CONFIDENCE IN SALARIES IN PETROLEUM ENGINEERING

Susan A. Peters
University of Louisville
s.peters@louisville.edu

AnnaMarie Conner
University of Georgia
aconner@uga.edu

Published: October, 2016

Overview of Lesson

This lesson introduces students to bootstrapping methods for making inferences about a population parameter using a randomly selected sample from the population. Students use random samples of salaries for petroleum engineering graduates - —graduates employed in the profession earning the highest mean starting salary in 2014 - —and technology tools to calculate and interpret interval estimates for the mean population starting salary. They also explore the effects of sample size and confidence level on margin of error. Students draw conclusions using both the context of the activity and bootstrapping distributions generated from simulations. Students' explorations conclude with drawing inferences about a population proportion.

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*. The four components are: formulate a question, design and implement a plan to collect data, analyze the data, and interpret results in the context of the original question.
This is a **GAISE Level C** activity.

Common Core State Standards for Mathematical Practice

2. Reason abstractly and quantitatively.
3. Construct viable arguments and critique the reasoning of others.
4. Model with mathematics.
5. Use appropriate tools strategically.

Learning Objectives Alignment with Common Core and NCTM PSSM

Learning Objectives	Common Core State Standards	NCTM Principles and Standards for School Mathematics
Students will describe the effects of sampling variability for samples randomly selected from a population.	7.SP.A.2. Use data from a random sample to draw inferences about a population with an unknown characteristic of	

	interest. Generate multiple samples (or simulated samples) of the same size to gauge the variation in estimates or predictions.	
Students will construct a bootstrapping distribution of means using a random sample.	S-IC.B.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.	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.
Students will use a bootstrapping distribution to find a margin of error for estimating a population mean or proportion.	S-IC.B.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.	Develop and evaluate inferences and predictions that are based on data: • understand how sample statistics reflect the values of population parameters and use sampling distributions as the basis for informal inference.
Students will describe the effects of sample size and confidence level on margin of error.	S-IC.B.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.	
Students will interpret the meaning of 95% confidence.		

Prerequisites

Students should know how to calculate and interpret numerical summary values for one variable data (mean, standard deviation, median, interquartile range) and know how to construct and interpret graphical displays of data including dot plots. Students should have some familiarity with data collection methods such as surveying and important constructs and concepts related to

data collection including random sampling and representative samples. Students who have previously conducted simulations and encountered sampling distributions will benefit most from this lesson.

Time Required

This extensive two-part lesson will require about 100-150 minutes for each part, with Part 1 requiring two 50-minute class periods and Part 2 requiring two to three 50-minute class periods. One class period in each part should be devoted to formulating questions and data collection.

Materials and Preparation Required

- Pencil and paper
- One deck of cards per student group
- Bootstrapping software or Internet access (directions in this lesson will refer to a StatKey applet at http://lock5stat.com/statkey/)
- Calculator or statistics software for computing statistics and graphing data
- Data file
- Post-it notes to record class data
- Large number lines to display dotplots of class data

Confidence in Salaries in Petroleum Engineering
Teacher's Lesson Plan

Part 1: Informal Inference

Describe the Context and Formulate a Question

Ask students to speculate about professions they potentially should consider for the future and why. Have students articulate specific job characteristics that would be important to consider. Students may identify many characteristics, but focus on characteristics that students would want to know as part of determining whether the profession might be a good choice for the future. Ask students to state specific questions about these characteristics that could be answered with data.

After students have a chance to discuss various characteristics, inform them that one job that might be appealing is that of a petroleum engineer. According to a survey conducted by the National Association of Colleges and Employers (NACE), bachelor degree graduates from the class of 2014 who earned the highest average (mean) starting salary of $86,266 were those who majored in petroleum engineering. Focus students on salaries and job availability as two important characteristics for considering the viability of the major and graduates' likelihood of achieving this mean salary.

The series of activities that follows is based on answering the following question related to salary: What is the average starting salary for graduates majoring in petroleum engineering?

Collect Data

This lesson does not involve direct data collection. Instead, students will consider how data were collected by NACE to determine the average starting salary for 2014 petroleum engineering graduates. In particular, students will consider the activity questions for "Setting the Context."

Distribute the "Setting the Context" activity sheet, and ask students to work in pairs or in groups to answer items (1) and (2) on the activity sheet. These items are intended to focus students on the difference between a sample and a population and the importance of using random and representative samples to make inferences about a population. After students have a chance to answer the questions, discuss their responses. If students suggest that NACE should have worked with the population of all 2014 petroleum engineering graduates, ask students to generate reasons why collecting data from the population of all graduates, particularly data about salaries, may not be feasible or possible. As students discuss methods that NACE might have used to collect salary information from a sample of 2014 petroleum engineering graduates, ask students to justify why the resulting sample from using these methods is or is not likely to be representative of the population. Poll students about what methods they believe would yield representative samples.

An important aspect of any data collection method that should be mentioned is random selection. Although students might suggest different methods to increase representativeness such as stratifying graduates according to the type of institution from which they graduated, there still may be lurking variables that interfere with selecting a representative sample. Random selection is designed to control the effects of unidentified factors by ensuring equal probabilities for selecting units exhibiting these factors (or not).

Ask students to consider NACE's actual data collection methods by first describing the methods to them. Some of the methods are specified on the activity sheet; other information can be obtained from the NACE (2015b) publication, *First destinations for the college class of 2014.* Ask students to consider whether the methods would produce a random and/or representative sample of 2014 graduates and why or why not (#3 from activity sheet). Positive aspects of the data collection methods include diversity in the institutions responding to the survey, the likelihood of truthful responses through anonymous reporting through institution contacts, and the presumably large size of the sample. Negative aspects of the data collection methods include an absence of randomization and a low response rate to the survey (190 institutions). Although the names of institutions providing data for the survey are provided by NACE in their report, the representativeness of these institutions in comparison with all undergraduate institutions granting degrees in petroleum engineering cannot be determined without extensive effort.

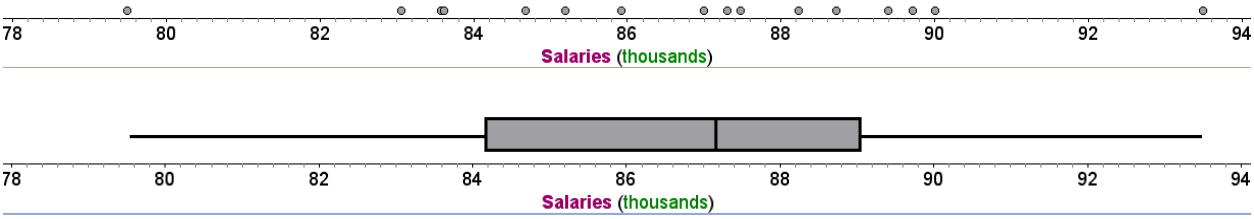
Lead ~~into~~ data analysis by asking students to focus on the meaning of an average salary of $86,266 if we were to assume that the NACE survey produced a representative sample (#4 from activity sheet) and what the average salary reveals about the larger set of data from which it was calculated. Focus on the fact that the data are likely to be variable but that the mean in and of itself tells nothing about the variability. Then, project the following quote, and ask students to comment on its meaning: "You can't fix by analysis what you bungled by design" (Light, Singer, & Willett, 1990, p. *v*.). Remind students that we perform analyses under the assumption that our sample data are representative of the larger population from which they are drawn. Randomization provides the best means for achieving samples representative of their respective populations.

Analyze Data

Ask students to work in pairs or in larger groups to respond to items (1) through (7) on the "Analyzing Data from a Single Sample" activity sheet. Students analyze salaries for a random sample of 16 petroleum engineering graduates and begin thinking about drawing inferences for the larger population of all starting salaries using this sample. The sample size of 16 was chosen specifically to align with the number of face cards used in the simulation, "Using Cards to Bootstrap." Make sure that students have a calculator or software to calculate summary values. Item (1) provides a good opportunity to review some basic statistics with students. If students are

well versed in exploratory data analysis methods and with describing distributions, then little time needs to be devoted to this first item. Note the following summary values, dotplot, and boxplot for these data.

N	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
16	86684.06	847.37	3389.46	79499.00	84152.50	87154.00	89605.50	93499.00



Interpret Results

These data are fairly symmetric, and although the sample size is relatively small, the symmetry and lack of outliers suggest that the mean and standard deviation are appropriate for describing the data. Use a Whip Around strategy to have groups share their descriptions from (1)—randomly select groups to share one observation about the distribution and continue in this manner until all ideas have been shared. As students present their responses, press them to not only report statistics but also to interpret the meaning of the measures. For example, the mean of approximately $86,684 means that if every one of the 16 engineers earned the same salary, they would each earn a salary of $86,684. This is not the case, however, as the approximate average deviation from the mean is $3,389. The middle 50% of salaries fall in the interval between $84,152.50 and $89,605.60. The person earning the least in this sample earns $79,499, which is $14,000 less than the person earning the maximum of $93,499.

As students continue to share their responses to items (2) through (7), focus students on the idea that sample characteristics rarely, if ever, are equivalent to the population characteristics whether the population is salaries from the NACE survey, salaries from some other population, or units different from salaries. Therefore, a sample mean is not likely to equal a population mean; however, without additional information about a population, a sample mean provides a reasonable estimate for the population mean. Introduce the idea of ***sampling variability***—that samples and their characteristics such as shape, measures of center, and measure of variation are likely to vary from sample to sample in repeated sampling—to suggest that this sample of size 16 could have been selected from the same population as the NACE sample. If students consider the variability in means to be too great for the sample of size 16 and the NACE sample to have been selected from the same population, ask students to speculate about what difference in means

would suggest samples selected from the sample population. Point out that inference techniques present criteria for making these types of decisions.

Items (6) and (7) set up the idea of using samples to make inferences about populations. Point out to students that because we typically don't expect sample means to equal population means, we typically estimate a population characteristic using an interval of values—an ***interval estimate***. Inform students that they will explore one method for finding interval estimates: bootstrapping.

