

MORE 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

Student Handouts

More Confidence in Salaries in Petroleum Engineering Student Handouts

Analyzing Data from a Single Sample

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



<http://www.occupational-resumes.com/Petroleum-Engineer-Resume-Finding-a-qualified-Resume-Writer-for-a.php>

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. In this series of activities, you will explore whether the drop in crude oil prices at the end of 2015 was accompanied by a drop in starting salaries for recent petroleum engineering graduates.



<http://www.resumeok.com/engineering-manufacturing-resume-samples/petroleum-engineer-resume-template/>

1. Suppose a sample of 16 petroleum engineering majors who graduated after 2014 reported the following starting salaries: \$35000, \$42500, \$55125, \$64875, \$65299, \$67750, \$71750, \$93750, \$93750, \$94125, \$94500, \$99875, \$100125, \$101250, \$103500, and \$106475. Represent and describe these sample data.

2. Is the mean salary from this sample equal to the mean salary from 2014 that was reported by NACE? Should it be? Why or why not?

Hypothesizing about Salaries

In reality, to definitively determine whether the mean starting salary for recent petroleum engineering graduates is \$86,266, we would need to survey every recent petroleum-engineering graduate about their starting salary. Realistically, 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 are prohibitive. We get our best guesses about characteristics of a population from using a sample randomly selected from the population.

We are interested in whether the actual mean starting salary for recent petroleum engineering graduates is \$86,266 because we suspect that the mean may have decreased after crude oil prices dropped drastically. The only data that we have available at this point are the 16 salaries from recent petroleum-engineering graduates, which you just analyzed (“Analyzing Data from a Single Sample”).

Assume that this sample was randomly selected from salaries from a representative group of recent graduates. Although the sample mean does not equal \$86,266, does it provide evidence to suggest that the mean starting salary for recent graduates is less than \$86,266? Or, could this sample mean have occurred by chance?

To answer these questions, we need to conduct a *test of significance*. A significance test yields an estimate of the probability that an observed data characteristic occurred by chance if the hypothesized value is indeed correct.



**“I’ve narrowed it to two hypotheses:
it grew or we shrunk.”**

<http://eugenieteasley.com/hypothesis/>

To be sure that we are clear about what we are testing, we begin by stating our hypotheses in terms of the population characteristics we are testing.

1. What population characteristic or parameter is our focus in this setting?

2. We begin significance tests with a hypothesis—the **null hypothesis (H_0)**—that our observed results occurred by chance, in this case, that the sample mean does not provide evidence of a reduced population mean. If the sample mean occurred by chance, what do we hypothesize as the population mean starting salary for recent petroleum-engineering graduates?

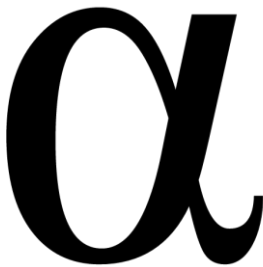
H_0 :

3. We conduct a significance test to determine whether evidence exists to cast doubt on the null hypothesis to the point where we reject the null hypothesis. The alternative to our null hypothesis is called the **alternative hypothesis**, notated as H_1 or H_a , and is the hypothesis about what we believe to be the case about the population characteristic and the hypothesis that we accept when we reject the null hypothesis. What do we hypothesize about the population mean starting salary for recent petroleum-engineering graduates?

H_1 :

4. As indicated above, a significance test yields an estimate of the probability that an observed data characteristic occurred by chance if the null hypothesis is true. What probability value(s) might cause us to question whether an observed characteristic such as a sample mean could have occurred by chance?

A low probability that a sample statistic occurred by chance raises questions about the truth or validity of the null hypothesis. Many statisticians begin to question chance occurrence with probabilities that are less than 0.05 or 0.01, which are typical threshold values that statisticians use when considering the null hypothesis. This threshold probability value is called the **alpha**



<http://www.clipartpanda.com/categories/alpha-clipart>

level or the **significance level** and is typically noted as α . When we observe probabilities less than α , we typically reject the null hypothesis in favor of the alternative hypothesis. Alternatively, when we observe probabilities greater than or equal to α , we have not proven that our null hypothesis is true but fail to reject the null hypothesis because we do not have sufficient evidence to accept the alternative.

