# Lab #3: Bootstrapping and Confidence Intervals

This lab will use both Minitab and Statkey software to explore sampling distributions and bootstrapping. In doing so, we will analyze data from the “College Scorecard”, a government run database that collects information from all degree-granting higher education institutions each year. The original dataset contains thousands of cases (higher ed institutions) and hundreds of variables. We will only be looking at a subset cases consisting of small colleges (1000-5000 students) that primarily offer bachelor’s degrees and require the ACT as part of their admission process. While we won’t exhaustively explore everything in the data here, we will use it again in an upcoming lab on hypothesis testing. The dataset will be looking at contains the following variables:

* INSTNM The name of the institution
* CITY The city where the institution is located
* STABBR Abbreviation of the state where the institution is located
* LOCALE A categorical variable describing the location of the institution

|  |  |
| --- | --- |
| 11 | City: Large (population of 250,000 or more) |
| 12 | City: Midsize (population of at least 100,000 but less than 250,000) |
| 13 | City: Small (population less than 100,000) |
| 21 | Suburb: Large (outside principal city, in urbanized area with population of 250,000 or more) |
| 22 | Suburb: Midsize (outside principal city, in urbanized area with population of at least 100,000 but less than 250,000) |
| 23 | Suburb: Small (outside principal city, in urbanized area with population less than 100,000) |
| 31 | Town: Fringe (in urban cluster up to 10 miles from an urbanized area) |
| 32 | Town: Distant (in urban cluster more than 10 miles and up to 35 miles from an urbanized area) |
| 33 | Town: Remote (in urban cluster more than 35 miles from an urbanized area) |
| 41 | Rural: Fringe (rural territory up to 5 miles from an urbanized area or up to 2.5 miles from an urban cluster) |
| 42 | Rural: Distant (rural territory more than 5 miles but up to 25 miles from an urbanized area or more than 2.5 and up to 10 miles from an urban cluster) |
| 43 | Rural: Remote (rural territory more than 25 miles from an urbanized area and more than 10 miles from an urban cluster) |

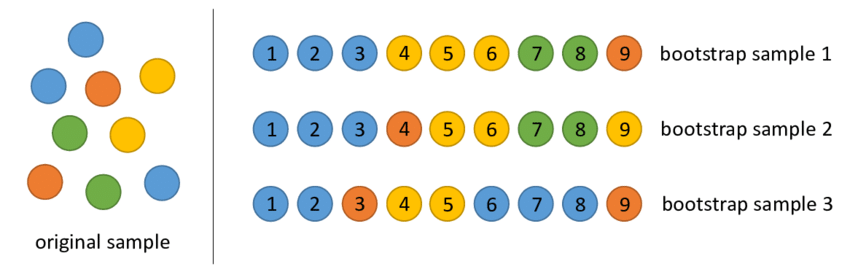
* ADM\_RATE Percentage of applicants that are admitted in to the institution
* ACTCMMID Median cumulative ACT score of enrolled students
* ACTENMID Median English ACT subscore of enrolled students
* ACTMTMID Median Math ACT subscore of enrolled students
* UGDS Total undergraduate enrollment
* PFTFAC Percent of faculty that are full time
* PCTPELL Percent of students receiving a Pell Grant
* COSTT4\_A Average yearly cost of attending
* AVGFACSAL Average faculty salary

## Bootstrapping:

So far the way that we’ve approached the construction of confidence intervals requires us to know the standard error of the sampling distribution. Basically this means that we’d be required to take thousands of different samples, a very impractical approach.

**Bootstrapping** is an ingenious idea that allows us to estimate the sampling distribution from just a single sample. When bootstrapping, we treat our original sample as if it were the population. We then sample from our sample, with replacement, to create **bootstrap samples**. Within each bootstrap sample, some of the cases from original sample could have many replicates, some might occur once, and some might not be present at all.

