impact on the surrounding environment (Murphy, 2013; USGS, 2005).

Students use software to generate a set of samples, display each of these locations on a Google map with a circle of radius 1 mile centered at each point, and record relevant information about each location.

These data are used to calculate an estimate of the proportion of the continental United States within one mile of a road and a 95% confidence interval from the students' samples of locations (generally each pair of students can collect about 20 in a class period). Each pair of students can also share their number of successes out of the total number of observations that fell within the contiguous United States. The teacher can then tally the class collective number of successes out of all observations within the contiguous United States and calculate the class-wide estimate of the proportion of the entire United States within one mile of a road (and associated 95% confidence interval for this parameter).

Students observe how an innovative sampling method can shed light on an interesting question and demonstrate the impact that sample size has on the width of the confidence interval.

## **GAISE Components**

This investigation follows the four components of statistical problem solving put forth in the *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report* (Franklin et al, 2007). The four components are: formulate a question, design and implement a plan to collect data, analyze the data by measures and graphs, and interpret the results in the context of the original question. This is a GAISE Level C activity.

## **Common Core State Standards for Mathematical Practice**

1. Make sense of problems and persevere in solving them.
2. Reason abstractly and quantitatively.
3. Construct viable arguments and critique the reasoning of others.
4. Model with mathematics.
5. Use appropriate tools strategically.
6. Look for and express regularity in repeated reasoning.

## **Common Core State Standards Grade Level Content (High School)**

S-IC. 1. Understand statistics as a process for making inferences about population parameters based on a random sample from that population.

S-IC. 4. Use data from a sample survey to estimate a population mean or proportion; develop a margin of error through the use of simulation models for random sampling.

## **NCTM Principles and Standards for School Mathematics**

### **Data Analysis and Probability Standards for Grades 9-12**

**Formulate questions that can be addressed with data and collect, organize, and display relevant data to answer them:**

- understand the differences among various kinds of studies and which types of inferences can legitimately be drawn from each;
- know the characteristics of well-designed studies, including the role of randomization in surveys and experiments;
- compute basic statistics and understand the distinction between a statistic and a parameter.

**Select and use appropriate statistical methods to analyze data:**

- **for univariate measurement data, be able to display the distribution, describe its shape, and select and calculate summary statistics.**

**Develop and evaluate inferences and predictions that are based on data:**

- use simulations to explore the variability of sample statistics from a known population and to construct sampling distributions;
- understand how sample statistics reflect the values of population parameters and use sampling distributions as the basis for informal inference.

**Understand and apply basic concepts of probability:**

- use simulations to construct empirical probability distributions.

**Prerequisites**

Students will have knowledge of sampling techniques, estimating unknown parameters, and confidence intervals.

**Learning Targets**

Students will observe differences in sample proportions and corresponding confidence intervals in accordance with having a variety of samples. They will understand the impact of sample size on confidence intervals. They will be able to interpret a confidence interval in the context of the problem at hand.

**Time Required**

1 class period.

**Materials Required**

One computer for every group of two students, with either R or RStudio (open source statistical software, see the RESOURCES link provided with this lesson) and the "mosaic" package installed as well as a copy of the Activity Worksheet (provided at the end of this lesson).

**Instructional Lesson Plan****The GAISE Statistical Problem-Solving Procedure****I. Formulate Question(s)**

One of the largest human constructions on Earth is the road network of the United States (USGS, 2005). Roads provide a key role as the circulatory system of our economy. As the US Geological Survey (USGS) notes, "Spatial relationships between the network and intervening roadless areas are important to ecological and hydrological resources."

Ask students how the presence of a road affects the surrounding environment. Discuss with students how the USGS has shown that roads and traffic eliminate forest canopy, elevate the temperature, and introduce vehicular noise and pollution, among other effects. It is estimated that roads affect the ecology of at least 22% of the land area of the continental United States (USGS, 2005).

A conservation policy called *roadless area conservation* aims to limit road construction in order to halt negative environmental impact on designated public lands. The United States Forest Service has formalized the concept of Inventoried Roadless Areas as lands identified by governmental agencies without roads that could be candidates for roadless area conservation (Murphy, 2013). According to the United States Forest Service, inventoried roadless areas comprise only 2% of the land in the continental United States (US Forest Service, 2013).

Explain that it is impossible to know precisely the proportion of the continental United States that is within one mile of a road, but that sampling is one approach to estimating this parameter.

## II. Design and Implement a Plan to Collect the Data

Within R or RStudio, have the students run the commands:

```
require(mosaic)  
source("http://www.amherst.edu/~nhorton/roadless-setup.R")
```

This assumes that the "mosaic" package is installed (see the RESOURCES link for further details).

This code will generate a list of twenty random pairs of latitude and longitude coordinates in and around the contiguous United States. The resulting sampled locations are saved as an R object called "myroadless". Note that sampling from latitude and longitude requires accounting for the spherical shape of the Earth. The `rgeo()` function from the "mosaic" package takes this into account.

It is helpful to remind students that because the perimeter of the United States is quite irregular, it is not straightforward to ensure that samples are taken from this exact region. As a result, judgment from the student will be needed to determine first whether the point is within the continental US, and only if it is, to determine whether the location is within a mile of a road. Oceans do not count as within the continental United States, but lakes contained within the United States do. If the instructor prefers to have each group have 20 samples within the continental United States, they can have students continue to sample until each student selects 20 locations that are within the US.