5. If the probability of obtaining a sample mean as low as our sample mean or lower is 0.025, what should we conclude about our hypotheses?
6. How might you go about determining the probability of obtaining a sample mean as low as our sample mean or lower?

We could select additional samples of engineers and calculate their mean starting salaries to estimate the probability of obtaining a sample mean that differs from \$86,266 as much as or more than our sample 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 randomly select 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 at random. We will use the term, *randomization sample*, for each randomly selected sample formed by resampling from the original sample.



<http://www.petroleumengineer.at/petroleum-engineer/profile.html>

7. Describe a process for sampling with replacement that could be used to randomly select 16 salaries from the 16 salaries given in “Analyzing Data from a Single Sample”: \$35000, \$42500, \$55125, \$64875, \$65299, \$67750, \$71750, \$93750, \$93750, \$94125, \$94500, \$99875, \$100125, \$101250, \$103500, and \$106475.

Using Cards to Test Hypotheses



<http://beaed.com/Products/Signage/SafetySigns/tabid/1097/CategoryID/434/List/0/Level/a/ProductID/5426/Default.aspx?SortField=ProductName%2CProductName>

We wish to test whether the value of this sample mean is too much less than the value of the hypothesized mean, $H_0: \mu = 86266$, to believe that the population mean could be \$86,266. To estimate a reasonable probability for the chance of obtaining a mean as low or lower than our sample mean, we need to select many samples and calculate their sample means.

Because we want to examine a distribution of means centered at the hypothesized value of \$86,266 to see where our sample mean falls in this distribution and because we are using our sample as a best estimate for a population, we need a sample that has this hypothesized population mean of \$86,266. Because our sample mean is \$5,663 less than the hypothesized mean, we will add \$5,663 to each data value in our sample to simulate a

distribution centered at \$86,266—a set of data now consistent with the null hypothesis—and sample with replacement from this distribution. (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 “Resampling Simulation” page. To begin, remove the aces and face cards from your deck of cards.



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

1. Record the values of the sample of 16 salaries that is consistent with the null hypothesis and that we will use for resampling.

Resampling Simulation

Card	Hearts ♥				Clubs ♣				Diamonds ♦				Spades ♠			
	Ace	King	Queen	Jack	Ace	King	Queen	Jack	Ace	King	Queen	Jack	Ace	King	Queen	Jack
Salary	\$40,663	\$48,163	\$60,788	\$70,538	\$70,962	\$73,413	\$77,413	\$99,413	\$99,413	\$99,788	\$100,163	\$105,538	\$105,788	\$106,913	\$109,163	\$112,138

Sample 1

Card	Hearts ♥				Clubs ♣				Diamonds ♦				Spades ♠			
	Ace	King	Queen	Jack	Ace	King	Queen	Jack	Ace	King	Queen	Jack	Ace	King	Queen	Jack
Salary	\$40,663	\$48,163	\$60,788	\$70,538	\$70,962	\$73,413	\$77,413	\$99,413	\$99,413	\$99,788	\$100,163	\$105,538	\$105,788	\$106,913	\$109,163	\$112,138
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	\$40,663	\$48,163	\$60,788	\$70,538	\$70,962	\$73,413	\$77,413	\$99,413	\$99,413	\$99,788	\$100,163	\$105,538	\$105,788	\$106,913	\$109,163	\$112,138
Tally																
Mean																

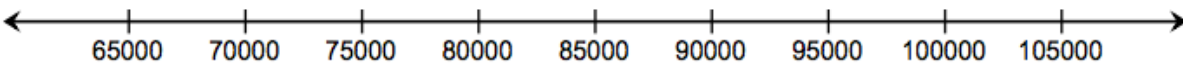
Sample 3

Card	Hearts ♥				Clubs ♣				Diamonds ♦				Spades ♠			
	Ace	King	Queen	Jack	Ace	King	Queen	Jack	Ace	King	Queen	Jack	Ace	King	Queen	Jack
Salary	\$40,663	\$48,163	\$60,788	\$70,538	\$70,962	\$73,413	\$77,413	\$99,413	\$99,413	\$99,788	\$100,163	\$105,538	\$105,788	\$106,913	\$109,163	\$112,138
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	\$40,663	\$48,163	\$60,788	\$70,538	\$70,962	\$73,413	\$77,413	\$99,413	\$99,413	\$99,788	\$100,163	\$105,538	\$105,788	\$106,913	\$109,163	\$112,138
Tally																
Mean																

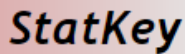
8. Record the value of each mean you calculated on a separate post-it note. Use your post-it notes to plot your four means on the class display. Examine the class distribution of means, and record it below.