The diagram below illustrates the process of bootstrap sampling:

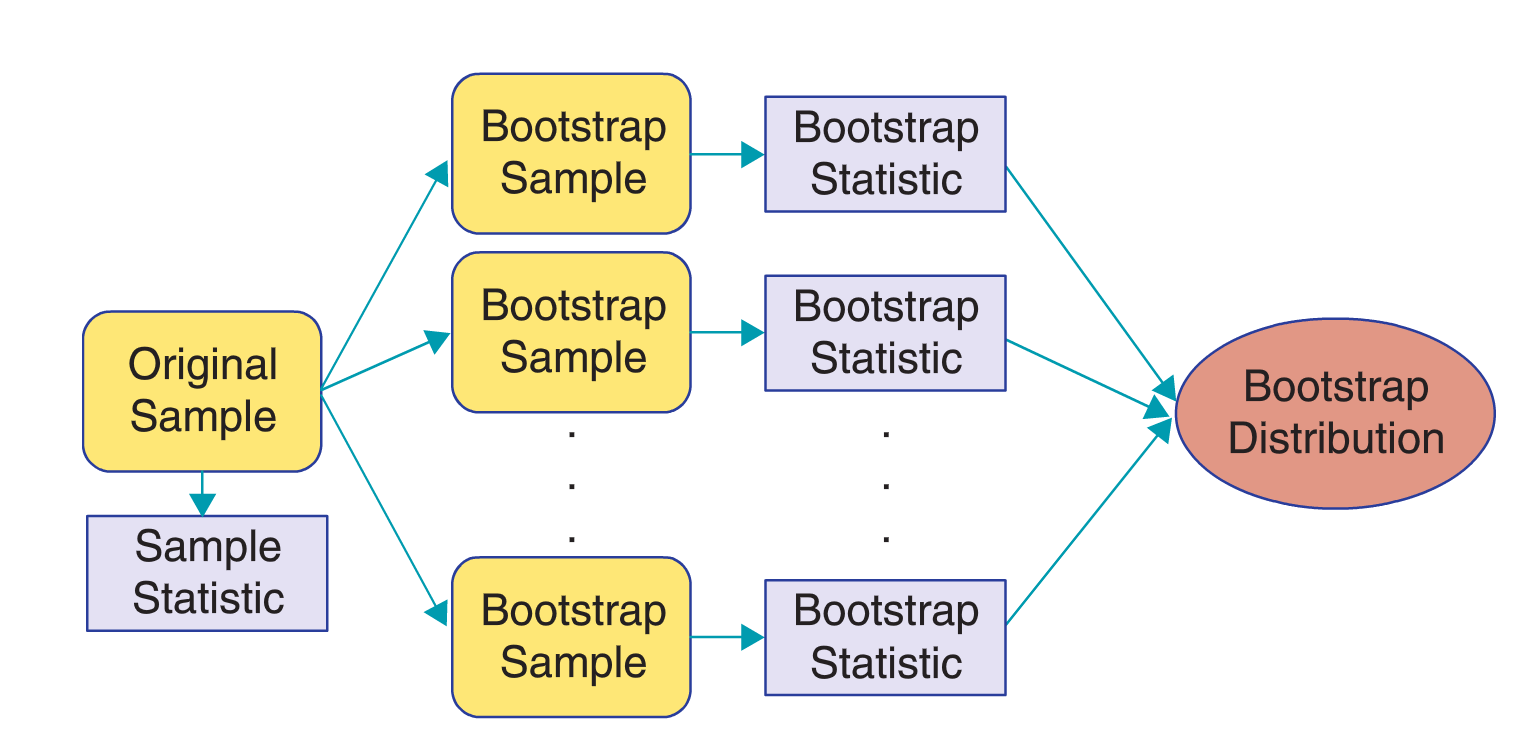


Notice how bootstrap sample 2 contains more yellow cases than were present in the original sample, this is due to bootstrapping sampling the cases with replacement.

## Question #1

Why do bootstrap samples need to be drawn with replacement? Think about the goal of replicating the sampling distribution and the role of sample size in determining the shape of the sampling distribution.