Part 2: Bootstrapping

Describe the Context and Formulate a Question

Before introducing bootstrapping to students, revisit the statistical question that was answered by the previous series of activities. In particular, ask students to identify the question, namely: What is the average starting salary for graduates majoring in petroleum engineering? Remind students that in the first series of activities, they answered the question using a single value. In the next series of activities, they will answer the question using an interval of values.

Ask students to again consider other questions they might wish to answer by using data about petroleum engineering graduates. Focus students on job availability as an important characteristic for considering the likelihood of graduates achieving this mean salary. In addition to constructing interval estimates for a population mean, students will construct interval estimates in response to the following question: What proportion of petroleum engineering graduates is unemployed?

Collect Data

This lesson does not involve direct data collection for the initial sample of 16 salaries. However, students will use sampling with replacement to select additional samples towards estimating the average starting salary for graduates majoring in petroleum engineering.

Return to the question from Part 1 that asked students to consider how close their estimate of the mean starting salary for all 2014 petroleum engineering graduates might be to the population mean. Ask students to assume that they only had their sample and the statistics they calculated from the sample to draw conjectures about an interval of values that would be reasonable for the population mean. Ask what information students considered when deciding upon this interval. Also ask students how confident they are that the population mean would be in this interval. Then ask students what additional information they might want to be more confident in constructing an interval to estimate the population mean. Depending upon their previous

experiences, students might suggest obtaining a larger sample or additional samples. Point out to students that the only data they have available to them is the data from this single sample. Revisit the idea of representativeness to have students consider what the population might be if the sample truly were representative of the population. Lead into the idea of bootstrapping by asking students to consider how they might use this single sample to obtain additional samples.

Bootstrapping and Sampling with Replacement

Give students time to read the boxed information that appears on the "Bootstrapping and Sampling with Replacement" page and to respond to the question on the page. Engage students in a think-pair-share to think about, discuss, and share methods for sampling with replacement. Students may suggest strategies such as creating slips of paper for each salary and selecting slips (with replacement) from a hat. If students previously worked with random number tables, they may suggest assigning numbers to each possible outcome and using a random number table to simulate sampling with replacement. For each strategy presented, ask students to be explicit in describing how each of the 16 salaries is represented, how the process incorporates randomization (so that each of the 16 salaries has the same probability of being selected), and how the process incorporates the idea of replacement so that each of the 16 values can be selected for each of the 16 selections.

Note that the bootstrapping process and using sample data as if it were population data may not be intuitive for students. Remind students that ideally we would work with the population directly; because we realistically can only work from the sample, we try to approximate characteristics of the population as closely as possible by using sample data. In this way, students can consider the population to be many copies of the sample. Rather than making copies of the sample, we use sampling with replacement to repeatedly select samples from our sample. In the case of salaries for 2014 petroleum engineering graduates, we assume that each starting salary from the sample represents many similar starting salaries from the population of all petroleum engineering graduates.

From this point forward, we refer to the 16 salaries as a sample of salaries. Our use of this terminology is an abbreviated way of saying that the salaries were reported by a sample of 2014 petroleum engineering graduates from the population of all 2014 petroleum engineering graduates. The sampling unit is the graduate, but the observational unit is the salary. Students may pick up on this slight change in wording.

Using Cards to Bootstrap

Students use a deck of cards to simulate sampling with replacement from the given sample of 16 salaries [items (1) through (5) for "Using Cards to Bootstrap"]. Research suggests that performing simulations by hand before using technology can aid students in understanding the conceptual ideas that underlie statistical inference (Pfannkuch, Forbes, Harraway, Budgett, & Wild, 2013). Students will combine their results with those from the rest of the class to

conjecture appropriate interval estimates for the population mean starting salary [items (6) through (10) for "Using Cards to Bootstrap"]. Prior to beginning this activity, display a large number line such as the number line displayed in item (6) in the classroom. Students should record their sample means on post-it notes and position the post-it notes as dots above the number line to create a dotplot of simulation results from the class.

As part of the simulation process, students compare one or more simulated bootstrap sample salaries with the original sample salaries to reinforce the notion of sampling variability [item (2)]. While they are working, encourage students to consider the shape, center, and variation of the bootstrap samples in comparison with the original sample. Students should observe differences in these characteristics, but ask them to focus on the variability in the samples in comparison with the variability in characteristics. Students create dotplots from the bootstrap sample means to begin creating a bootstrap distribution; asking students to compare the variability of the samples with the variability of the bootstrap distribution can help students to see the reduced variation in a distribution of means.

After students complete the activity, focus discussion on items (8) and (9). If students do not express greater confidence for suggesting an interval estimate from the class display, question students about how the size of a sample affects their confidence for describing distribution characteristics. Just as larger samples instill greater confidence for drawing inferences about populations, larger distributions of statistics instill greater confidence for drawing inferences about parameters. The distribution of bootstrap sample means from (6) is a distribution of a sample of sample means. Students should have greater confidence in estimating an interval estimate for the population mean from the dotplot displaying sample means from the class, which should then motivate additional simulation.

Analyze Data

Bootstrapping for Confidence

Students will need Internet access and computing technology in the form of a laptop or tablet to access the StatKey applets (http://lock5stat.com/statkey) to create a bootstrap distribution for 1000 sample means. Students should work with a partner on items (1) through (5) from "Bootstrapping for Confidence" and use the applet to find a 95% confidence interval for the population mean starting salary of petroleum engineering graduates.

When all students have recorded 95% confidence intervals for the population mean salary, ask students to share their intervals with the class. Then ask students to respond to the following questions.

1. How many of the class intervals capture the mean of $86,266 reported by NACE?

2. Would these intervals suggest that $86,266 could be the population mean? Why or why not?
3. How are the class intervals similar?
4. How do the class intervals differ?

These questions are intended to lead students into considering the meaning of 95% confidence. Most, if not all, of the class intervals will capture the mean reported by NACE to suggest that the population mean could be $86,266. The intervals likely will have different centers (and endpoints) but similar widths. The center of each interval likely will be close to the mean of the original sample, $86,684. A common misinterpretation of confidence intervals stems from a desire to interpret 95% as the percentage of sample data captured in the interval (West & Ogden, 1998). Focus students on the original sample used to conduct the simulation to generate a confidence interval, and ask students whether they see any relationship between the data and the interval. In particular, ask students to compare individual salaries with the confidence interval. The important point to emphasize is that the confidence interval tells us nothing about individual data values—the interval only provides an estimate for a population parameter, which in this case is a mean. Analysis of these intervals reveals that in the long run, approximately 95% of the intervals capture the population mean. Associating 95% with multiple intervals can prevent students from falling into common traps for interpreting confidence intervals. Students should focus on 95% as the percentage of intervals that capture the parameter and associate this percentage with level of confidence. Because confidence intervals are constructed from a sample, students tend to want to interpret the intervals in terms of a sample characteristic rather than the population parameter for which the interval is intended to provide an interval of plausible estimate values (West & Ogden, 1998).

Gaining or Losing Confidence

Students next explore the effects of confidence level on the width of confidence intervals by responding to items (1) through (5) on "Gaining or Losing Confidence." Although not necessary for meeting the standards addressed by this lesson, students' understanding of confidence can be enhanced by focusing on the intervals and repeatedly considering the intervals to be estimates for the population mean. Focus discussion on item (4). The larger number of population estimates resulting from a wider interval of estimated values should translate into more confidence for capturing the population mean, a notion that tends to be counterintuitive for students.

Considering the Effects of Sample Size

Students also should consider the effects of sample size on the width of confidence intervals. Students will work with samples of size 64 and 256, which were chosen to better enable students to discover the inverse square root relationship between sample size and margin of error. Students should work in pairs to develop responses for items (1) through (7) on "Considering the Effects of Sample Size". Time can be saved by providing students with a file that contains the

data for 64 starting salaries so that students can copy and paste the data into the applet without needing to key in all 64 values. Create a text file containing a heading of "Salary" and each of the 64 salaries on separate lines. Provide students with the file, and make sure that they know the name of the file.

Note the following summary values for this sample. If students need additional time describing distributions, spend time graphing these data and describing the distribution, paying particular attention to shape, center, and variation.

N	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
64	85784.56	466.23	3729.80	75773.00	83443.00	85446.50	88397.00	95127.00

After students respond to item (7), focus discussion on items (2), (5), and (7). For (2), pay particular attention to students' interpretation of the 95% confidence interval to potentially address common pitfalls such as claims that the confidence interval provides a 95% probability for capturing the population mean. A single interval is not associated with probability for capturing the mean—the interval either does or does not capture the population mean. In the long run, 95% of confidence intervals constructed in a similar manner would capture the population mean; 95% confidence does not tell us anything about whether a single confidence interval captures the population parameter but rather refers to the long-term behavior of the process.

For (5), focus on similarities such as the interval centers being close in value to the sample mean. Also focus on differences such as the widths of the intervals. The width of the interval for the sample of size 64 is smaller than the width of the interval for the sample of size 16. Focus on intervals and their widths leads naturally into discussing margin of error (7), which is also known as the half-width of the interval.

Students should further consider the effects of sample size on the margin of error by responding to items (8) through (12). When discussing these items, focus on the inverse square root relationship to suggest that halving the margin of error would occur again if we took a random sample of 256 starting salaries. Also reinforce the relationship between the sample mean and the mean of the bootstrap distribution, proper interpretation of the 95% confidence interval, the relationship between the population mean and the confidence interval, and how sample size affects the margin of error.

Interpret Results

Bootstrapping for Confidence in Employment

Throughout students' analyses, students interpreted results in the context of the data with guidance from the instructor. To consider using bootstrapping procedures to estimate population

proportions, ask students to work independently on items (1) through (5) of "Bootstrapping for Confidence in Employment." Students find confidence intervals for a population proportion using StatKey and interpret the intervals within the context of the data. Before starting, students may need some help connecting the bootstrapping process for estimating a population mean with the bootstrapping process for estimating a population proportion. Students may need to consider a much smaller sample and resample by hand from that smaller sample before using StatKey to simulate resampling from the larger sample.