9. Use the class means to estimate the probability of observing a mean as low or lower than our observed sample mean. Record your estimate here. What would you conclude about your hypotheses based on this estimate?
10. Compare and contrast this probability and your conclusions with your probability and conclusions from #6.
11. With which estimate are you more confident for drawing conclusions about recent petroleum engineering graduates' starting salaries and why?
12. How many means did you record on your dotplot in #8?

Randomizing for Significance

To estimate the probability of selecting a sample with a mean as low as or lower than our original sample mean when the null hypothesis is true, we need hundreds of randomization 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. Instead, we use computing technology to simulate the selection of 1000 or more samples and calculate their means to form a randomization distribution of means. A nice collection of applets for resampling, StatKey, is freely available at <http://lock5stat.com/statkey/>



<http://lock5stat.com/statkey/>

Go to the StatKey website, and under the heading of “Randomization Hypothesis Tests,” select the option of “Test for Single Mean.” To draw inferences about a population, we use our original sample data with \$5,663 added to each value, for the sample to have a mean of \$86,266. (As a reminder, these adjusted salaries are: \$40663, \$48163, \$60788, \$70538, \$70962, \$73413, \$77413, \$99413, \$99413, \$99788, \$100163, \$105538, \$105788, \$106913, \$109163, and \$112138.) We wish to test whether the mean of the original sample is too much less than the hypothesized mean, $H_0: \mu = 86266$, to believe that the value of \$86,266 could be the population mean. We want to examine a distribution of means centered at the hypothesized value of \$86,266 and examine where our original sample mean would fall in this distribution.



<https://freemansoftware.wordpress.com/tag/work/>

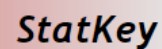
$$H_0: \mu = 86266$$

$$H_1: \mu < 86266$$

We use these data that are consistent with the null hypothesis as if they were the population data with a mean of \$86,266 to generate a probability estimate for testing the null hypothesis, $H_0: \mu = 86266$, against the alternative hypothesis, $H_1: \mu < 86266$. We then resample from these data, record the means, and plot the means to form a randomization distribution.

We use StatKey to create this distribution by following the steps listed below.

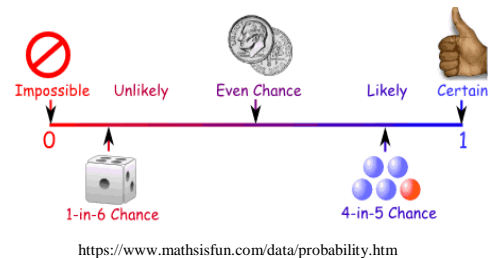
- Click on the “Edit Data” tab at the top of the screen.
- Select and delete the data that appear in the “Edit data” window.
- On the first line, enter the heading of “Salary.”
- Enter each of the 16 salaries without the dollar signs on a separate line below the heading.
- Double-check your entries, and then click “OK.”
- Enter the correct value for the null hypothesis by clicking on the displayed value for mu above the graph and then enter the value of 86266.



<http://lock5stat.com/statkey/>

1. Our adjusted sample data is now displayed in the graph labeled as “Original Sample.” Click on the “Generate 1 Sample” tab to select a single randomization sample. You should see the sample displayed in the graph labeled as “Randomization Sample.” The mean of this sample is plotted on the “Randomization Dotplot of \bar{x} ” graph. As we noted, we would like 1000 or more randomization sample means from which to estimate the probability of selecting a sample with a mean as low as or lower than our original sample mean when the null hypothesis is true. 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?