To display the values, type the command:

```
print(myroadless)
```

```
> print(myroadless)
```

	X	sample	latitude	longitude	withinContinent	within1mile	location
1	1	1	29.5320	-80.1230	NA	NA	NA
2	2	2	40.7566	-73.9863	NA	NA	NA
3	3	3	35.0140	-104.0230	NA	NA	NA
4	4	4	34.1224	-118.1233	NA	NA	NA
5	5	5	49.2782	-122.1340	NA	NA	NA
6	6	6	44.8976	-123.4023	NA	NA	NA
7	7	7	44.2288	-114.4680	NA	NA	NA
8	8	8	40.8641	-74.2761	NA	NA	NA
9	9	9	26.6087	-94.2477	NA	NA	NA
10	10	10	36.4753	-106.3966	NA	NA	NA
11	11	11	29.9672	-75.6496	NA	NA	NA
12	12	12	48.8301	-77.7243	NA	NA	NA
13	13	13	27.8969	-124.1354	NA	NA	NA
14	14	14	49.3266	-73.4116	NA	NA	NA
15	15	15	32.2116	-118.7470	NA	NA	NA
16	16	16	31.8356	-68.4139	NA	NA	NA
17	17	17	47.3755	-117.0904	NA	NA	NA
18	18	18	42.6263	-77.3050	NA	NA	NA
19	19	19	31.1202	-96.3120	NA	NA	NA
20	20	20	25.6109	-90.0864	NA	NA	NA

Figure 1. Sample data frame.

This sample data frame can be found in the file:

samplemyroadless.csv

If the teacher wants to collate individual location results from the class, then some additional effort is required. Each student's randomly sampled data frame needs to be edited as the student works through the twenty locations to determine whether: (1) it is within the continental US (coded 1 for TRUE and 0 for FALSE), and (2), if so, whether it is within 1 mile of a road (coded 1 for TRUE and 0 for FALSE: values not within the continental US should be left blank).

The "location" field should be filled in with the name of the state (if within the US), the body of water (Pacific Ocean, Gulf of Mexico or Atlantic Ocean), or country (Mexico, Canada) for other locations.

The students will have a handout with a table of 20 locations to take down their results as they proceed. If students are also submitting their results to the teacher in electronic form, they can use this paper record to populate the values in their own copy of "myroadless".

To determine whether their samples are within 1 mile of a road, students will repeatedly call the "getLocation()" function. This takes a single argument: the location (between 1 and 20) to plot. Figure 2 displays the results for a point in the middle of the ocean (individual results will vary):

getLocation(1)

When the "getLocation(1)" command is run, a pop-up window should open that contains a Google map centered on the first sample location (Figure 2). [Hint: Be sure to have everyone in the class confirm that the window displays, as it may be necessary to disable pop-up blockers during this activity.] If a student has trouble seeing the circle of radius one mile, have them try zooming in or out to see if it appears.

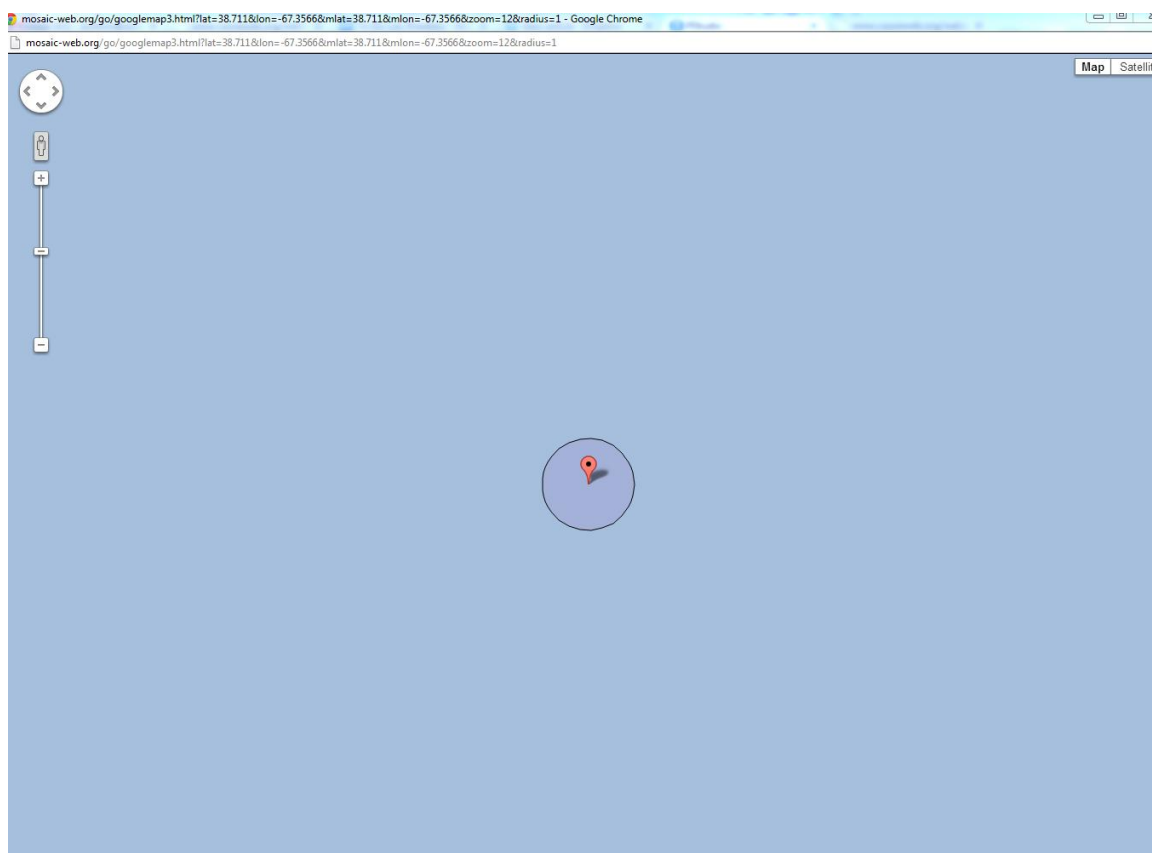


Figure 2. Location outside the contiguous United States.