Students should work in groups of four to interpret the confidence interval in terms of an interval estimate for the population proportion of petroleum engineering graduates currently seeking employment. They also consider ways in which they could find more precise estimates for the population proportion.

Students should share their results in groups in such a way as to highlight the variations in their results. Each student is assigned a number between one and four. All students assigned number one meet to discuss their group's responses to items (1) through (5). Similarly, students form groups from those assigned numbers two, three, and four to discuss results. Inform students that they should come to agreement on their responses to (3), (4) and (5) and be prepared to share their responses or questions they have about these items with the rest of the class. As groups discuss their results, circulate among the groups to ensure proper interpretations of intervals. After groups have a chance to discuss their results and come to consensus, ask students to return to their original groups to discuss any potential differences that arose from meeting with members of other groups. Bring the class back together as a whole, and discuss items (1) through (5) as needed.

Now that students have analyzed considerable data, return to the NACE survey and the data collection methods used by NACE. Tell students that the data they explored essentially were random samples from the NACE survey data. Ask students: do you believe that the average starting salary for petroleum engineers and the proportion of petroleum engineering majors currently seeking employment put forth by NACE generalizes to *all* 2014 petroleum engineering graduates? Why or why not? If students do not believe that results from NACE data generalize to the larger population, ask students for implications from their analyses.

Suggested Assessment

Ask students to complete items (1) and (2) from "Try This on your Own." Students should provide complete interpretations of the processes used to find the confidence interval, the confidence interval, and the margin error in their explanations.

Possible Differentiation

The lesson in general is targeted for students at GAISE Level C; however, lesson activities associated with the Part 1 could be implemented with students at Level B. These students may need some additional guidance for representing and describing sample data when "Analyzing Data from a Single Sample" such as being told which representations and statistics to use.

Students at Level B may need greater differentiation for activities associated with Part 2. Specifically, they may need to discuss the concepts introduced in "Bootstrapping and Sampling with Replacement" in conjunction with completing the first step of "Using Cards to Bootstrap." After using cards to sample with replacement, students may be able to consider additional processes that could be used to sample with replacement. Similarly, after completing the second step of "Using Cards to Bootstrap," students may observe that different samples yield different population estimates to suggest why using an interval estimate might be better than using point estimates for a population characteristic. Students at Level B also will need to spend some time comparing the sample distribution with the distribution of means that emerges in "Using Cards to Bootstrap." Similarly, they should make multiple comparisons between the sample distribution and the distribution of means in "Bootstrapping for Confidence." Rather than immediately generating 1000 samples using the software, students should generate many samples and examine the emerging distribution of means to compare characteristics of the sample distribution with characteristics of the distribution of sample means. Students at Level B might skip "Gaining or Losing Confidence" and work more slowly through parts (1) through (5) of "Considering the Effects of Sample Size" by again comparing the sample distribution with the emerging distribution of sample means. Differentiation needed for "Bootstrapping for Confidence in employment" and "Try this on your Own" similarly should focus more on making the situation as concrete as possible and slowly generating the distribution of statistics and thus focus more on the beginning steps of the activities than on the later steps.

Possible Extension

Statistics educators argue that statistical activities should involve use of real data. Others go further in suggesting that activities "should use real data that matters" (Tintle et al., 2016), data that can be used to make decisions. Although students may not make decisions about their future career by examining statistics related to petroleum engineering, they can use the strategies and techniques presented in the context of petroleum engineering graduates to make decisions about their futures. One extension that can benefit students would be to have students conduct research for an occupation of interest to them. Although NACE does not provide free access to their data on recent college graduates, salary and unemployment data are freely available from the Bureau of Labor Statistics through their Current Population Survey (http://www.bls.gov/cps/data.htm). Students can use data and statistics reported by the Bureau of Labor Statistics for occupations of interest to find interval estimates for parameters.

References for Confidence in Petroleum Engineering Activities

Cleophas, T. J., Zwinderman, A. H., Cleophas, T. F., & Cleophas, E. P. (2009). *Statistics Applied to Clinical Trials*. Dordrecht, The Netherlands: Springer.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for Assessment and Instruction in Statistics Education* (GAISE) *Report*. Alexandra, VA: American Statistical Association. Retrieved from http://www.amstat.org/education/gaise/

Kemp, J. (2014, July 17). Peak petroleum engineer? Or time to still join the boom? *Reuters.* Retrieved from http://www.reuters.com/article/2014/07/17/us-usa-engineering-oilandgas-employment-idUSKBN0FM28520140717

Light, R. J., Singer, J. D., & Willett, J. B. (1990). *By Design: Planning Research on Higher Education*. Cambridge, MA: Harvard University Press.

Lock, R. H., Lock, P. F., Morgan, K. L., Lock, E. F., & Lock, D. F. (2013). *Statistics: Unlocking the Power of Data*. Hoboken, NJ: Wiley.

National Association of Colleges and Employers. (2015a). *Spring 2015 Salary Survey Executive Summary*. Bethlehem, PA: Author.

National Association of Colleges and Employers. (2015b). *First Destinations for the College Class of 2014*. Bethlehem, PA: Author. https://www.naceweb.org/uploadedFiles/Pages/surveys/first-destination/nace-first-destination-survey-preliminary-report-022015.pdf

National Association of Colleges and Employers. (2015c). *Class of 2014 Bachelor Degree Results*. Bethlehem, PA: Author.

National Council of Teachers of Mathematics (NCTM). (2000). *Principles and Standards for School Mathematics*. Reston, VA: Author.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for Mathematics*. Washington, DC: Authors. Retrieved from http://www.corestandards.org/wp-content/uploads/Math_Standards.pdf

Payscale Human Capital. (2015). *Petroleum Engineer Salary (United States).* Retrieved from http://www.payscale.com/research/US/Job=Petroleum_Engineer/Salary

Pfannkuch, M., Forbes, S., Harraway, J., Budgett, S., & Wild, C. (2013). *"Bootstrapping" Students' Understanding of Statistical Inference.* Auckland, NZ: Teaching & Learning Research Initiative. Retrieved from http://www.tlri.org.nz/sites/default/files/projects/9295_summary%20report.pdf

Tintle, N., Chance, B. L., Cobb, G. W., Rossman, A. J., Roy, S., Swanson, T., & VanderStoep, J. (2016). *Introduction to Statistical Investigations*. Hoboken, NJ: Wiley.

West, R. W., & Ogden, R. T. (1998). Interactive Demonstrations for Statistics Education on the World Wide Web. *Journal of Statistics Education, 6*(3). Retrieved from http://www.amstat.org/publications/jse/v6n3/west.html

Further Reading About the Topic

Tintle, N., Chance, B. L., Cobb, G. W., Rossman, A. J., Roy, S., Swanson, T., & VanderStoep, J. (2016). *Introduction to statistical investigations*. Hoboken, NJ: Wiley.

Wild, C. (2011, November 22). Bootstrapping and randomization: Seeing all the moving parts [Webinar]. In *CAUSEweb Activity Webinar Series*. Retrieved from https://www.causeweb.org/webinar/activity/2011-11/

Zieffler, A., & Catalysts for Change. (2015). *Statistical Thinking: A simulation approach to uncertainty* (3rd edition). Minneapolis, MN: Catalyst Press. Downloadable from https://github.com/zief0002/Statistical-Thinking

Confidence in Petroleum Engineering
Student Handouts

Setting the Context


http://www.forbes.com/pictures/efkk45eghj/1-petroleum-engineering/

For the class of 2014, bachelor's degree graduates earning the highest average (mean) starting salary of $86,266 were those who majored in petroleum engineering (National Association of Colleges and Employers [NACE], 2015a). Petroleum engineers often work for oil companies and oversee retrieval and production methods for oil and natural gas (Payscale, 2015). The demand for petroleum engineers tends to rise and fall with oil prices. As oil prices increase, consumer demands for cheaper production increase; as oil prices decrease, so do demands for innovation. Challenges such as increased environmental regulations, however, suggest that demands for petroleum engineers will not taper off soon (Kemp, 2014).


http://www.pete.lsu.edu/research/pertt/photos

1. The National Association of Colleges and Employers (NACE) reported a mean starting salary of $86,266 for bachelor degree graduates in petroleum engineering. How might the Association have collected data to determine the figure of $86,266?

2. If all 2014 petroleum engineering graduates were surveyed, would their mean starting salary be $86,266? Why or why not?

NACE surveyed its member higher education institutions, and the member institutions collected survey information from their graduates to form the sample from which information about petroleum engineers was drawn (NACE, 2015b). NACE membership consists of nearly 2000 colleges and universities in the United States representing diverse geographic areas, public and private institutions, urban and rural institutions, and small to large organizations. A total of 190 schools and career centers responded to the survey.

3. How representative is the sample of petroleum engineers surveyed by NACE in relation to the population of all 2014 graduates with degrees in petroleum engineering? On what are you basing this belief?

4. Assume that the NACE sample is a random, or at least representative, sample of 2014 petroleum engineering graduates. Would each graduate within the sample earn $86,266 annually? Why or why not?

Analyzing Data from a Single Sample

1. Suppose a random sample of 16 petroleum engineering majors who graduated in 2014 reported the following salaries: $93499, $90008, $89719, $89401, $88730, $88238, $87475 $87306, $87002, $85923, $85193, $84682, $83623, $83584, $83063, and $79499. Represent and describe these sample data.

2. Is the mean salary from this sample equal to the mean salary reported by NACE? Should it be? Why or why not?

3. If the actual mean starting salary for petroleum engineers equals the NACE estimate of $86,266, could the salaries from #1 have been reported from a sample of graduates from the population of all petroleum engineering graduates? Why or why not?

