Bootstrapping

Caelan Bryan, Jenna Dufresne, Jamielee Jimenez Perez

Fall 2022

Literature Review

Bootstrap Methods: Another Look at the Jackknife (B. Efron 1979) is credited as being the creation of bootstrapping by author B. Efron. It starts by explaining that bootstrapping is a step above the jackknife method, and that the jackknife is a linear expansion for estimating the bootstrap. Starting in the introduction, it is stated that the bootstrap is shown to estimate the variance of the sample median, which is an area that the jackknife fails at. It is also mentioned that the bootstrap does well at estimating the error rates in certain problems, which outperforms other non-parametric estimation methods. The problem attempting to be solved is estimating the sampling distribution based on the observed data. The idea behind the bootstrap method is listed in three parts. First, construct the sample probability distribution. Second, draw a random sample of size n. Lastly, approximate the sampling distribution by the bootstrap distribution. It is mentioned that that the difficult part of bootstrapping is calculating the bootstrap distribution and three methods are given to accomplish this; direct Monte Carlo approximations, and Taylor series expansion methods. A few applications are listed. These include estimating the median, error rate estimation in discrimination analysis, relationship with the jackknife, Wilcoxon's statistic, and regression models. Finally, a list of some remarks regarding the bootstrap method are listed. Some important ideas listed are that the calculation of the bootstrap distribution using the Monte Carl method is easy to implement on the computer and that the bootstrap and jackknife provide approximate frequency statements and not approximate likelihood statements.

Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap, and Other Methods (Bradley Efron 1981) starts by giving some background on what the purpose of the paper is. In this case, we want to estimate the standard error based on the data. This is normally done through parametric modeling methods, but in the paper it is being done through nonparametric methods. All methods are being tested using the same scenario which is bivariate normal distribution. This paper has four points. They are to describe the various methods, show how the methods derive from the same idea, relate the methods to each other, and finally show how the methods perform differently even though they are similar. The Monte Carlo experiment that the data was obtained from is then described. The bootstrap method is the first that is looked at. The steps are described to obtain the bootstrap samples. Random samples are created from the original sample and the bootstrap estimate is obtained. This process is than repeated multiple time to achieve enough samples to accurately estimate the standard error. 128 and 512 total samples were used separately, and while the 512 total sample provided a slightly more accurate result, the increase was minimal. on this result, the size of N samples is not overly important past 50-100. A smoothed bootstrap method is than looked at. This produced by compromising the normal theory maximum, likelihood estimate and the nonparametric maximum likelihood estimate. The results of the smoothed bootstrap method were overall better than those in the on-smoothed bootstrap method. Next, the jackknife, infinitesimal jackknife, half-sampling, and random sub sampling methods are discussed. All results are displayed in a table containing th estimated values, standard deviation, and confidence interval for each estimate. Overall, the bootstrap method produced the results that closest matches the theoretical values.

In Bootstrapping - An Introduction And Its Applications In Statistics (Hossain 2000) the authors take a high-level approach to introducing the reader to bootstrapping along with applications to different parts of statistics. The introduction begins to explain what bootstrapping is and how it is accomplished. They make use of an example to help explain what bootstrapping is. In this case, they look at performing analysis on the population of the United States. It would be difficult, costly, and timely to sample the entire population so instead you might sample a smaller subset of the population and create new bootstrap samples using replacement. In this sense the bootstrap samples might contain duplicates or even omit certain responses from the original sample. Doing so allows interpretation of the entire population based on a subset of the population. One of the first applications looked at is estimation of means and confidence intervals for the mean. Another application is constructing confidence intervals for regression coefficients. The authors also look at regression models, case re-sampling, estimating the distribution of sample mean, Bayesian bootstrap, smooth bootstrap, parametric bootstrap, and re-sampling residuals as applications of bootstrapping. Lastly, some of the advantages and disadvantages are listed. One advantage is the simplicity of deriving estimates of standard errors and confidence intervals where a disadvantage would be that the result still depends on the original sample.