A circle with radius one mile is drawn around the location. It is apparent that the location is not within the contiguous United States. Zooming out (by clicking on the "minus" sign on the left) reveals that we are off the coast of Florida (Figure 3). This location would be coded as 0 in the withinContinent column, with the within1mile column left blank, and a value of "Atlantic Ocean" in the location column.



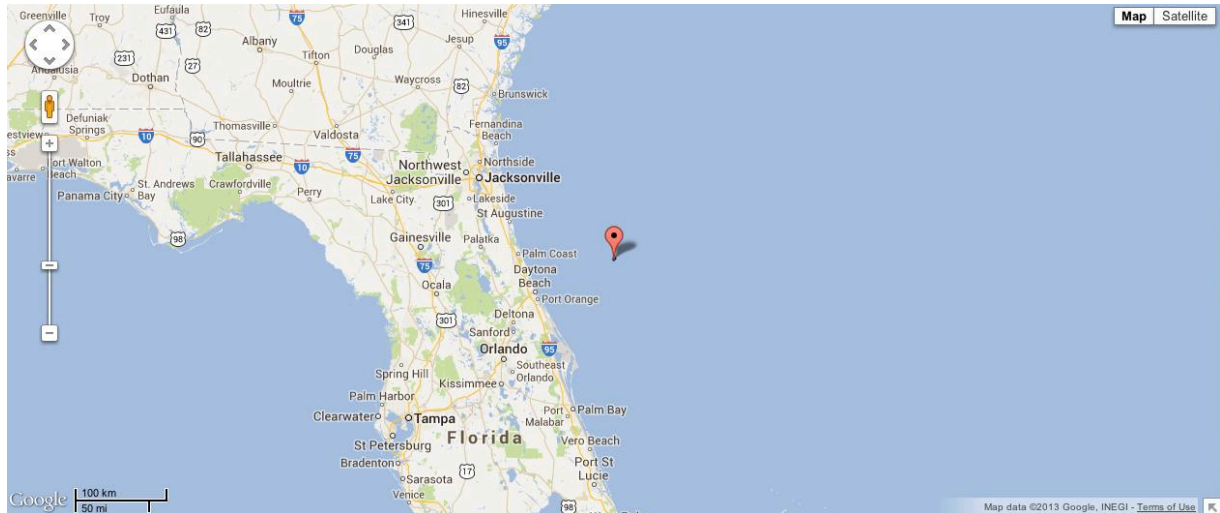


Figure 3. Zoomed out location outside of the United States.

Calling "getLocation(2)" reveals a more interesting location. Here we are on Times Square, well within one mile of a road (Figure 4). This location would be coded as 1 in the withinContinent column, 1 in the within1mile column, and a value of "NY" in the location column.

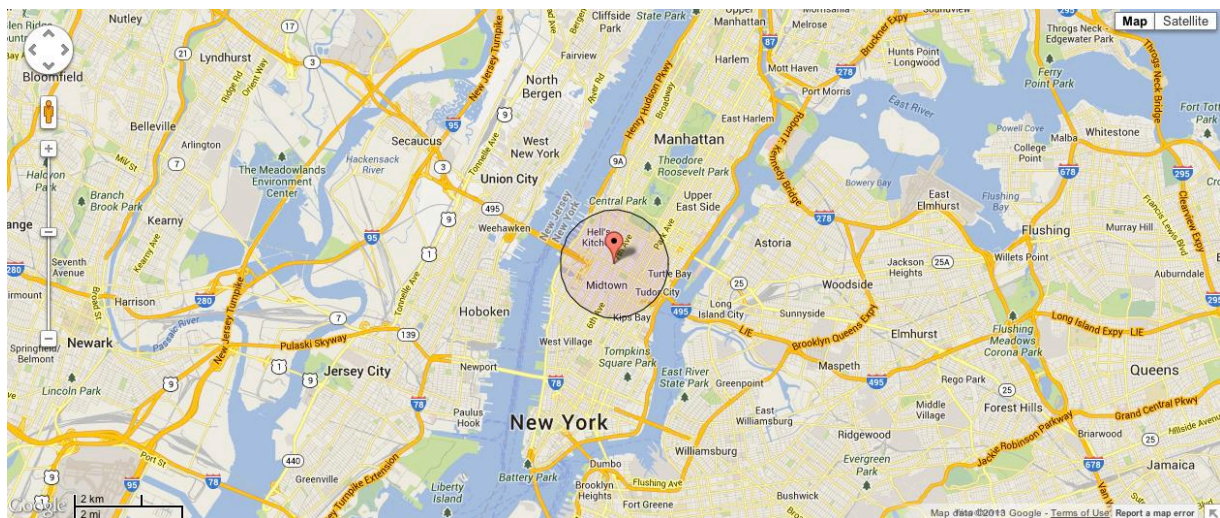


Figure 4. Location within one mile of a road.