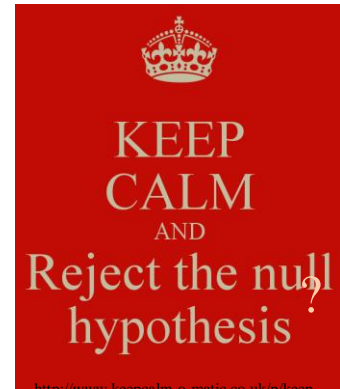
The value of the randomization distribution mean should be close to or approximately equal to our hypothesized population mean. We use the randomization distribution to determine the probability of selecting a sample with a mean as low as or lower than our original sample mean if the null hypothesis is true.



2. Locate the sample mean within the randomization distribution. Does it fall in the interval of values in the left tail, the right tail, or the middle of the randomization distribution?
3. Are there many randomization sample means that are less than or equal to the original sample mean?
4. Consider a significance level of $\alpha = 0.05$. Because the alternative hypothesis is $H_1: \mu < 86266$, you should consider only those randomization means in the left tail that are as low or lower than the observed sample mean. Click on the box at the top of the graph for “Left Tail.” The graph now displays a probability value (0.025 is the default left-tail probability) and highlights in red the means in the tail that correspond with that probability (the ratio of the number of highlighted means to the number of all randomization means displayed). The value for the rightmost of those means is listed. One way to determine whether the simulation provides sufficient evidence to doubt a population mean starting salary of \$86,266 is to change the probability value to correspond with the significance level of 0.05. To do so, click on the probability value displayed, and enter a value of 0.05. Is the observed sample mean one of means in the left tail that is highlighted in red?

5. What does your answer to #4 tell you about the probability of obtaining a mean starting salary equal to the original sample mean or even less if the null hypothesis is true?

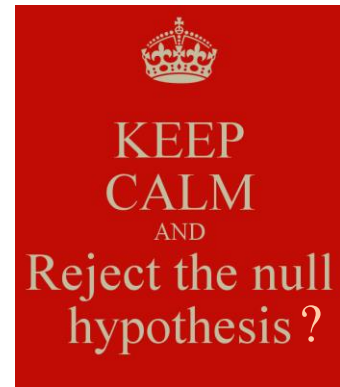
6. In terms of our hypotheses, should you reject the null hypothesis in favor of the alternative hypothesis or fail to reject the null hypothesis?



<http://www.keepcalm-o-matic.co.uk/p/keep-calm-and-reject-the-null-hypothesis/>

7. A second way to determine whether the sample provides sufficient evidence to doubt a population mean starting salary of \$86,266 is to enter the value of the original sample mean in the box displaying the value of the rightmost red mean value. Click on this value, and enter the original sample mean. What probability is displayed now?

8. In terms of our hypotheses, should you reject the null hypothesis in favor of the alternative hypothesis or fail to reject the null hypothesis?



<http://www.keepcalm-o-matic.co.uk/p/keep-calm-and-reject-the-null-hypothesis/>

9. What does your decision to reject or fail to reject the null hypothesis mean in terms of the starting salary for recent petroleum engineering graduates in relation to the starting salaries of 2014 graduates?

10. How would the process you followed differ if we had no expectation that the population mean might be less than \$86,266? In other words, how would you conduct a significance test for which the alternative hypothesis was $H_1: \mu \neq 86266$?

Connecting Confidence with Significance

The resampling process used to conduct a significance test is the same as the resampling process used to construct confidence intervals; the only difference is the shift in sample values used for resampling with the significance test.

We will use the bootstrapping method used in the “Confidence in Salaries in Petroleum Engineering” lesson to find a 95% confidence interval for the population mean starting salary for petroleum engineering graduates.