4. Based on your examination of salaries, could the given sample of petroleum engineering graduates have been drawn from the same population of engineers as the sample surveyed by NACE? Why or why not?

5. Estimate the mean starting salary for all 2014 petroleum-engineering graduates. Explain your answer?

6. Will this estimate for the mean starting salary of the population be equal to the population mean? Why or why not?

7. Would you expect the estimate to be reasonably close to the population mean? How close?

Bootstrapping and Sampling with Replacement

In reality, surveying an entire population typically cannot be done. In the case of surveying graduates to determine their starting salaries, privacy laws would prohibit colleges and universities from supplying researchers with graduates' contact information. Even if populations can be surveyed, the costs associated with doing so often would be prohibitive. We get our best guesses about characteristics of a population from using a sample randomly selected from the population.

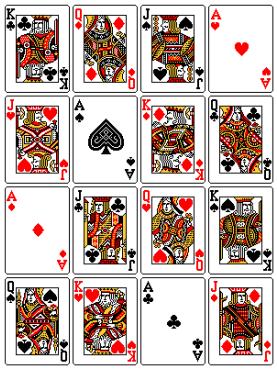
Because we do not anticipate that the sample will match the population exactly, we estimate population characteristics using intervals of values (***interval estimates***) rather than individual values (***point estimates***). One method for constructing interval estimates is known as the ***bootstrapping*** method. The method's name comes from the saying to "pull yourself up by your bootstraps" (Cleophas, Zwinderman, Cleophas, & Cleophas, 2009), which refers to using one's own efforts to get out of a difficult or impossible situation—to make the seemingly impossible become possible. In the case of statistics, the bootstrap method allows us to make estimates for the population through brute force—no formulas necessary!

https://omarsbrain.wordpress.com/2010/01/22/bootstrapping-and-artificial-intelligence/

We are interested in estimating the actual mean starting salary for 2014 petroleum engineering graduates. We could select additional samples of engineers and calculate their mean starting salaries to form an interval estimate for the population mean. Because sampling from the population can be expensive, however, we instead use our best estimate for the population—the sample—and use it as if it were the population. We select samples, called bootstrap samples, using the data from our sample, a process called ***resampling***. Because there are a finite number of values in our sample, we use ***sampling with replacement***, meaning that after being selected, each salary is recorded and returned to the collection before the next salary is selected.

Describe a process for sampling with replacement that could be used to randomly select 16 salaries from the 16 salaries given in # 1 from "Analyzing Data from a Single Sample."

Using Cards to Bootstrap



http://www.numericana.com/answer/cards.htm

To find a reasonable interval estimate for a population mean, we need to select many samples and calculate their sample means. (In reality, we would want to select all possible resamples to know all possible means that could result from samples of the population, but doing so often is impractical. Instead, we work with a large number of resamples.) We simulate the process for the sake of efficiency.

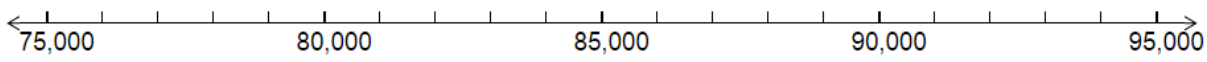
We will use 16 cards from a deck of cards to represent specific salaries in order to simulate sampling with replacement from our sample of 16 salaries. In particular, we will use the aces and face cards of the four card suits to represent each of the salaries as shown on the next page. To begin, remove the aces and face cards from your deck of cards.

1. Simulate the selection of a sample of size 16 using resampling.
   a. Shuffle the 16 aces and face cards, and randomly select one of the cards.
   b. Record a tally mark for this card in the appropriate box for Sample 1 on the next page.
   c. Replace the card.
   d. Repeat the selection and recording process (a-c) 15 more times until you have a total of 16 tally marks.
   e. Calculate the mean for the 16 salaries selected, and record the value in the table.

2. Compare and contrast this bootstrap sample with the original sample of size 16. Focus on the distribution of values and on the mean.

3. Repeat the resampling process (#1) three more times, recording your results in the tables on the next page.

4. Examine the four means that you calculated for your four bootstrap samples by first plotting the means on a dotplot. Use these means to suggest an interval estimate for the mean starting salary of the population of petroleum engineering graduates.



5. Would your estimate change if you had calculated additional means? Why or why not?

Resampling Simulation

Card	Hearts ♥				Clubs ♣				Diamonds ♦				Spades ♠			
	Ace	King	Queen	Jack	Ace	King	Queen	Jack	Ace	King	Queen	Jack	Ace	King	Queen	Jack
Salary	$93,499	$90,008	$89,719	$89,401	$88,730	$88,238	$87,475	$87,306	$87,002	$85,923	$85,193	$84,682	$83,623	$83,584	$83,063	$79.499

Sample 1

Card	Hearts ♥				Clubs ♣				Diamonds ♦				Spades ♠			
	Ace	King	Queen	Jack	Ace	King	Queen	Jack	Ace	King	Queen	Jack	Ace	King	Queen	Jack
Salary	$93,499	$90,008	$89,719	$89,401	$88,730	$88,238	$87,475	$87,306	$87,002	$85,923	$85,193	$84,682	$83,623	$83,584	$83,063	$79.499
Tally																
Mean																

Sample 2

Card	Hearts ♥				Clubs ♣				Diamonds ♦				Spades ♠			
	Ace	King	Queen	Jack	Ace	King	Queen	Jack	Ace	King	Queen	Jack	Ace	King	Queen	Jack
Salary	$93,499	$90,008	$89,719	$89,401	$88,730	$88,238	$87,475	$87,306	$87,002	$85,923	$85,193	$84,682	$83,623	$83,584	$83,063	$79.499
Tally																
Mean																

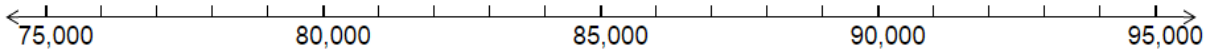
Sample 3

Card	Hearts ♥				Clubs ♣				Diamonds ♦				Spades ♠			
	Ace	King	Queen	Jack	Ace	King	Queen	Jack	Ace	King	Queen	Jack	Ace	King	Queen	Jack
Salary	$93,499	$90,008	$89,719	$89,401	$88,730	$88,238	$87,475	$87,306	$87,002	$85,923	$85,193	$84,682	$83,623	$83,584	$83,063	$79.499
Tally																
Mean																

Sample 4

Card	Hearts ♥				Clubs ♣				Diamonds ♦				Spades ♠			
	Ace	King	Queen	Jack	Ace	King	Queen	Jack	Ace	King	Queen	Jack	Ace	King	Queen	Jack
Salary	$93,499	$90,008	$89,719	$89,401	$88,730	$88,238	$87,475	$87,306	$87,002	$85,923	$85,193	$84,682	$83,623	$83,584	$83,063	$79.499
Tally																
Mean																

6. Record the value of each mean you calculated on a separate post-it note. Use your post-it notes to plot your four means on the class display. Examine the class distribution of means, and record it below.



7. Use the class means to suggest an interval estimate for the mean starting salary of the population of petroleum engineering graduates.

8. Compare and contrast this interval estimate with your estimate from #4.

9. With which estimate are you more confident that you have accurately captured the population mean starting salary for petroleum engineering graduates, and why?

10. How many means were recorded on your dotplot in #6?

Bootstrapping for Confidence

To find an accurate interval estimate for the population mean, we need hundreds of bootstrap sample means—realistically, 1000 or more. Even though the cards can help us to select samples quickly, the card process would be quite tedious and frustrating to use for finding 1000 sample means. We need many more means than we reasonably can gather from using simulations with materials such as cards.

http://aozoraaira.blogspot.com/2014/10/international-moment-of-frustration.html

http://lock5stat.com/statkey/

Instead, we use computing technology to simulate the selection of 1000 or more samples and calculate their means to form a bootstrap distribution of means. A nice collection of applets for resampling, StatKey, is freely available at http://lock5stat.com/statkey/

1. Go to the StatKey website, and under the heading of "Bootstrap Confidence Intervals," select the option of "CI for Single Mean, Median, Standard deviation". To generate an interval estimate for the population mean, often referred to as a confidence interval for the mean, you will first need to enter the 16 salaries from the original sample by following these steps. As a reminder, the salaries are: $93,499, $90,008, $89,719, $89,401, $88,730, $88,238, $87,475 $87,306, $87,002, $85,923, $85,193, $84,682, $83,623, $83,584, $83,063, and $79,499.
   a. Click on the "Edit data" tab at the top of the screen.
   b. Select and delete the data that appear in the "Edit data" window.
   c. On the first line, enter the heading of "Salary."
   d. Enter each of the 16 salaries on a separate line below the heading.
   e. Double-check your entries, and then click "OK."

2. This sample is now displayed in the graph labeled as "Original Sample". Click on the "Generate 1 Sample" tab to select a single bootstrap sample. You should see the sample displayed in the graph labeled as "Bootstrap Sample". The mean of this sample is plotted on the "Bootstrap Dotplot of Mean" graph. As we noted, we would like 1000 or more bootstrap sample means from which to estimate the population mean. Rather than repeat the generation of a single samples 1000 times, we instead will generate 1000 samples by clicking on the "Generate 1000 Samples" tab. You will not see all 1000 samples, but you will see all of the means plotted in the bootstrap distribution. What is the mean of these means?

http://www.awinningpersonality.com/self-improvement/emotional-intelligence/5-daily-exercises-boost-keep-confidence/

The value of the bootstrap distribution mean should be close to or approximately equal to the mean of our original sample. We use the bootstrap distribution to determine our interval estimate for the population mean. Interval estimates are associated with a level of confidence for capturing the value of interest from the population. A common confidence level is 95%, which we would associate with the interval of values for the middle 95% of bootstrap sample means.