Calling "getLocation(3)" gives us a roadless location (Figure 5). Zooming out we see that the location is near Santa Rosa Lake State Park in New Mexico (Figure 6). This location would be coded as 1 in the withinContinent column, 0 in the within1mile column, and a value of "NM" in the location column.



Figure 5. Roadless location.

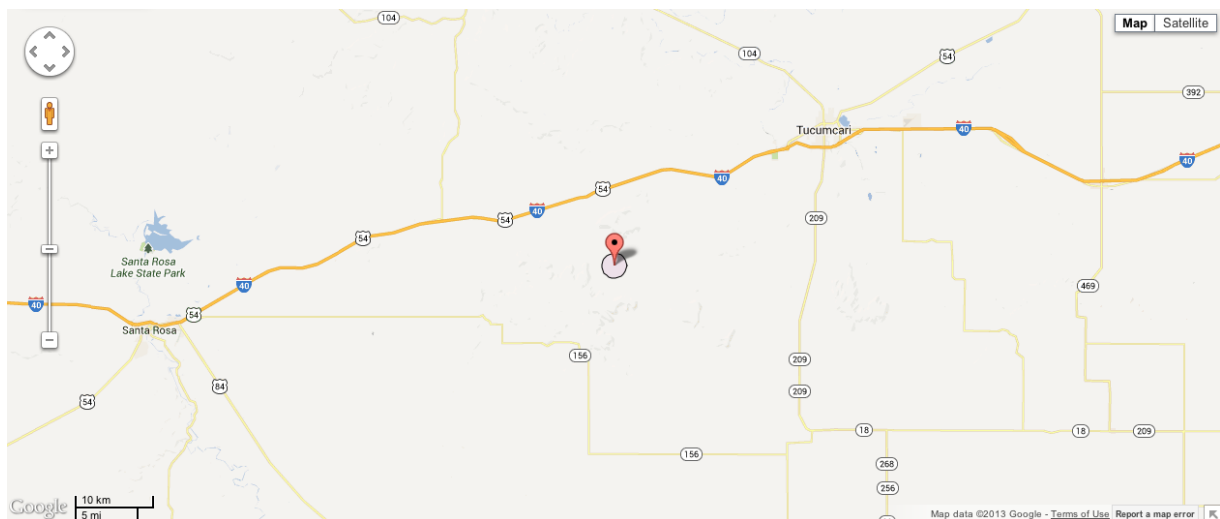


Figure 6. Zoomed out roadless location.

Students should update their records with information about the third sample location. After recording information for the rest of the twenty locations in a table provided in their worksheet (see Figure 7), students should continue working through the lesson template to analyze the data.

sample	withinContinent	within1mile	location
1	0		Atlantic Ocean
2	1	1	NY
3	1	0	NM
4			
5			
6			
7			



8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			

Figure 7. Sample student table.

### III. Analyze the Data

Each student can use the number of sampled locations within one mile of a road (denote this as  $X$ ) and the total number of sampled locations that landed inside the continental United States (denote this as  $n$ ) to calculate the observed proportion  $\hat{p} = X/n$  as well as their 95% confidence interval for the actual proportion of the continental United States within one mile of a road.

One assumption required to use a normal approximation confidence interval is that the sample size is sufficiently large, which can be checked with the Success/Failure condition (DeVeaux, 2013). We need at least 10 “successes” and at least 10 “failures,” so:  $10 \leq n\hat{p}$  and  $10 \leq n(1 - \hat{p})$ , where  $n$  represents sample size and  $\hat{p}$  represents sample proportion. Because the number of sample locations that fall within the continental US is typically  $n=11-15$  for samples of 20 locations, this is insufficient for the use of normal approximation methods. Note that the sample size for each student will not meet this assumption, though the condition needed for the assumption may be met when combining data from a whole class.

Another approach to finding an interval for small  $n$  is the use of exact binomial methods. In addition to not requiring the large sample size, the binomial method also works when a student has no failures (e.g. all of their sampled points are within 1 mile of a road).

Students can use the exact binomial procedure in R/RStudio implemented using the "binom.test()" function to calculate an interval using exact methods (Clopper, 1934):

```
binom.test(X, n)
```

where " $X$ " and " $n$ " are the number of sampled locations within one mile of a road and the number of sampled locations that landed inside the United States with  $X \leq n \leq 20$ .

As an example, consider a student who found  $X = 8$  and  $n = 12$ , for a sample proportion of 0.667. The output is shown in Figure 8.

```
> binom.test(8, 12)

Exact binomial test

data:  x and n
number of successes = 8, number of trials = 12, p-value = 0.3877
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.3488755 0.9007539
sample estimates:
probability of success
      0.6666667
```

Figure 8. Confidence interval from exact binomial procedure.

Their interval ranges from 0.35 to 0.90, indicating the extreme uncertainty regarding the population proportion when only 12 trials are available.

An alternative to the exact binomial test is the Agresti Coull method (Agresti, 1998). Students can use this method by using the following code:

```
install.packages("binom")
require(binom)
binom.confint(X, n, conf.level = 0.95,
methods="agresti-coull")
```

where "X" and "n" are the number of sampled locations within one mile of a road and the number of sampled locations that landed inside the United States with  $X \leq n \leq 20$ .