<http://www.pete.lsu.edu/research/perft/photos>

1. As before, we will use StatKey to perform the simulation efficiently.
 - a. From the main StatKey menu, select “CI for Single Mean” from the “Bootstrap Confidence Intervals” options.
 - b. You may need to click on “Edit Data,” and enter the 16 salaries from the original sample.
 - c. Generate 1000 samples.
 - d. We are interested in finding a 95% confidence interval for the average starting salary of recent petroleum engineering graduates. In particular, because we believe the salary may have gone down from the 2014 average, we are not specifically interested in the lower bound for the interval but will focus on the upper bound. As a result, you should check the box for “Right Tail” at the top of the graph and enter a value of 0.05 for the probability associated with the right tail. Record the interval, keeping in mind that there is no lower bound for the confidence interval.
2. Interpret the meaning of this interval.
3. Did your interval capture the 2014 mean of \$86,266?
4. Does the interval cause you to question whether the population mean starting salary for recent petroleum engineering graduates is less than \$86,266? Why or why not?
5. Recall your decision from the significance test you conducted and that you recorded in #9 of “Randomizing for Significance.” How does this decision compare with the conclusion you drew from the confidence interval?

6. If we had no expectation that the population mean might be less than \$86,266, only that it might be different from \$86,266, we would need to consider both lower and upper bounds for the confidence interval. Return to StatKey and check the box for “Two-Tail” at the top of the graph and enter a value of 0.025 for the probabilities associated with each tail. Record the interval.

7. Interpret the meaning of this interval.

8. Did your interval capture the 2014 mean of \$86,266?

9. Does the interval cause you to question whether the population mean starting salary for recent petroleum engineering graduates differs from \$86,266? Why or why not?

10. If you did not conduct the significance test associated with the hypotheses from #10 of “Randomizing for Significance,” conduct the significance test and record your decision here.

11. How does this decision compare with the conclusion you drew from the confidence interval?

12. Use your comparison of conclusions between significance tests and confidence intervals in #5 and #11 to draw a conjecture about the relationship between significance tests and confidence intervals.

Try This on your Own



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

84.5% of the petroleum-engineering graduates in 2014 were able to find employment (NACE, 2015a). In this activity, you will explore whether the drop in crude oil prices at the end of 2015 accompanied a drop in employment for recent petroleum engineering graduates. You select a random sample of 250 recent petroleum engineering graduates and find that 160 of them are employed. Use a randomization test to test whether the population proportion of recent petroleum engineering graduates who obtain employment is less than 84.5%

1. Record your null and alternative hypotheses for the test you will perform.
2. Use StatKey to perform the simulation efficiently.
 - a. From the main StatKey menu, select “Test for Single Proportion” from the “Randomization Hypothesis Test” options.
 - b. Click to “Edit Data,” and enter the appropriate count of graduates who are employed for a sample of size 250 consistent with the null hypothesis.
 - c. Enter the value for your null hypothesis proportion.
 - d. Generate 1000 samples.
 - e. Select the proper tail based on your alternative hypothesis, and use a significance level of $\alpha = 0.05$.

Locate the sample proportion within the randomization distribution. Does it fall in the left tail, the right tail, or in the middle of the randomization distribution?

3. Are there many randomization sample proportions that are less than or equal to the original sample proportion?
4. In terms of your hypotheses, should you reject the null hypothesis in favor of the alternative hypothesis or fail to reject the null hypothesis?

5. What does your decision to reject or fail to reject the null hypothesis mean in terms of the employment rate for recent petroleum engineering graduates in relation to the employment rate for 2014 graduates?

6. Use StatKey to find a 95% confidence interval for the population proportion employment rate for recent petroleum engineering graduates.
 - a. From the main StatKey menu, select “CI for Single Proportion” from the “Bootstrap Confidence Intervals” options.
 - b. You may need to click on “Edit Data,” and enter the count of 160 for the count of graduates who are employed and the sample of size 250.
 - c. Generate 1000 samples.
 - d. Check the appropriate box for “Left Tail,” “Two-Tail,” or “Right Tail” at the top of the graph and enter the probability associated with that option. Record the interval.

7. Interpret the meaning of this interval.

8. Did your interval capture the 2014 employment rate of 84.5%?

9. Does the interval cause you to question whether the population proportion for recent petroleum engineering graduates’ employment is less than 84.5%? Why or why not?

10. Recall your decision for the significance test you conducted and that you recorded in #4. How does the decision compare with the conclusion you drew from the confidence interval?

11. Did your conjecture for the relationship between significance tests and confidence intervals hold? If not, make a new conjecture.