In The Importance of Discussing Assumptions when Teaching Bootstrapping (Totty, Molyneux, and Fuentes 2021) the authors spend time ensuring that the readers understand when it is appropriate to use bootstrapping methods by presenting the math along with experimental results. The paper starts with an introduction to what bootstrapping is. They also mention that bootstrapping has increased in popularity since it's inception with applications in linear regression and neural networks among other fields. The methods focused on during the paper are studentized, basic, and percentile bootstrap intervals and their hypothesis tests. The next section focuses on why it's important to teach statistical computing and bootstrapping. One feature relevant to bootstrapping is that students understandings of confidence intervals and statistical inference relies on their understanding of sampling distributions. Next, some of the assumptions for bootstrapping are discussed. The assumptions were split based on interval estimation and hypothesis testing. For both cases, the largest assumption is that the distribution can be made approximately pivotal through shifting or studentization. Lastly, simulation-based performance was evaluated. This was done by using the metrics coverage proportion, significance level, and power. The most important results were that when the assumptions are broken, there can be differences in the performance of the different bootstrapping methods. There also was not a improvement between bootstrapping and other methods whose assumptions were also broken. The smaller the sample size and non-normalcy also impacted the performance of the methods. Lastly, an R package was created, which encompasses the functions used throughout the paper, that can be used to create intervals using bootstrapping methods.

Bootstrapping with R to make generalized inference for regression model (Sillabutra et al. 2016) looks at a specific application of bootstrapping, validating a generalized regression model to make generalization of statistical inference to different cases outside of the original sample. In the introduction, the authors explain the different types of regression models and explain the different ways to complete model validation. Some of those listed include cross-validation, Jackknife, and bootstrap methods. The idea of the bootstrap method is then also explained. The methodology and design of the experiment is then outlined. In this case, the original observations will be resampled leading to a set of bootstrap samples. The mean estimate and regression coefficient estimates can then be found for each bootstrap sample. Finally, confidence intervals can be created for the mean, regression coefficients, and standard errors for both the mean and regression coefficients. A table is provided to outline the values received from the original sample and bootstrap samples. Finally, the results are discussed in the conclusion. Based on the results of this experiment, the values found are very similar among the original and bootstrap samples, although the confidence intervals for the bootstrap samples are often wider. Some advantages to bootstrapping are also listed.

In the article, Bootstrapping ("Bootstrapping" 2022), it first begins with a statement from Benjamin Zimmer. He believes the origin of bootstrapping was impossible to do. This idea evolved from a man trying to jump a fence by the force of pulling up on his bootstraps. Fortunately, statistically bootstrapping is possible. Bootstrapping is taking a population, creating a small collection of data by using replacement and randomly resampling to analyze. The idea of replacement is very important when discussing bootstrapping. When replacement happens in bootstrapping it means every item that is drawn from the population, the same item exists in the sample. The importance behind sampling and replacement is when each sample or subset is made there will be statistical measurements made on each set or on all the sets together. Once the measurements are done the data is ready to plot. After the data is plotted analysis can be done. Furthermore, inferences can be made on the population as a whole. As the article continues, dives in to the importance of machine learning and how to implement bootstrapping in python. To further understand bootstrapping the article gives an example and states advantages and disadvantages.

Data Description

The dataset we will be using to explore Bootstrapping is the “Causes of Death - Our World in Data” dataset from kaggle (Chavez 2022). The causes of death dataset was also expanded with population data from the world bank (2022).

```
# Import the data
bootstrapping_data_cleaned = read.csv('bootstrapping_data_cleaned.csv',
                                      stringsAsFactors = TRUE)
```