See Figure 9 for a sample output for the same student who found  $X = 8$  and  $n = 12$ .

```
> binom.confint(8, 12, conf.level = 0.95, methods = "agresti-coull")
      method x  n      mean   lower   upper
1 agresti-coull 8 12 0.6666667 0.388011 0.864491
```

Figure 9. Confidence interval using Agresti-Coull Method.

This interval ranges from 0.39 to 0.86, still indicating the large amount of uncertainty regarding the population proportion when only 12 trials are available.

Each group of students then records their results on the chalkboard as well as handing over to their teacher their count of sampled locations within one mile of a road and their count of sampled locations that landed inside the United States. The teacher may then tally the class-wide totals and calculate an estimate of the proportion of the United States within one mile of a road based on the entire class's data. They may also utilize the `binom.test()` code as above, but with the class-wide tally, to calculate a (much narrower) 95% confidence interval, though the conditions for a normal approximation confidence interval will likely be met for a class-wide tally.

Having students draw their group's 95% confidence interval on the chalkboard, then aligning alongside them the 95% confidence interval calculated from the class-wide dataset provides a concrete and effective visual for comparing the impact that different samples and increased sample size has on the confidence interval (see Figure 10). The larger, blue confidence intervals are from individual students, and the small, red confidence interval on the right is from combining data from all students into one sample.

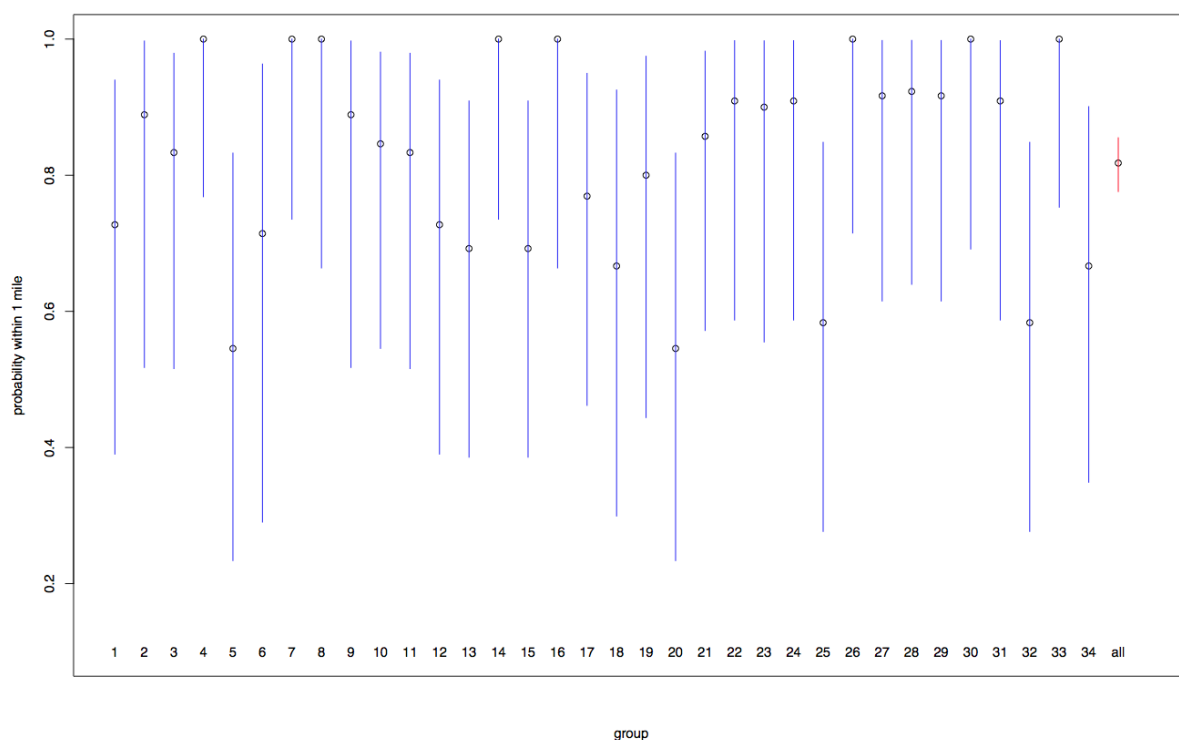


Figure 10. Group and class estimates with 95% confidence intervals for population proportion of locations within 1 mile of a road.

With some additional effort, by collecting individual student data, a graphical display of the class results can be created. Students will need to provide the specific locations and results to the teacher by saving their random locations and tallied results for each one to a spreadsheet. This can be done using the command: `write.csv(myroadless, file="myroadless.csv")`. Then they can edit the "myroadless.csv" file in a spreadsheet program and e-mail it to the teacher.

These data sets can be collated by the teacher to create a summary of the class on a map of the United States (see Figure 11 for a sample). The edge of the sampling area can be seen by the blue dots (those outside the continental US), with red dots indicating locations within a mile of a road, and the green those that are more distant. We see that nearly all of the roadless areas are in the western states (with one in northern Maine).