To use bootstrapping, we calculate the test statistic of interest for each of our bootstrap samples and use this collection to form the **bootstrap distribution**:



The bootstrap distribution is useful in approximating the variability (standard error) of the statistic from our original sample. The process involved is very similar to what we did when first discussing sampling distributions:

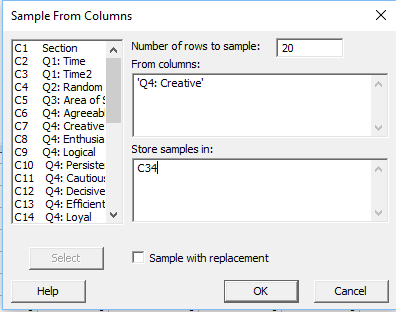
|  |  |
| --- | --- |
| **The Ideal World** | **The Bootstrap World** |
| Start with the population | Start with the original sample |
| Take many random samples from the population | Take many samples with replacement from the original sample |
| Calculate the Standard Error as the standard deviation of the sampling distribution | Calculate the Standard Error as the standard deviation of the bootstrap distribution |

The standard deviation of the bootstrap distribution turns out to be a shockingly good estimate of the standard deviation of the sampling distribution (the standard error). We will learn more about the reliability of bootstrapping in our next set of notes, but for now we will learn how to use bootstrapping to construct confidence intervals.

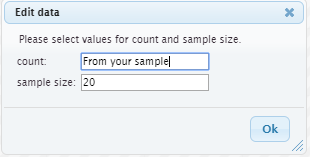
The program StatKey: <http://www.lock5stat.com/StatKey/index.html> allows us to conveniently construct bootstrap confidence intervals for a variety of different statistics, but it requires us to either use its pre-loaded datasets or to input our sample. Because the college data represent a population, we will see how to randomly generate a sampling in Minitab.

## Random Samples in Minitab:

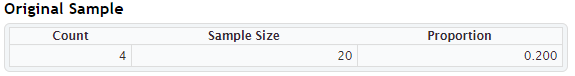
In this example our question of interest will “What percentage of students in STA209-02/04 describe themselves as creative?” Of the course this question could be answered definitively if we had access to the entire population, but we are going to limit ourselves to a random sample of 20 students. To draw a random in Minitab use: **Calc -> Random Data -> Sample from Columns**



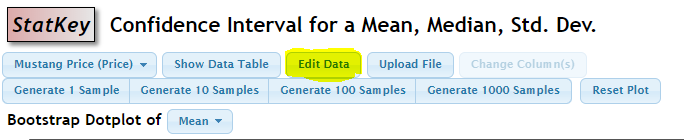
We will use this sample in StatKey, but in order to do so we need to know the number of self-described creative students in the sample. Calculate this number (ie: use a one-way frequency table) and input it into StatKey using: **Edit Data**

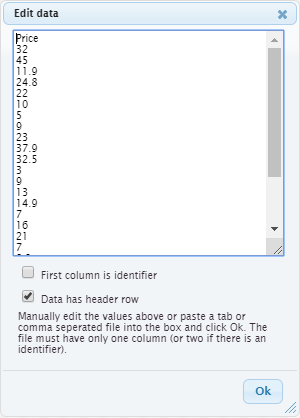


Notice that the original sample you input into StatKey will be described in the top right:



Other types of data need to be entered into StatKey differently. For example, quantitative data require you to copy and paste the entire data vector:





## Question #2

Find the true population mean cumulative ACT score and mean admissions rate for small undergraduate colleges. In your write-up, express these population parameters using the proper notation (hint: use the “equations” option under the “insert” tab in word).

## Question #3

Using Minitab, obtain a random sample of n = 20 from the College Data, paste the names of the colleges in your sample into your write-up.