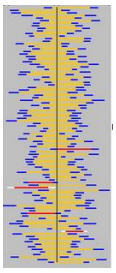
3. To find the 95% confidence interval for the mean starting salary for petroleum engineers, click to select the "two-tail" distribution inside the graph area on the upper left side in StatKey. The endpoints of the interval are displayed on the bootstrap distribution. Record your interval. We would interpret the interval as follows: We are 95% confident that the mean starting salary for all petroleum engineers graduating in 2014 is between <lower endpoint of interval> and <upper endpoint of interval>. Record the interpretation for the interval you found.

4. Did your interval capture the mean starting salary of $86,266 reported by NACE?

5. Does your interval cause you to question the NACE estimate of $86,266 for the population mean? Why or why not?

Gaining or Losing Confidence

In general, level of confidence is based on long-term behavior associated with sampling. Our interval estimate was determined by using data from a sample randomly selected from the population. 95% confidence means that if we were to repeat the sampling process and the process used to calculate confidence intervals, in the long run, we would expect 95% of our intervals to capture the value of the population mean. Because we have no way of knowing the population mean, we will never know whether our 95% confidence interval successfully captured the mean; we only can state our confidence level for capturing the mean.



1. Suppose that you would prefer to have a higher level of confidence such as 99%. How do you think a 99% confidence interval would differ from a 95% confidence interval?

2. Return to StatKey, and click on the value of 0.95 displayed in a blue square in the window for the bootstrap distribution. Enter "0.99" for 99% confidence, and click "Ok." Record the 99% confidence interval. How does the 99% confidence interval differ from the 95% confidence interval?

3. Suppose that you were interested in a lower level of confidence such as 90%. How would a 90% confidence interval differ from the 95% and 99% confidence intervals? Use StatKey to determine whether the 90% confidence interval matches your expectation.

4. You should have found that a 99% confidence interval is wider than a 95% confidence interval found using the same sample. Likewise, a 95% confidence interval is wider than a 90% confidence interval. Describe why this relationship makes sense.

5. A sample size of 16 is relatively small. What effect do you think a larger sample size would have on the 95% confidence interval if the sample characteristics remained the same?

Considering the Effects of Sample Size

http://qbdworks.com/sample-size-question/

Small samples reveal greater variability in shape, center, and variation than larger samples. As a result, smaller samples also reveal greater variability in estimates for population characteristics. In general, bootstrap distributions will center on the value of the original sample under consideration; this is less likely to be the case when bootstrapping from a small sample. For this reason (and others), statisticians prefer working with larger samples. You may have noticed a difference between your sample mean and the center of the bootstrap distribution when you worked with your sample of size 16.

1. Suppose we had selected a random sample of 64 petroleum engineering majors who graduated in 2014 who reported the following salaries.

   89931, 84424, 88501, 84486, 85420, 82618, 91187, 80356, 84020, 85926, 89339, 79041, 82948, 85134, 83727, 84456, 81966, 91112, 80547, 88365, 88232, 88429, 88798, 85410, 87297, 81404, 83795, 83013, 85473. 86270, 83367, 86989, 81648, 85934, 89716, 95127, 84599, 84260, 83519, 92175, 88451, 88036, 79892, 85785, 83022, 91979, 84542, 86263, 79596, 89283, 81663, 86479, 87399, 92901, 84531, 86860, 84135, 83282, 84599, 75773, 91970, 87136, 90598, 87078

   Click on the "Edit Data" tab in StatKey. Enter a heading of "Salary" and then each salary on separate lines OR open a text file that contains the data, and copy and paste the contents into the data window. Find and record the 95% confidence interval using the bootstrap distribution from these data, following the same process used for the sample of size 16.

2. Interpret the 95% confidence interval.

3. Did your interval capture the mean starting salary of $86,266 reported by NACE?

4. Does your interval cause you to question the NACE estimate of $86,266 for the population mean? Why or why not?

5. Compare and contrast the 95% confidence interval from the bootstrap distribution for the sample of size 16 and from the bootstrap distribution for the sample of size 64.

6. Was the bootstrap distribution for the sample size of 64 approximately centered at the mean of the sample ($85,785)?

7. Calculate the half-width of the 95% confidence interval (half of the interval width or half of the difference between the upper and lower endpoints) from the sample of size 16 and the half-width of the 95% confidence interval from the sample of size 64. How are the two half-widths related?

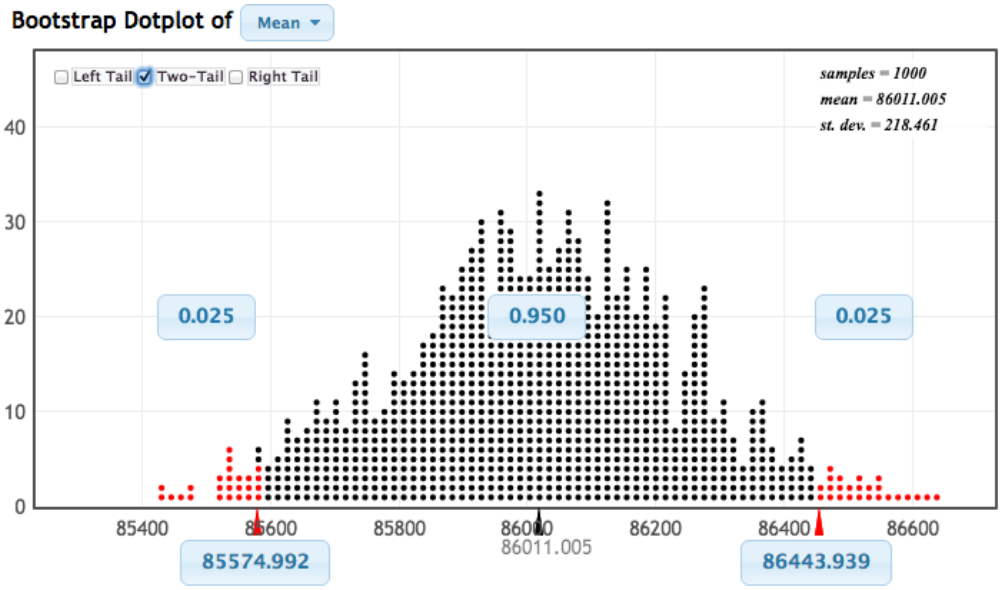
Another name for the half-width of a confidence interval is ***margin of error***. The margin of error gives the maximum expected difference between the population value of interest and the sample estimate for that value. You may have heard this terminology reported on the news in relation to election polls. A report stating that a poll predicts a candidate's support at 42% with a 5.7% margin of error means that 42% of the sample supported the candidate, and the population support is estimated to be between 36.3% and 47.7% with typically 95% confidence. There is in **inverse square root relationship** between the margin of error and sample size. If you wish to halve the width of your confidence interval or the margin of error, you will need to quadruple the size of your sample.

http://www.zazzle.com.au/error+cushions

http://oiecommunity.com/election/election-polls/

8. What sample size would be needed to halve the margin of error for a confidence interval estimate of the population mean starting salary yet again?

9. A precocious student selected a random sample of 256 starting salaries for 2014 petroleum engineering graduates. The student then followed the process outlined above to create a bootstrap distribution and 95% confidence interval and achieved the following results.



What approximate value would you expect for the mean of this student's sample of size 256?

10. Provide an interpretation of the 95% confidence interval for the population mean.

11. Does the interval cause you to question the NACE estimate of $86,266 for the population mean? Why or why not?

12. What is the margin of error? How does this margin of error relate to the margins of error you calculated in #7?

Bootstrapping for Confidence in Employment

A skeptical student makes the following observation: "Big deal! So you convinced me that petroleum engineering graduates have average starting salaries over $85,000. That salary assumes that they actually can find jobs! What about all of those graduates who aren't employed? What good is the salary if I can't get a job to earn that salary?"

http://www.theepochtimes.com/n3/603148-what-is-unhealthy-skepticism/

To gain a sense of the percentage of petroleum engineering graduates that cannot obtain employment, you select a random sample of 277 petroleum engineering graduates and find that 43 of those graduates are seeking employment currently.

We will use the bootstrapping method to find a 95% confidence interval for the population proportion of petroleum engineering graduates who are seeking employment. The beauty of bootstrapping methods is that we follow the same procedures no matter what population characteristic we wish to estimate! In this case, that means that we would start with our sample of size 277; treat this sample as the population; and sample with replacement from this sample. In our simulation, 43 of the 277 sampling units would represent "seeking employment" whereas 234 would represent "employed." We would select a unit 277 times with replacement and record the number of "seeking employment" units selected out of the 277 units. We would repeat this process 1000 or more times and construct a bootstrap distribution of proportions. [The figure of 43 out of 277 graduates currently seeking employment is the number reported by NACE (2015c).]

1. As before, we will use StatKey to perform the simulation efficiently.
   a. From the main StatKey menu, select "CI for Single Proportion" from the "Bootstrap Confidence Intervals" options.
   b. Click to "Edit Data," and enter the count of 43 for graduates seeking employment from the sample of size 277.
   c. Generate 1000 samples.
   d. Find (and record) the 95% confidence interval for the proportion of petroleum engineering graduates seeking employment.

2. Interpret this 95% confidence interval for the population proportion.

3. Does the interval cause you to question whether a majority of graduates are seeking employment currently? Why or why not?

4. What is the margin of error? What does the margin of error tell you?

5. What steps could you take to decrease the margin of error?

Try This on your Own

In reality, even though 43 of the 277 petroleum engineering graduates were seeking employment, not all of the remaining 234 graduates obtained standard petroleum engineering jobs. Some pursued continuing education studies, while others entered into military service. Some were only employed part time. Of the 277 graduates, 68.7% were employed full time as petroleum engineers (NACE, 2015c).

http://woman.thenest.com/chemical-engineer-vs-petroleum-engineer-14124.html

1. Find a 95% confidence interval for the proportion of petroleum engineering graduates who are employed full time. Interpret this interval.

