BSTAT 3322 Homework H04

By Craig W. Slinkman

September 24, 2015

# Objectives

There are three major objectives of this homework:

1. To continue becoming familiar with Markdown.
2. To learn how to use the R-packages and boot to compute confidence intervals.
3. To interpret such confidence intervals and use them to assess the quality of your point estimates.

The points to this homework are distributed as follows:

|  |  |
| --- | --- |
| Requirement | Points |
| Professionalism | 20 |
| R Markdown | 10 |
| 1. Recidivism | 35 |
| 2. Yellowstone | 35 |
| Total | 100 |

# Q1. Recidivism

Recidivism is one of the most fundamental concepts in criminal justice. It refers to a person's relapse into criminal behavior, often after the person receives sanctions or undergoes intervention for a previous crime. Recidivism is measured by criminal acts that resulted in rearrest, re-conviction or return to prison with or without a new sentence during a three-year period following the prisoner's release.

Bureau of Justice Statistics studies have found high rates of recidivism among released prisoners. One study tracked 404,638 prisoners in 30 states after their release from prison in 2005. The researchers found that:

Within three years of release, about two-thirds (67.8 percent) of released prisoners were rearrested.

* Within five years of release, about three-quarters (76.6 percent) of released prisoners were rearrested.
* Of those prisoners who were rearrested, more than half (56.7 percent) were arrested by the end of the first year.
* Property offenders were the most likely to be rearrested, with 82.1 percent of released property offenders arrested for a new crime compared with 76.9 percent of drug offenders, 73.6 percent of public order offenders and 71.3 percent of violent offenders.

## Q1.1 Point estimate

What is the point estimate of the *proportion* that are arrested within 3 years?

## Q1.2 Create synthetic sample

Create a synthetic sample using the size of 404,638 where 67.8% of the released prisoners are coded with a 1 and the rest are coded as 0.

## 1.3 Create a bootstrap sampling distribution

Create a bootstrap sampling distribution by bootstrapping the sample using 1000 replications. Note that this will take some time. How long will depend upon the age of your computer. It should take under 10 minutes. I have a new computer and it took about 30 seconds.

## 1.4 Assess the shape of the sampling distribution

Use the package to draw a histogram of the bootstrap sampling distribution. Give an assessment of the sampling distribution.

## 1.5 Point estimates and bias evaluation.

* What is the point estimate of the *proportion* of released prisoners who are re-arrested within after release?
* What is the bootstrap sampling bias and the absolute percentage bias? Do you need to increase the number of bootstrap replications? If so increase the number of replications.

## 1.6 Confidence interval

Select the appropriate bootstrap confidence interval. Then compute the interval. Round the interval to three decimal places. Use a 99% level of significance.

## 1.7 Interpretation of the confidence interval and analysis.

Give the point estimate value of the proportion, and the 99% confidence bounds. Write this in a short paragraph. Discuss the quality of the estimate.

# Q2. Estimating the inter-eruption interval pf the Yellowstone super volcano

The Yellowstone Caldera is the volcanic caldera and super-volcano located in Yellowstone National Park in the United States, sometimes referred to as the Yellowstone Super-volcano.

The super-volcano beneath Yellowstone National Park has not erupted for around 70,000 years. When it next erupts, scientists say most of the US will be covered by a blanket of ash, while there will be major climate changes felt across the globe.

Currently, experts say they are 99.9% confident Yellowstone will not blow in the 21st century but no one knows exactly when the next event will be.

However, there are many individuals who claim that Yellowstone will erupt very soon. In this homework we will investigate the mean time between the large-impact eruptions at the Yellowstone caldera.

The data for this problem can be found in the data folder for this homework. It is called .

Note that you will need to delete the first tow in the table because we cannot compute the interval between eruptions for the first eruption. Sample code to import the data and delete the first row is shown below.

dataFile <- # Specify file to read.  
 file.path(   
 getwd(),  
 "data",  
 "Yellowstone.csv" )  
#  
YellowStone <-   
 read.csv( dataFile ) # Import data.  
#  
YellowStone <-  
 YellowStrone[-1,] # Delete first row.

The variable of interest is . The inits of are millions of years/

## Q2.1 Graphical analysis

Draw a histogram of the Interval between eruptions. Make sure the histogram meets with the standards we have been using in this class.

## 2.2 Create a bootstrap sampling distribution

Create a bootstrap sampling distribution by using the R-function in the R-package. Use 1,000 replications.

## 2.3 Assess the shape of the sampling distribution

Use the R-package to draw a histogram of the bootstrap sampling distribution. Give an assessment of the sampling distribution. Also draw a normal Q-Q plot of the sample statistics to assess the normality of the sampling distribution. You should also carry out the Anderson-Darling test for normality.

## 2.4 Point estimates and bias evaluation

### 2.4.1 Point estimate

What is the point estimate of the *mean interval between eruptions* of the Yellowstone Caldera? Be sure to include the units in your answer. Write your answer in a sentence.

### 2.4.2 Bias analysis

What is the bootstrap sampling bias and the absolute percentage bias? Do you need to increase the number of bootstrap replications? If so, increase the number of replications till the absolute sampling bias is less than 1%. When I do this I double the sample size each time I need to rerun. Show your all work in the R-Markdown document.

## 2.5 Confidence interval

Select the appropriate bootstrap confidence interval. Then compute the interval. Round the interval to three decimal places. Use a 95% level of significance.

## 2.6 Interpretation of the confidence interval

Give the point estimate value of the interval between eruptions. Also include the 95% confidence interval and the appropriate interpretation of the interval. This should be a short paragraph of two or three sentences. Be sure to include the appropriate units.

## 2.7 Analysis

Does it appear that the estimate of the width of the 95% confidence interval is narrow enough to be useful? We do this by comparing the margin of error to the size of the estimate. You can compute the margin of error, using the following equation:

is the upper confidence limit and is the lower confidence limit.