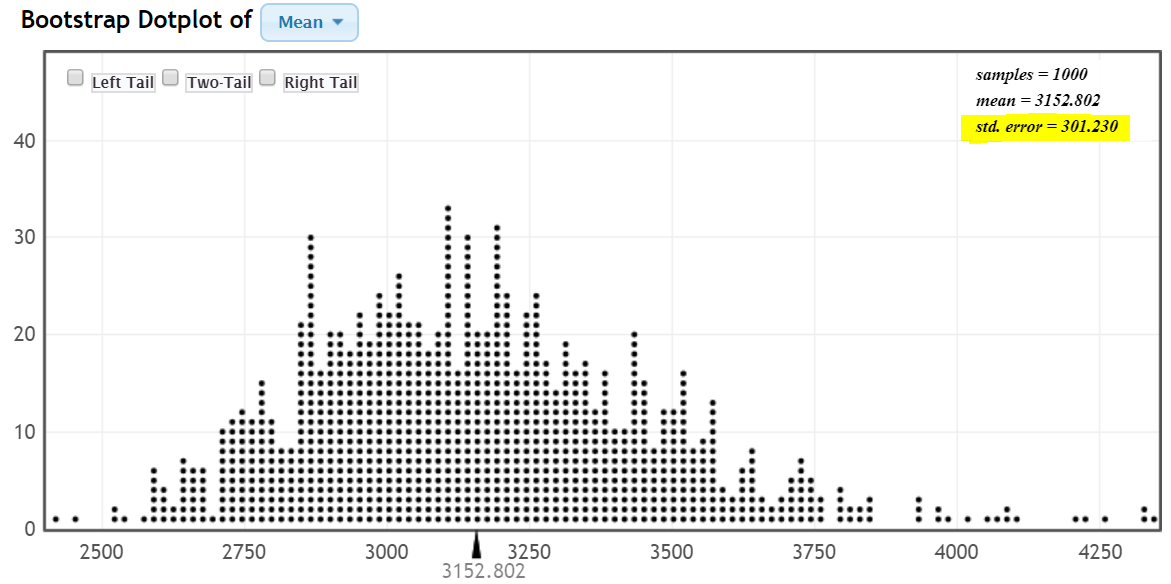
## Question #4

Find the best estimates of the mean cumulative ACT score and mean admissions rate for all small colleges using your sample. Include these in your write-up using the proper notation.

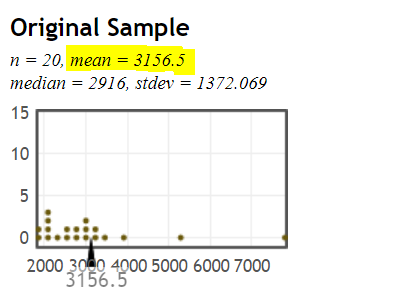
## Bootstrap Confidence Intervals using Standard Error

As previously mentioned, the standard deviation of the bootstrap distribution is a good approximation of the standard error of that statistic. Provided the bootstrap distribution is symmetric and bell shaped, this allows us to construct 95% confidence intervals using:

Here the statistic is our best estimate from the original sample and SE is the standard error estimated from the bootstrap distribution. We can use StatKey to get an estimate of the standard error, in this example I use the pre-loaded “Manhattan Apartments Rent” dataset:



We also need the estimate from the original sample:



We can construct a 95% confidence interval for the mean of the population:

## Question #5:

Use Statkey to construct a 95% bootstrap confidence interval for the mean cumulative ACT score using your sample of size n = 20. Show your work, including a copy of the bootstrap distribution you generated using StatKey.

## Question #6:

Repeat this process of using StatKey to construct a 95% bootstrap confidence interval for the mean cumulative ACT score, this time using a new random sample of n = 50. Show your work, including a copy of the bootstrap distribution.

## Question #7:

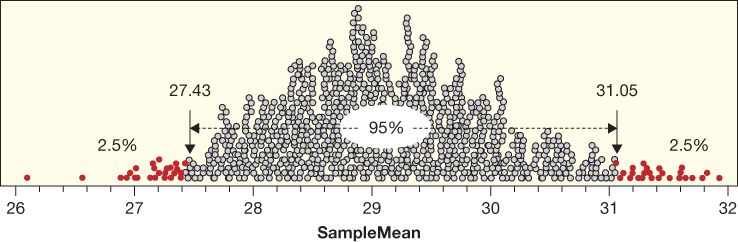
How does the width of the 95% confidence interval in Question #5 compare with the interval of Question #6? Explain in 2-3 sentences why the width of these two intervals is different.