The raw data contains a multitude of death statistics broken down by the continent, region, country, and territory. The statistics for cause of death are given in number of deaths and split up by the cause of death. Data is presented over the span of multiple years. The population dataset from the world bank is a list of countries, regions, and territories but contains the population for each year from 1960 to 2021 where available.

```
# Display a sample of the data
head(bootstrapping_data_cleaned)
```

```
##        Entity Population Code Year Deaths_Meningitis Deaths_Neoplasms
## 1 Afghanistan   12412311  AFG 1990              2159            11580
## 2 Afghanistan   13299016  AFG 1991              2218            11796
## 3 Afghanistan   14485543  AFG 1992              2475            12218
## 4 Afghanistan   15816601  AFG 1993              2812            12634
## 5 Afghanistan   17075728  AFG 1994              3027            12914
## 6 Afghanistan   18110662  AFG 1995              3102            13106
##   Deaths_FireHeatHotSubstances Deaths_Malaria Deaths_Drowning
## 1                          323             93            1370
## 2                          332            189            1391
## 3                          360            239            1514
## 4                          396            108            1687
## 5                          420            211            1809
## 6                          434            175            1881
##   Deaths_InterpersonalViolence Deaths_HIVAIDS Deaths_DrugUseDisorders
## 1                         1538             34                      93
## 2                         2001             41                     102
## 3                         2299             48                     118
## 4                         2589             56                     132
## 5                         2849             63                     142
## 6                         2969             71                     151
##   Deaths_Tuberculosis Deaths_RoadInjuries Deaths_MaternalDisorders
## 1                4661                4154                     2655
## 2                4743                4472                     2885
## 3                4976                5106                     3315
## 4                5254                5681                     3671
## 5                5470                6001                     3863
## 6                5628                6211                     4035
##   Deaths_LowerRespiratoryInfections Deaths_NeonatalDisorders
## 1                             23741                    15612
## 2                             24504                    17128
## 3                             27404                    20060
## 4                             31116                    22335
## 5                             33390                    23288
## 6                             34030                    23722
##   Deaths_AlcoholUseDisorders Deaths_ExposureToForcesOfNature
## 1                         72                               0
## 2                         75                            1347
## 3                         80                             614
## 4                         85                             225
## 5                         88                             160
## 6                         91                             381
##   Deaths_DiarrhealDiseases Deaths_EnvironmentalHeatAndColdExposure
## 1                     4235                                     175
## 2                     4927                                     113
## 3                     6123                                      38
## 4                     8174                                      41
## 5                     8215                                      44
## 6                     9566                                      46
##   Deaths_NutritionalDeficiencies Deaths_Selfharm Deaths_ConflictAndTerrorism
## 1                           2087             696                        1490
## 2                           2153             751                        3370
## 3                           2441             855                        4344
## 4                           2837             943                        4096
## 5                           3081             993                        8959
## 6                           3131            1032                        5525
##   Deaths_DiabetesMellitus Deaths_Poisonings Deaths_ProteinEnergyMalnutrition
## 1                    2108               338                             2054
## 2                    2120               351                             2119
## 3                    2153               386                             2404
## 4                    2195               425                             2797
## 5                    2231               451                             3038
## 6                    2248               467                             3087
##   Deaths_CardiovascularDiseases Deaths_ChronicKidneyDisease
## 1                         44899                        3709
## 2                         45492                        3724
## 3                         46557                        3776
## 4                         47951                        3862
## 5                         49308                        3932
## 6                         50158                        3974
##   Deaths_ChronicRespiratoryDiseases Deaths_CirrhosisOtherChronicLiverDiseases
## 1                              5945                                      2673
## 2                              6050                                      2728
## 3                              6223                                      2830
## 4                              6445                                      2943
## 5                              6664                                      3027
## 6                              6823                                      3076
##   Deaths_DigestiveDiseases Deaths_AcuteHepatitis
## 1                     5005                  2985
## 2                     5120                  3092
## 3                     5335                  3325
## 4                     5568                  3601
## 5                     5739                  3816
## 6                     5843                  3946
##   Deaths_AlzheimersDiseaseOtherDementias Deaths_ParkinsonsDisease
## 1                                   1116                      371
## 2                                   1136                      374
## 3                                   1162                      378
## 4                                   1187                      384
## 5                                   1211                      391
## 6                    1225                    394
```

To get more meaningful numbers, the population for each year 1990 through 2019 was populated into the causes of death dataset. In order to perform a more predictable experiment, the causes of death dataset was cleaned by first removing all non-country entries. The number of executions and terrorism deaths columns were also removed due to a lack of data for the majority of countries. As a final note, the Vatican and Liechtenstein are the two countries that did not have cause of death statistics available in the dataset.