2. How would you explain your findings to the skeptical student who questioned whether graduates could get jobs in petroleum engineering? As part of your response, be sure to provide complete interpretations of the processes used to find the confidence interval and why, the confidence interval, and the margin of error.

Student Handouts
Solutions for the Teacher

Confidence in Salaries in Petroleum Engineering

1. The National Association of Colleges and Employers (NACE) reported a mean starting salary of $86,266 for bachelor degree graduates in petroleum engineering. How might the Association have collected data to determine the figure of $86,266?
   Answers will vary. Students may suggest that NACE should have collected salary information from the population of all 2014 petroleum-engineering graduates. Others may suggest that NACE should collect data from a sample of 2014 petroleum-engineering graduates. Hopefully, some students will focus on methods that produce representative samples of graduates from the larger population of 2014 petroleum-engineering graduates such as by selecting a simple random sample of graduates.

2. If all 2014 petroleum engineering graduates were surveyed, would their mean starting salary be $86,266? Why or why not?
   Because NACE likely did not collect salary data from the population of all 2014 petroleum-engineering graduates, the mean starting salary from the sample of salaries is not likely to match the mean from the population of all salaries; however, the value is likely to be relatively close to the population value. Sample characteristics rarely, if ever, are equivalent to the population characteristics.

3. How representative is the sample of petroleum engineers surveyed by NACE in relation to the population of all 2014 graduates with degrees in petroleum engineering? On what are you basing this belief?
   Answers will vary in that some students may believe that the methods used will yield a representative sample while others will not. Positive aspects of the data collection methods include diversity in the institutions responding to the survey, the likelihood of truthful responses through anonymous reporting through institution contacts, and the presumably large size of the sample. These positive aspects may lead students to believe that the sample is representative of the population of all 2014 petroleum-engineering graduates. Negative aspects of the data collection methods include an absence of randomization and a low response rate to the survey (190 institutions). These negative aspects may lead students to believe that the sample is not representative of the population of all 2014 petroleum-engineering graduates.

4. Assume that the NACE sample is a random, or at least representative, sample of 2014 petroleum engineering graduates. Would each graduate within the sample earn $86,266 annually? Why or why not?

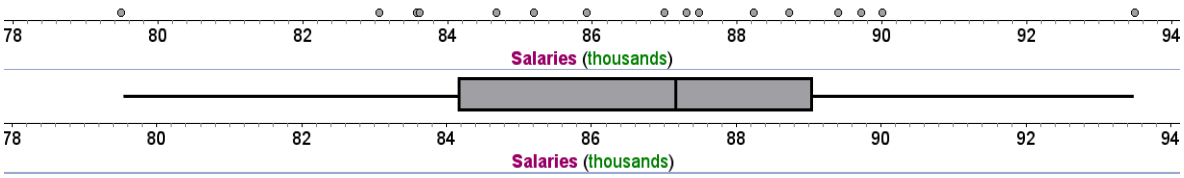
We would not expect each graduate to earn $86,266 annually. Data are likely to be variable, and the mean in and of itself tells us nothing about the variability. The mean only provides a representative value of the salaries—the salary that all sample graduates would earn if each earned the same salary. We would not actually expect graduates to each earn the same salary.

Analyzing Data from a Single Sample

1. Suppose a random sample of 16 petroleum engineering majors who graduated in 2014 reported the following salaries: $93499, $90008, $89719, $89401, $88730, $88238, $87475 $87306, $87002, $85923, $85193, $84682, $83623, $83584, $83063, and $79499. Represent and describe these sample data.

   Note the following summary values, dotplot, and boxplot for these sample data.

N	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
16	86684.06	847.37	3389.46	79499.00	84152.50	87154.00	89605.50	93499.00



   The distribution of sample salaries is fairly symmetric with no outliers. Although the sample size is relatively small, the symmetry and lack of outliers suggest that the mean and standard deviation are appropriate for describing the data. The mean salary is $86,684.06 with a standard deviation of $3389.46. The mean of approximately $86,684 means that if every one of the 16 engineers earned the same salary, they would each earn a salary of $86,684. Data on average deviate approximately $3389.46 from the mean. The lowest salary in the sample is $79,499.00, which is $14,000 less than the highest salary of $93,499. The middle 50% of salaries (8) fall between $84,152.50 and $89,605.50, and the median salary is $87,154.00. Eight salaries are greater than the median salary, and eight are less than the median salary.

2. Is the mean salary from this sample equal to the mean salary reported by NACE? Should it be? Why or why not?

   Sample characteristics rarely, if ever, are equivalent to the population characteristics whether the population is salaries from the NACE survey, salaries from some other population, or units different from salaries. Therefore, a sample mean is not likely to equal a population mean; however, without additional information about a population, a sample mean provides a reasonable estimate for the population mean.

3. If the actual mean starting salary for petroleum engineers equals the NACE estimate of $86,266, could the salaries from #1 have been reported from a sample of graduates from the population of all petroleum engineering graduates? Why or why not?

Sampling variability is to be expected; samples and their characteristics such as shape, measures of center, and measure of variation are likely to vary from sample to sample in repeated sampling and hence vary from characteristics of the population. As a result, sampling variability suggests that this sample of size 16 could have been selected from the population with a mean of $86,266.

4. Based on your examination of salaries, could the given sample of petroleum engineering graduates have been drawn from the same population of engineers as the sample surveyed by NACE? Why or why not?
   Sampling variability is to be expected; samples and their characteristics such as shape, measures of center, and measure of variation are likely to vary from sample to sample in repeated sampling. As a result, sampling variability suggests that this sample of size 16 could have been selected from the same population as the NACE sample.

5. Estimate the mean starting salary for all 2014 petroleum engineering graduates. On what are you basing this estimate?
   Because we typically don't expect sample means to equal population means, we typically estimate a population's characteristics using an interval of values—an ***interval estimate—*** that centers on the value of the sample as our best estimate for the population mean. Thus, students may select an estimate close to the sample mean, an interval of values around the sample mean, or an interval of values close to the sample mean, all based on the value of the sample mean. A sample mean provides a reasonable estimate for the population mean, so a reasonable estimate for the population mean starting salary for 2014 petroleum engineering graduates is $86684.06.

6. Will this estimate for the mean starting salary of the population be equal to the population mean? Why or why not?
   Sampling variability is to be expected; samples and their characteristics such as shape, measures of center, and measure of variation are likely to vary from sample to sample in repeated sampling and hence vary from characteristics of the population. We typically do not expect sample means to equal population means and thus would not expect an estimate based on a sample mean to equal the value of the population mean.

7. Would you expect the estimate to be reasonably close to the population mean? How close?
   A sample mean provides a reasonable estimate for the population mean. As a result, an estimate for a population mean based on a sample mean should be reasonably close to the population mean. With respect to how close, the larger the sample size, the closer the sample estimate should be to the value from the population.

Bootstrapping and Sampling with Replacement

Describe a process for sampling with replacement that could be used to randomly select 16 salaries from the 16 salaries given in # 1 from "Analyzing Data from a Single Sample."

Students may suggest strategies such as creating slips of paper for each salary and selecting slips (with replacement) from a hat. If students previously worked with random number tables, they may suggest assigning numbers to each possible outcome and using a random number table to simulate sampling with replacement. In their responses, students should be explicit in describing how each of the 16 salaries is represented, how the process incorporates randomization (so that each of the 16 salaries has the same probability of being selected), and how the process incorporates the idea of replacement so that each of the 16 values can be selected for each of the 16 selections.

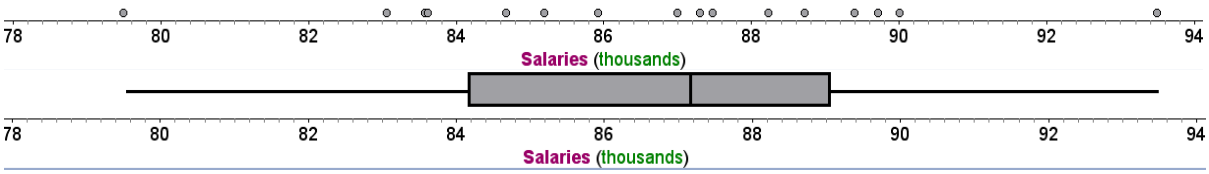
Using Cards to Bootstrap

1. Simulate the selection of a sample of size 16 using resampling.

   a. Calculate the mean for the 16 salaries selected, and record the value in the table.
   Responses will vary. Results from one simulation are recorded in the table for Sample 1.

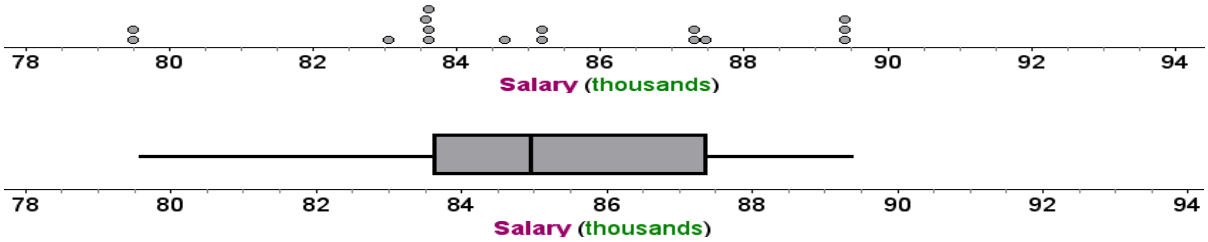
2. Compare and contrast this bootstrap sample with the original sample of size 16. Focus on the distribution of values and on the mean.
   Original Sample:

N	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
16	86684.06	847.37	3389.46	79499.00	84152.50	87154.00	89605.50	93499.00



Bootstrap Sample:

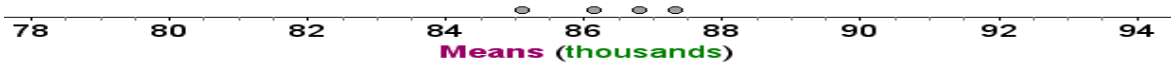
N	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
16	85117.00	782.07	3128.27	79499.00	83603.50	84937.50	87390.50	89401



The bootstrap sample has less variability and tends to include lower salaries than the original sample. The mean, median, lower quartile, upper quartile, and maximum are all lower in value than the corresponding values from the original sample. The mean is approximately $1500 lower whereas the median is approximately $2000 lower. The data in both distributions are fairly symmetric.