## Confidence Intervals using Bootstrap Percentiles

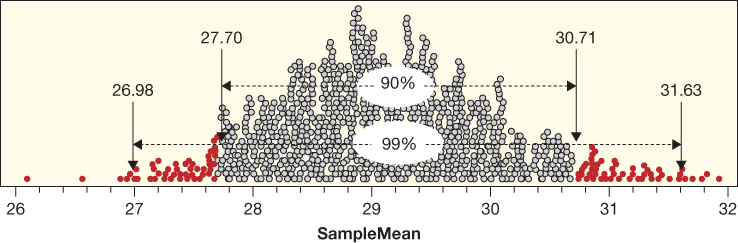
The “2SE” approach to confidence intervals can be somewhat limiting:

1. It requires the bootstrap distribution to be symmetric and bell-shaped
2. It only allows us to construct 95% confidence intervals. Sometimes the situation might warrant more or less confidence (ie: 90% confidence intervals or 99% confidence intervals)

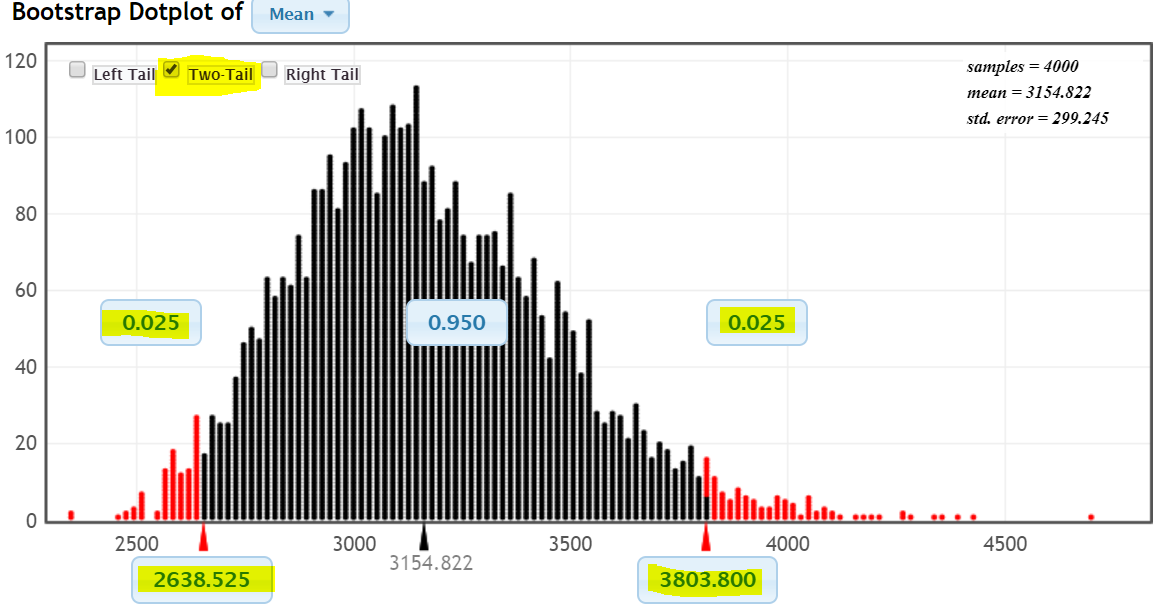
Fortunately, bootstrapping also provides a more general method using percentiles to address these concerns. Rather than using the 2 standard deviation rule to approximate the middle 95% of the bootstrap distribution, we can just chop off the lowest 2.5% and highest 2.5% of the bootstrap distribution to directly determine the middle 95% that we use to form 95% confidence intervals. This idea, called the **percentile bootstrap**, is illustrated below:



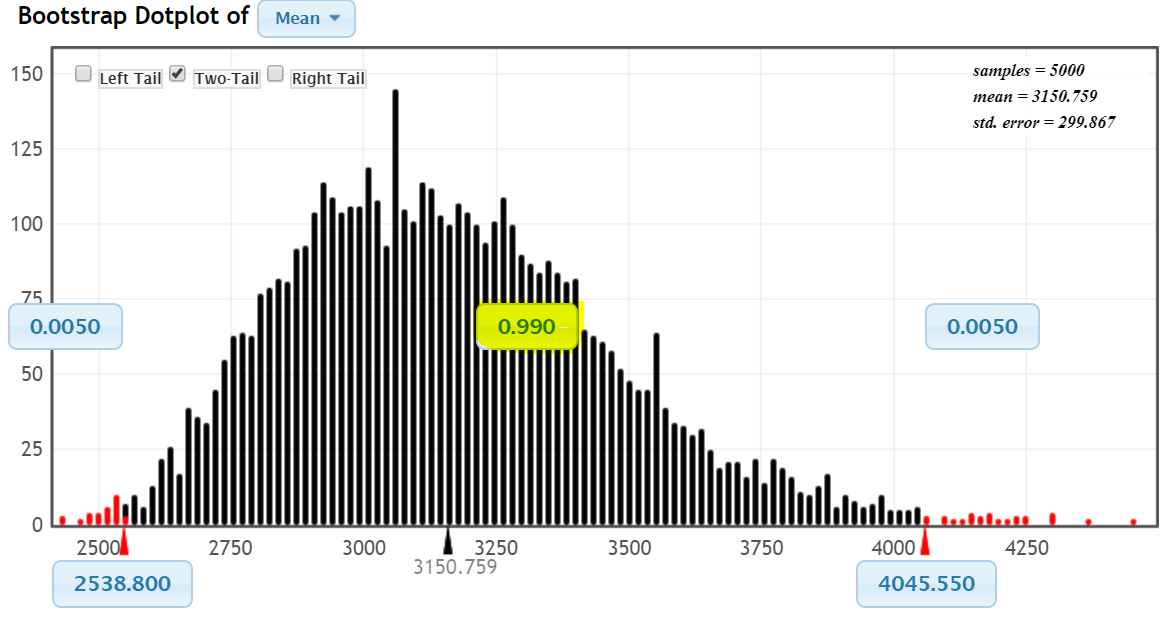
The percentile bootstrap offers us the flexibility to find valid confidence intervals of any confidence level, for bootstrap distributions of any shape. Below is an illustration of intervals at 90% and 99% confidence levels:



The percentile bootstrap confidence intervals can be found by checking the “Two Tail” box in StatKey. The example below illustrates a 95% percentile bootstrap confidence interval for the Manhattan Apartment Data:



Clicking on the box near the center of the distribution allows us to change the desired confidence level:



## Question #8:

Fill out the following table relating confidence level and interval width using the percentile bootstrap approach on your most recent sample (the one of n = 50) to construct interval estimates for the mean cumulative ACT score

|  |  |  |
| --- | --- | --- |
| **Confidence Level:** | **Interval (A, B)** | **Length (B – A)** |
| 50% |  |  |
| 70% |  |  |
| 80% |  |  |
| 90% |  |  |
| 95% |  |  |
| 99% |  |  |

## Question #9:

Create a scatter plot relating “Confidence Level” and “Length”. Is the relationship linear or non-linear? Explain what you see in 2-3 sentences, try to relate these findings to the shape of the bootstrap distribution.

## Question #10:

Based upon the various bootstrap intervals we’ve investigated in this lab, fill out the following table summarizing the impact of changing various factors on confidence interval width:

|  |  |
| --- | --- |
| **Change:** | **Impact (wider interval / narrower interval / negligible impact):** |
| Sample size increases |  |
| Number of bootstrap samples increases |  |
| Confidence level increases |  |
| Standard error increases |  |
| Confidence level decreases |  |

Question #11:

Given what you’ve seen, do you think that all values in a 95% confidence interval are equally plausible? Explain your answer in 1-2 sentences.

Question #12:

Bootstrapping is very flexible and applies to a wide range of parameters. For this question I’d like you come up with your own research question requiring you to use one of the following: correlation/regression, differences in means, or differences in proportions. Recall that these options respectively corresponds with: relating two quantitative variables, relating a quantitative and categorical variable, relating two categorical variables. In your write-up, include 3-5 sentences addressing:

1. Your research question and the two variables
2. Your sample size
3. Your bootstrap confidence interval
4. What your bootstrap confidence interval tells you about the association between those variables.

(Note: you might want to consider creating your own binary variable(s) if you choose to look at a difference in proportions or difference in means)