### **Roadless America Sample (green=roadless, red=within 1 mile of a road)**

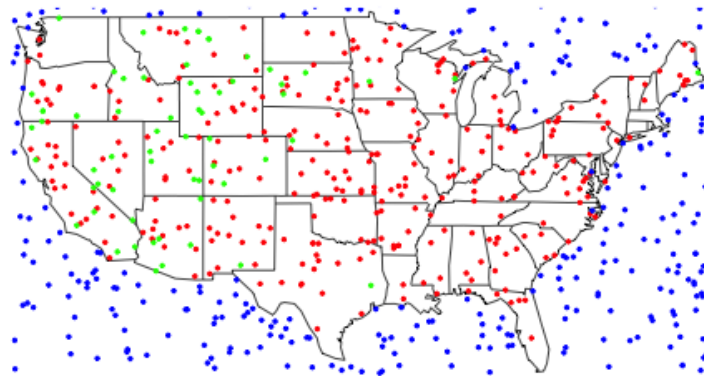


Figure 11. Data plotted on a United States map.

### **IV. Interpret the Results**

Ask students to use confidence intervals to describe the uncertainty associated with the proportion of the continental US within 1 mile of a road. Ask students: What does the confidence level mean? Students should note that the confidence level describes the uncertainty related to sampling, and if the confidence level is 95%, we would expect approximately 95% of the intervals to include the true proportion of the continental US within one mile of a road. Discuss the fact that the confidence interval for the population proportion generated from the whole class is narrower than the one from each individual group. Ask students: Using the provided formula, what do you observe when comparing those confidence intervals? Students should acknowledge that the sample size is negatively associated with the standard error through

the given formula:  $\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}^*(1-\hat{p})}{n}}$ , where  $\hat{p}$  represents the sample estimate, 1.96 is the

critical value for a 95% confidence level,  $n$  represents the sample size, and  $\sqrt{\frac{\hat{p}^*(1-\hat{p})}{n}}$  is the standard error of  $\hat{p}$ . As the sample size increases, the standard error decreases. Since the sample size of the whole class is much larger than that of each individual group, the confidence interval generated from the combined data for the whole class would have a narrower range than that of each group.

Ask students: What are the assumptions and the conditions under which we can “trust” a confidence interval using the large sample approximation? Students should acknowledge that we assume that each sample is independent and that we must check this condition with the randomization and 10% condition (detailed below). Students should also acknowledge that we assume that the sample size is large enough to make the sampling model approximately normal and that we must check this condition with the Success/Failure Condition (also detailed below). (DeVeaux, 2013).

Ask students: Are these conditions met? Is the sample independent? Students should note that the data values do not affect each other; the probability of a success for one coordinate does not affect the probability of a success for another.

Ask students: Were the data sampled at random? Students should acknowledge that randomization helps to ensure independence. Ask students: Does the sample exceed 10% of the population? Students should understand that if the sample exceeds this proportion, the probability of success of the later samples may be skewed. This is because the process of sampling without replacement changes the remaining population negligibly only if our sample size is less than 10% of the population. Hence if it exceeds this proportion, then the probability of “success” changes drastically. Students should note that our sample size does not exceed 10% of the population and that this condition is not really relevant here because essentially we have an infinite population. Also note, that technically we are sampling with replacement; there is nothing stopping the random number generator from choosing the exact same coordinate pair more than once in a sample. However, this is extremely unlikely in a sample of 20 coordinates taken from the entire area of the United States. Students should note that while the points may overlap on the map, that this is just due to the resolution of the plot.

Ask students: Is the sample large enough to make the sampling distribution approximately normal? Although our sample size cannot exceed 10% of the population, it also cannot be too small. More data is needed as the proportion approaches either 0 or 1. Our assumption that sample size is adequate can be checked with the Success/Failure condition. We need at least 10 “successes” and at least 10 “failures,” so:  $10 \leq n\hat{p}$  and  $10 \leq n(1 - \hat{p})$ , where  $n$  represents the sample size and  $\hat{p}$  represents the sample proportion. Students should note that even though the sample size of each group may not meet this assumption, this problem would be addressed if the data from the whole class is used. Note that since the sample size of each group is low, after dropping off all samples outside the continental US, it is possible for some groups to have a sample proportion that is exactly 1. This occurs when all of the group’s samples that were within the contiguous US were within one mile of a road.



### **Assessment**

A number of assessment questions might be considered in relation to this activity. Sample questions are given below.

1. A student reports that their confidence interval for the proportion of the US that is within a mile of a road ranges from 0.38 to 0.84. They state that they are 95% confident that if they took another sample, the sample proportion would be in this interval. What is wrong with their reasoning?
  
2. A student finds that out of the 17 points that they sampled, only 11 of them were within a mile of a road. Calculate and interpret the sample proportion and 95% confidence interval for the population parameter.
  
3. A student finds that out of the 17 points that they sampled, only 11 of them were within a mile of a road. Is it appropriate for this student to use the large sample approximation? Why or why not?

4. A student reports that their 95% confidence interval for the proportion of the US that is within a mile of a road ranges from 0.38 to 0.86. They state that it's unlikely that the true proportion is equal to 0.9. Is their reasoning correct?

5. One group finds that out of the 17 points they sampled, 11 of them were within a mile of a road. A student from a different group reports that she also found 11 points within a mile of a road, but sampled 19 points. She concludes that her standard error must be greater. Is this student correct?

6. Is it possible for a group to satisfy the Success/Failure condition? Why or why not?

## Answers

1. The correct interpretation for a confidence interval is that if we go through the process of collecting a sample and calculating its confidence interval 100 more times, about 95 of the confidence intervals will contain the true proportion of the US that is within a mile of a road. We cannot place the 95% confidence on a single new sample, but we can use it to give a sense of what we would expect to see out of many new samples.
2. Though the sample size is small, and the normal approximation may not hold, we will proceed with it (and compare to the results using the exact binomial).