3. Repeat the resampling process (#1) three more times, recording your results in the tables on the next page.
   Sample simulation results are recorded in the table.

4. Examine the four means that you calculated for your four bootstrap samples by first plotting the means on a dotplot. Use these means to suggest an interval estimate for the mean starting salary of the population of petroleum engineering graduates.



   The mean of these four means is $86,353.70, so an appropriate point estimate would be $86,353,70 or an interval estimate of values about this mean, between, for example, $85,117 and $87339.30.

5. Would your estimate change if you had calculated additional means? Why or why not?
   The estimate likely would change based on a change in the overall mean of the means. However, the change is not likely to be great.

Resampling Simulation

Card	Hearts ♥				Clubs ♣				Diamonds ♦				Spades ♠			
	Ace	King	Queen	Jack	Ace	King	Queen	Jack	Ace	King	Queen	Jack	Ace	King	Queen	Jack
Salary	$93,499	$90,008	$89,719	$89,401	$88,730	$88,238	$87,475	$87,306	$87,002	$85,923	$85,193	$84,682	$83,623	$83,584	$83,063	$79.499

Sample 1

Card	Hearts ♥				Clubs ♣				Diamonds ♦				Spades ♠			
	Ace	King	Queen	Jack	Ace	King	Queen	Jack	Ace	King	Queen	Jack	Ace	King	Queen	Jack
Salary	$93,499	$90,008	$89,719	$89,401	$88,730	$88,238	$87,475	$87,306	$87,002	$85,923	$85,193	$84,682	$83,623	$83,584	$83,063	$79,499
Tally				III			I	II			II	I	III	I	I	II
Mean	$85,117.00															

Sample 2

Card	Hearts ♥				Clubs ♣				Diamonds ♦				Spades ♠			
	Ace	King	Queen	Jack	Ace	King	Queen	Jack	Ace	King	Queen	Jack	Ace	King	Queen	Jack
Salary	$93,499	$90,008	$89,719	$89,401	$88,730	$88,238	$87,475	$87,306	$87,002	$85,923	$85,193	$84,682	$83,623	$83,584	$83,063	$79,499
Tally	II	I	II	I	I			II		II			I	I	III	
Mean	$87,339.30															

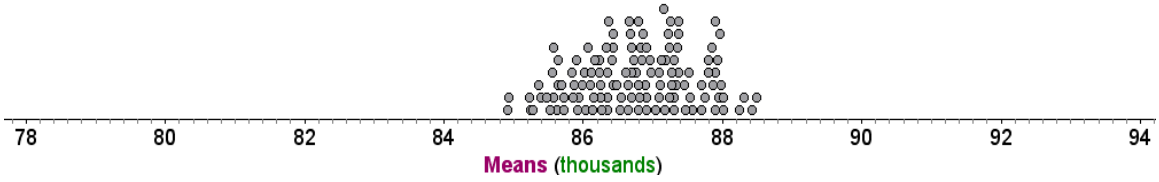
Sample 3

Card	Hearts ♥				Clubs ♣				Diamonds ♦				Spades ♠			
	Ace	King	Queen	Jack	Ace	King	Queen	Jack	Ace	King	Queen	Jack	Ace	King	Queen	Jack
Salary	$93,499	$90,008	$89,719	$89,401	$88,730	$88,238	$87,475	$87,306	$87,002	$85,923	$85,193	$84,682	$83,623	$83,584	$83,063	$79,499
Tally		II	I		I	II		I	II	I	I	I	II			II
Mean	$86,143.30															

Sample 4

Card	Hearts ♥				Clubs ♣				Diamonds ♦				Spades ♠			
	Ace	King	Queen	Jack	Ace	King	Queen	Jack	Ace	King	Queen	Jack	Ace	King	Queen	Jack
Salary	$93,499	$90,008	$89,719	$89,401	$88,730	$88,238	$87,475	$87,306	$87,002	$85,923	$85,193	$84,682	$83,623	$83,584	$83,063	$79,499
Tally	II			I		IIII	I	I				IIII	I		I	I
Mean	$86,815.30															

6. Record the value of each mean you calculated on a separate post-it note. Use your post-it notes to plot your four means on the class display. Examine the class distribution of means, and record it below.
   Sample means from four simulations for 30 students (120 means) are displayed below.



7. Use the class means to suggest an interval estimate for the mean starting salary of the population of petroleum engineering graduates.
   The overall mean of the means is $86,745.70. An appropriate point estimate for the mean starting salary of the population of petroleum engineering graduates would be $86,745.70. Students might provide an interval estimate that encompasses the entire span of means, in this case from $84926.30 to $88,500.80, or a tighter interval based in the assumption that the population mean is more likely to be near the majority of sample means.

8. Compare and contrast this interval estimate with your estimate from #4.
   Theoretically, the more data that we have, the more confidence we should have in the inferences we draw from our data. Therefore, the better interval estimates should come from #7. The point estimate for this series of simulations is slightly higher, and the interval estimate is slightly wider. Most simulations will produce a slightly wider interval estimate if the minimum and maximum sample mean values are used to construct the intervals. Additionally, students are likely to notice that the distribution of means is less variable than the salaries in the sample distribution.

9. With which estimate are you more confident that you have accurately captured the population mean starting salary for petroleum engineering graduates, and why?
   Theoretically, the more data that we have, the more confidence we should have in the inferences we draw from our data. Therefore, the better estimates should come from #7

10. How many means were recorded on your dotplot in #6?
    The dotplot in #6 displays values for 120 means.

Bootstrapping for Confidence

2. This sample is now displayed in the graph labeled as “Original Sample.” Click on the “Generate 1 Sample” tab to select a single bootstrap sample. You should see the sample displayed in the graph labeled as “Bootstrap Sample.” The mean of this sample is plotted on the “Bootstrap Dotplot of Mean” graph. As we noted, we would like 1000 or more bootstrap sample means from which to estimate the population mean. Rather than repeat the generation

of a single samples 1000 times, we instead will generate 1000 samples by clicking on the "Generate 1000 Samples" tab. You will not see all 1000 samples, but you will see all of the means plotted in the bootstrap distribution. What is the mean of these means?
$86,677.21 is the mean for one simulation of 1000 samples.

3. To find the 95% confidence interval for the mean starting salary for petroleum engineers, click to select the "two-tail" distribution inside the graph area on the upper left side in StatKey. The endpoints of the interval are displayed on the bootstrap distribution. Record your interval. We would interpret the interval as follows: We are 95% confident that the mean starting salary for all petroleum engineers graduating in 2014 is between <lower endpoint of interval> and <upper endpoint of interval>. Record the interpretation for the interval you found.
   A 95% confidence interval for the population mean that results from one simulation of 1000 samples is $85,122.47 to $88,196.38. We would interpret the interval to say that we are 95% confident that the mean starting salary for all petroleum engineers graduating in 2014 is between $85,122.47 and $88,196.38.

4. Did your interval capture the mean starting salary of $86,266 reported by NACE?
   Yes, the preceding interval, $85,122.47 to $88,196.38, captures $86,266.

5. Does your interval cause you to question the NACE estimate of $86,266 for the population mean? Why or why not?
   Because $86,266 is included in the interval, we do not have reason to question the NACE estimate of $86,266 for the population mean. The center of the interval, $86,677.21, is within $500 of the NACE estimate.

Gaining or Losing Confidence

1. Suppose that you would prefer to have a higher level of confidence such as 99%. How do you think a 99% confidence interval would differ from a 95% confidence interval?

A 99% confidence interval should be wider than a 95% confidence interval because of the increased confidence for capturing the population mean. A 99% confidence interval will capture more values than a 95% confidence interval, and more values translates to greater confidence that one of the values might be the population mean.

2. Return to StatKey, and click on the value of 0.95 displayed in a blue square in the window for the bootstrap distribution. Enter "0.99" for 99% confidence, and click "Ok." Record the 99% confidence interval. How does the 99% confidence interval differ from the 95% confidence interval?

A 99% confidence interval for the population mean that results from one simulation of 1000 samples and sample means spans from $84571.19 to $88,725.00. The lower bound of this interval is less than the lower bound of the 95% confidence interval, and the upper bound is higher than the upper bound of the 95% confidence interval. Hence, this 99% confidence interval is wider than the 95% confidence interval.

3. Suppose that you were interested in a lower level of confidence such as 90%. How would a 90% confidence interval differ from the 95% and 99% confidence intervals? Use StatKey to determine whether the 90% confidence interval matches your expectation.
   Because a 90% confidence interval is associated with less confidence, we would expect a 90% confidence interval to be narrower than a 95% confidence interval or a 99% confidence interval. A 90% confidence interval for the population mean reported from StatKey is from $85,371.09 to $87,984.250, which is narrower than either of the previous intervals associated with 95% and 99% confidence.

4. You should have found that a 99% confidence interval is wider than a 95% confidence interval found using the same sample. Likewise, a 95% confidence interval is wider than a 90% confidence interval. Describe why this relationship makes sense.

   Greater confidence should be associated with a larger number of possibilities for the population mean, and lesser confidence should be associated with a smaller number of possibilities.

5. A sample size of 16 is relatively small. What effect do you think a larger sample size would have on the 95% confidence interval if the sample characteristics remained the same?
   A larger sample size should provide a better estimate for the population mean and thus should yield a tighter confidence interval.

Considering the Effects of Sample Size

1. Suppose we had selected a random sample of 64 petroleum engineering majors who graduated in 2014 who reported the following salaries.

   89931, 84424, 88501, 84486, 85420, 82618, 91187, 80356, 84020, 85926, 89339, 79041, 82948, 85134, 83727, 84456, 81966, 91112, 80547, 88365, 88232, 88429, 88798, 85410, 87297, 81404, 83795, 83013, 85473. 86270, 83367, 86989, 81648, 85934, 89716, 95127, 84599, 84260, 83519, 92175, 88451, 88036, 79892, 85785, 83022, 91979, 84542, 86263, 79596, 89283, 81663, 86479, 87399, 92901, 84531, 86860, 84135, 83282, 84599, 75773, 91970, 87136, 90598, 87078
   Click on the “Edit Data” tab in StatKey. Enter a heading of “Salary” and then each salary on separate lines OR open a text file that contains the data, and copy and paste the contents into the data window. Find and record the 95% confidence interval using the bootstrap distribution from these data, following the same process used for the sample of size 16.

A 95% confidence interval for the population mean that results from one simulation of 1000 samples of size 64 is from $84,863.90 to $86,630.98 and centered at $85,781.56.

2. Interpret the 95% confidence interval.
   For this interval, we can be 95% confident that the population mean salary for all 2014 petroleum engineering graduates is between $84,863.90 and $86,630.98.

3. Did your interval capture the mean starting salary of $86,266 reported by NACE?
   The interval did capture the value of $86,266 reported by NACE.

4. Does your interval cause you to question the NACE estimate of $86,266 for the population mean? Why or why not?
   Because $86,266 is included in the interval, we do not have reason to question the NACE estimate of $86,266 for the population mean.

5. Compare and contrast the 95% confidence interval from the bootstrap distribution for the sample of size 16 and from the bootstrap distribution for the sample of size 64.
   A 95% confidence interval for the population mean that results from one simulation of 1000 samples of size 16 is $85,122.47 to $88,196.38. We would interpret the interval to say that we are 95% confident that the mean starting salary for all petroleum engineers graduating in 2014 is between $85,122.47 and $88.196.38 and centered at $86,677.21. A 95% confidence interval for the population mean that results from one simulation of 1000 samples of size 64 is from $84,863.90 to $86,630.98 and centered at $85,781.56. We can be 95% confident that the population mean salary for all 2014 petroleum engineering graduates is between $84,863.90 and $86,630.98.

   Both interval centers are close in value to the sample mean for the samples of size 16 and 64 with values of $86,684.06 and $85,784.56, respectively. The width of the interval for the sample of size 64 is smaller than the width of the interval for the sample of size 16.

6. Was the bootstrap distribution for the sample size of 64 approximately centered at the mean of the sample ($85,785)?
   The center of $85,781.56 is within $5 of the sample center of $85,785.

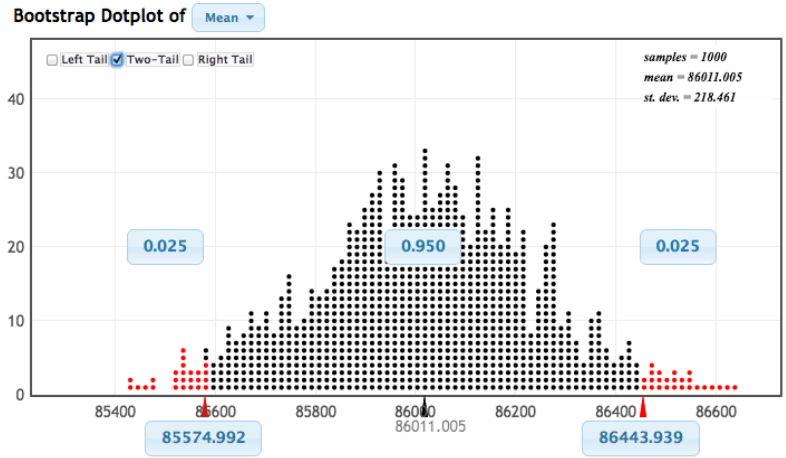
7. Calculate the half-width of the 95% confidence interval (half of the interval width or half of the difference between the upper and lower endpoints) from the sample of size 16 and the half-width of the 95% confidence interval from the sample of size 64. How are the two half-widths related?
   The half-width of the 95% confidence interval for the population mean that results from one simulation of 1000 samples of size 16 is $1519.17, and the half-width for the 95% confidence interval for the sample of size 64 is $849.42. The half-width for the 95% confidence interval for the sample is size 64 is smaller than the half-width for the 95%

confidence interval for the sample of size 16 but more than half that of that 95% confidence interval for the sample of size 16.

8. What sample size would be needed to halve the margin of error for a confidence interval estimate of the population mean starting salary yet again?

   The sample size would need to be 64*4=256 to halve the margin of error.

9. A precocious student selected a random sample of 256 starting salaries for 2014 petroleum engineering graduates. The student then followed the process outlined above to create a bootstrap distribution and 95% confidence interval and achieved the following results.



   What approximate value would you expect for the mean of this student's sample of size 256?
   We would expect the mean of the distribution of sample means to be close to the mean of the sample, so we would expect the sample mean to be close in value to $86,011.01.

10. Provide an interpretation of the 95% confidence interval for the population mean.
    We can be 95% confident that the population mean salary for all 2014 petroleum engineering graduates falls between $85,574.99 and $86,443.94.

11. Does the interval cause you to question the NACE estimate of $86,266 for the population mean? Why or why not?
    Because $86,266 is included in the interval, we do not have reason to question the NACE estimate of $86,266 for the population mean.

12. What is the margin of error? How does this margin of error relate to the margins of error you calculated in #7?
    The half-width of the 95% confidence interval for the population mean that results from one simulation of 1000 samples of size 256 is $432.93, which is approximately half of the half-width for the 95% confidence interval for the sample means for samples of size 64, which was $849.42.

Bootstrapping for Confidence in Employment

1. As before, we will use StatKey to perform the simulation efficiently.

e. Find (and record) the 95% confidence interval for the proportion of petroleum engineering graduates seeking employment.

The 95% confidence interval for the population proportion of petroleum engineering graduates seeking employment for the simulation of 1000 samples is between 0.112 and 0.199.

2. Interpret this 95% confidence interval for the population proportion.
   We can be 95% confident that the population proportion of all 2014 petroleum engineering graduates who are seeking employment is between 11.2% and 19.9%.

3. Does the interval cause you to question whether a majority of graduates are seeking employment currently? Why or why not?
   The confidence interval from #2 raises questions about whether a majority of graduates are seeking employment currently because no proportion greater than or equal to 50% is included in the interval, assuming that majority means 50% or more.

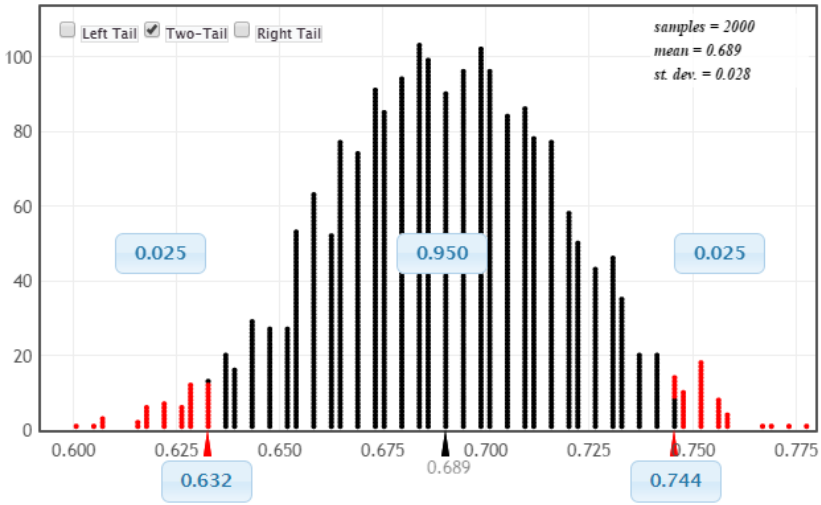
4. What is the margin of error? What does the margin of error tell you?
   The margin of error is about 4.5%. The margin of error provides an indication of the width of the confidence interval in which we would expect to find the population proportion, which would be 9%. A half-width of 4.5% suggests a cushion of 4.5% about the point estimate for the population proportion, which in the case of this simulation was 15.4%.

5. What steps could you take to decrease the margin of error?
   To decrease the margin of error, we could collect data from a larger random sample of petroleum engineering graduates or decrease the level of confidence we use to construct the confidence interval.

Try This on your Own

1. Find a 95% confidence interval for the proportion of petroleum engineering graduates who are employed full time. Interpret this interval.

   The results from one simulation for a 95% confidence interval for the proportion of petroleum engineering graduates who are employed full time appear below.



A proper interpretation of the interval would be that we can be 95% confident that the true proportion of petroleum engineering students who are employed full time is between 63.2% and 74.4% or within a margin of error of 7.5% from our sample proportion of 68.7%.

2. How would you explain your findings to the skeptical student who questioned whether graduates could get jobs in petroleum engineering? As part of your response, be sure to provide complete interpretations of the processes used to find the confidence interval and why, the confidence interval, and the margin of error.

   Because we cannot possibly ask every petroleum engineering graduate the status of his/her employment, we use a random sample of graduates and infer the actual proportion of graduates who are employed full time from the sample. By randomly selecting the sample, we achieve the highest probability for selecting a sample that is representative of the population. Our sample of 277 graduates is a fairly large sample that allows us to infer population characteristics with confidence. We consider the proportion of full-time employed graduates from this sample in relation to many samples to find an interval of reasonable values for the population proportion. If we were to collect many more samples and repeat the process (which would be quite a costly endeavor), 95% of our interval estimates would successfully capture the population proportion. Thus, we can be very confident that a majority of petroleum engineering graduates (50% or more) are able to obtain full-time employment.