Sample Proportion =  $11/17 = 0.647$

Confidence Interval:  $0.647 \pm 1.96 * \sqrt{\frac{0.647 * (1 - 0.647)}{17}}$

$[0.420, 0.874]$

We are 95% confident that the true proportion of the US that is within one mile of a road lies between 42% and 87.4%.

```
> binom.test(11, 17)
```

Exact binomial test

data: x and n

number of successes = 11, number of trials = 17

95 percent confidence interval:

0.3832837 0.8579025

sample estimates:

probability of success

0.6470588

3. No, check the Success/Failure Condition:

$$17 * (11/17) = 11$$

$$17 * (1 - 11/17) = 6 < 10$$

Therefore the condition is not met.

4. Yes, we are 95% confident that the true proportion of the US that is within a mile of a road lies within 0.38 and 0.86. Therefore, 0.9 is a proportion that we do not consider likely to be the true proportion as it lies outside of this interval.

5. No, because the second student has a larger sample size, her standard error will actually be smaller. This is because the standard error formula is:  $\sqrt{\frac{\hat{p}*(1-\hat{p})}{n}}$  and when  $n$  increases, the overall standard error decreases.

6. If the proportion of the US that is within one mile of a road is 50% and each group has 20 samples then the Success/Failure condition is just barely met.

$$20*(0.5) = 10$$

$$20*(1-.5)=10$$

However, we have a good sense that the proportion of the US that is within one mile of a road is greater than 50%, so it is unlikely that a single group can satisfy this condition. As the sample proportion  $\hat{p}$  increases,  $1 - \hat{p}$  decreases and the failure part of the condition will not be met.

### Possible Extensions

A number of extensions of this activity may be considered. The `map.where()` function in the "maps" package can be used to determine whether a sampled point is within the continental US: this could be used to automate the sampling of 20 locations within the desired region. For example, the following code shows that we are in Manhattan for location 2:

```
require(maps); require(mosaic)
map.where(database='usa', -73.9863, 40.7566)
[1] "manhattan"
```

We have not done this because we believe that it is pedagogically appropriate for students to learn that sampling  $n$  locations may not yield  $n$  observations of the quantity they wish to study. This is analogous to a clinical trial, where some subjects who are screened decline to participate.

Other countries may be sampled (to allow a comparison of the density of the road network).

R Markdown ([http://www.rstudio.com/ide/docs/authoring/using\\_markdown](http://www.rstudio.com/ide/docs/authoring/using_markdown)) enables easy authoring of reproducible web reports from R and RStudio (Baumer et al, 2014). This technology could be used to facilitate the commands needed by students and instructors. We have provided sample Markdown files along with the other sample files in the RESOURCES directory.

## Acknowledgements

We wish to thank Daniel Kaplan of Macalester College for the original idea for this project (which builds on Gelman and Nolan's beach ball example from "Teaching statistics: a bag of tricks.") and Randall Pruim for the mosaic package and the implementation of the "rgeo()" and "googleMap()" functions. We also acknowledge support from Project MOSAIC (NSF grant 0920350) and the Smith College Provost/Dean of the Faculty Office.

## References

1. Agresti, Alan, and Brent A. Coull. "Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions." *American Statistician*. Vol. 52 No. 2. (1998): 119-126. Web. 17 Nov. 2013. <<http://www.jstor.org/stable/2685469> >.
2. Baumer, Ben, Çetinkaya-Rundel, Mine, Bray, Andrew, Loi, Linda, and Horton, Nicholas J. "R Markdown: Integrating A Reproducible Analysis Tool into Introductory Statistics.", 2014, Technology Innovations in Statistics Education, <http://escholarship.org/uc/item/90b2f5xh>.
3. Clopper, C. and Pearson, S. The use of confidence or fiducial limits illustrated in the case of the Binomial. *Biometrika* 26: 404-413, 1934.
4. DeVeaux, Velleman and Bock. *Intro Statistics* (4th edition), 2013, pages 436-437.
5. Franklin et al. "Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report", ASA, 2007. <http://www.amstat.org/education/gaise/>, accessed November 17, 2013.
6. Gelman, Andrew, and Deborah Nolan. *Teaching Statistics: A Bag of Tricks*. New York, NY: Oxford University Press, Inc, 2002.
7. Murphy, Kim. "Judge upholds roadless protections on US forests," *Los Angeles Times*, March 26, 2013.
8. US Forest Service, "Roadless Area Conservation," US Department of Agriculture, 2013.
9. US Geological Survey, "Distance to Nearest Road in the Conterminous United States," US Department of the Interior, January 2005, [www.fort.usgs.gov/products/publications/21426/21426.pdf](http://www.fort.usgs.gov/products/publications/21426/21426.pdf), accessed November 17, 2013.

## **What Percent of the Continental US is Within 1 Mile of a Road? Activity Sheet (Estimating Population Proportions Using Sampling)**

### **Motivation:**

One of the largest human constructions on Earth is the road network of the United States (USGS, 2005). Roads provide a crucial role as the circulatory system of our economy.

Wilderness is, by definition, roadless. Untouched areas provide important habitats and play a key role in ensuring ecological balance. According to the U.S. Geological Survey (USGS), roads affect the ecology of at least 22% of the land area of the continental United States. Knowing what proportion of the continental US is within 1 mile of a road and mapping these roadless areas can help determine wilderness areas because there is a link between access by road and human presence.

A conservation policy called *roadless area conservation* aims to limit road construction in order to halt negative environmental impacts on designated public lands. The United States Forest Service has formalized the concept of Inventoried Roadless Areas as lands identified by governmental agencies without roads that could be candidates for roadless area conservation. According to the United States Forest Service, inventoried roadless areas comprise only 2% of the land in the continental United States (US Forest Service, 2013).

### **Objective:**

This activity is intended to help reinforce concepts of sampling and confidence intervals while simultaneously understanding how to estimate the proportion of the continental United States that is within 1 mile of a road, an unknown parameter.

You will work with a partner or in a small group to sample random locations within the continental United States and determine whether or not they are within 1 mile of a road. You will calculate a confidence interval with a sample proportion and then compare your results with those of other members of your class and your class as a whole.

### **Citations:**

US Forest Service, "Roadless Area Conservation," US Department of Agriculture, 2013.

US Geological Survey, "Distance to Nearest Road in the Conterminous United States," US Department of the Interior, January 2005.



**Instructions:****Set Up:**

First open R or RStudio.

In the command line run the following commands:

```
install.packages("mosaic")  
require(mosaic)
```

This will give you access to the functions needed to complete this activity.

Next run the command:

```
source("http://www.amherst.edu/~nhorton/roadless-setup.R")
```

This code will generate a list of twenty random latitude and longitude coordinates in and around the contiguous United States, which is stored in the R/RStudio object called "myroadless".

To display the values, type the command:

```
print(myroadless)
```

**Collect Data:**

Due to the irregular perimeter of the United States, some of these locations may lie in an ocean, Mexico, or Canada. For each location you must first determine whether or not the location is within the continental US. If some portion of the circle with radius 1 around a sample point lies within the US, this can be counted as within the US. This point may be discussed or debated further as a class. If a sample point is not within the continental US, there is no need to determine whether or not the location is within 1 mile of a road. A location outside of the continental US will not be included in your sample proportion calculation.

To determine whether each sample is within 1 mile of a road, repeatedly use the "getLocation()" function. This takes a single argument: the location (between 1 and 20) to plot. Run the following one at a time:

```
getLocation(1)  
getLocation(2)  
getLocation(3)  
...  
getLocation(20)
```

When the "getLocation()" command is run, a pop-up window should open that contains a Google map centered on the sample location. Be sure to enable pop-ups so that you can see the Google map. A circle of radius one mile will be drawn around your location. Zoom in or out as necessary to answer the following questions and record them in the table below.

1.) Is your location within the continental United States?

Yes: Put a 1 in the withinContinent column of your table in the row corresponding to your trial.

No: Put a 0 in the withinContinent column of your table in the row corresponding to your trial.

2.) Is your location within one mile of a road?

This only applies if you were within the continental United States.

Yes: Put a 1 in the within1mile column of your table in the row corresponding to your trial.

No: Put a 0 in the within1mile column of your table in the row corresponding to your trial.

3.) In the location column of your table record one of the following:

- the state abbreviation of the state your location is in
- Pacific Ocean
- Atlantic Ocean
- Mexico
- Canada

sample	withinContinent	within1mile	location
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			

**Analyze the Data:**

Now you want to determine your sample proportion of the continental US that is within 1 mile of a road and its corresponding confidence interval.

Count how many ones you have in the withinContinent column of your table. This is your sample size,  $n$ .

Now count how many ones you have in the within1mile column of your table. This is your number of successes,  $X$ .

Your sample proportion,  $\hat{p}$  is:  $X/n$ .

Your standard error, SE is:  $\sqrt{\frac{\hat{p}*(1-\hat{p})}{n}}$ .

Your 95% confidence interval is:  $\hat{p} \pm 1.96* \sqrt{\frac{\hat{p}*(1-\hat{p})}{n}}$ .

One assumption required to use the above normal approximation confidence interval is that the sample size is sufficiently large, which can be checked with the Success/Failure condition. We must expect at least 10 “successes” and at least 10 “failures,” so:  $10 \leq n*(\hat{p})$  and  $10 \leq n*(1-\hat{p})$ . Your group’s results may not meet this assumption, but when you combine your data as a class, the assumption will be met.

Another way to calculate a confidence interval when the sample size is small is to use exact binomial methods. You can do this in R/RStudio by running the following line:

```
binom.test(X, n)
```

using your values for  $X$  and  $n$ .

Interpret your confidence interval in context. What are the other conditions that need to be met in order to use this confidence interval?

**Display the Data:**

Draw the confidence interval that you got from R/RStudio on the board, marking your  $\hat{p}$  as a dot, and the interval as a line. Compare your results with those of other groups in your class. Beside your confidence interval, write your  $X$  and  $n$  values. This way your results can be added to those of your classmates in order to get a confidence interval for your class as a whole. What do you notice is the difference between your confidence interval and that of the entire class? Why do you think this difference occurs?

**Submitting your Data (Optional):**

If your instructor wants to get the details of your locations, you will want to save your data frame to your computer so that you can edit it in a spreadsheet program.

Run the following command:

```
write.csv(myroadless,file="myroadless.csv")
```

This will save your data frame as a CSV file called “myroadless.csv” so that you can edit it in a spreadsheet program and send your results to your instructor.

Run the following command to see where on your computer your data frame is saved:

```
getwd()
```

This will return a pathname telling you where your data frame is located.

Open “myroadless.csv” in a spreadsheet program and fill in the withinContinent, within1mile, and location columns to match your table on this worksheet. Save your work and e-mail it to your instructor.