The columns in the cleaned dataset are as follows:

Name	Description	Type
Entity	Name of Country	Nominal
Population	Population of Country at Specific Year	Discrete
Code	Three Letter Country Code	Nominal
Year	Year for Causes of Deaths	Nominal
Deaths_Meningitis	Number of Deaths Caused by Meningitis	Discrete
Deaths_Neoplasms	Number of Deaths Caused by Neoplasms	Discrete
Deaths_FireHeatHotSubstances	Number of Deaths Caused by Fire, Heat, or Hot Substances	Discrete
Deaths_Malaria	Number of Deaths Caused by Malaria	Discrete
Deaths_Drowning	Number of Deaths Caused by Drowning	Discrete
Deaths_InterpersonalViolence	Number of Deaths Caused by Interpersonal Violence	Discrete
Deaths_HIVAIDS	Number of Deaths Caused by HIV/AIDS	Discrete
Deaths_DrugUseDisorders	Number of Deaths Caused by Drug Use Disorders	Discrete
Deaths_Tuberculosis	Number of Deaths Caused by Tuberculosis	Discrete
Deaths_RoadInjuries	Number of Deaths Caused by Road Injuries	Discrete
Deaths_MaternalDisorders	Number of Deaths Caused by Maternal Disorders	Discrete
Deaths_LowerRespiratoryInfections	Number of Deaths Caused by Lower Respiratory Infections	Discrete
Deaths_NeonatalDisorders	Number of Deaths Caused by Neonatal Disorders	Discrete
Deaths_AlcoholUseDisorders	Number of Deaths Caused by Alcohol Use Disorders	Discrete
Deaths_ExposureToForcesOfNature	Number of Deaths Caused by Exposure to Forces of Nature	Discrete
Deaths_DiarrhealDiseases	Number of Deaths Caused by Diarrheal Diseases	Discrete
Deaths_EnvironmentalHeatAndColdExposure	Number of Deaths Caused by Environmental Heat and Cold Exposure	Discrete
Deaths_NutritionalDeficiencies	Number of Deaths Caused by Nutritional Deficiencies	Discrete
Deaths_Selfharm	Number of Deaths Caused by Self-Harm	Discrete
Deaths_ConflictAndTerrorism	Number of Deaths Caused by Conflict and Terrorism	Discrete
Deaths_DiabetesMellitus	Number of Deaths Caused by Diabetes Mellitus	Discrete
Deaths_Poisonings	Number of Deaths Caused by Poisoning	Discrete
Deaths_ProteinEnergyMalnutrition	Number of Deaths Caused by Protein Energy Malnutrition	Discrete
Deaths_CardiovascularDiseases	Number of Deaths Caused by Cardiovascular Diseases	Discrete
Deaths_ChronicKidneyDisease	Number of Deaths Caused by Chronic Kidney Disease	Discrete
Deaths_ChronicRespiratoryDiseases	Number of Deaths Caused by Chronic Respiratory Diseases	Discrete
Deaths_CirrhosisOtherChronicLiverDiseases	Number of Deaths Caused by Cirrhosis or Other Chronic Liver Diseases	Discrete
Deaths_DigestiveDiseases	Number of Deaths Caused by Digestive Diseases	Discrete
Deaths_AcuteHepatitis	Number of Deaths Caused by Acute Hepatitis	Discrete
Deaths_AlzheimersDiseaseOtherDementias	Number of Deaths Caused by Alzheimers Disease or Other Dementias	Discrete
Deaths_ParkinsonsDisease	Number of Deaths Caused by Parkinsons Disease	Discrete

In order to fit our data into our experiments below we will need to further modify the dataset. The first thing we want to do is add a new column that is sum of all causes of death columns, or the total number of reported deaths. Second, we will extract only the data from the most recent year, in the case of this dataset it is 2019. Next, we are only going to look at a subset of the available causes of death. To make things easy we will select the most common causes of death, cardiovascular diseases and neoplasms (new and abnormal growth of tissue in some part of the body). Lastly, to normalize the data between countries, we will use the population to obtain the rate of death caused by cardiovascular diseases and neoplasms.

```
# Create the total number of reported deaths column
bootstrapping_data_cleaned$Number_Of_Deaths <- rowSums(bootstrapping_data_cleaned[,(5:35)])

# Filter down the data to just include 2019 and just the columns we want
boot_data = bootstrapping_data_cleaned[bootstrapping_data_cleaned$Year == '2019',
                                       c('Entity', 'Population',
                                         'Number_Of_Deaths', 'Deaths_Neoplasms',
                                         'Deaths_CardiovascularDiseases')]

# Create the rates columns
boot_data$Deaths_CardiovascularDiseasesRate <- boot_data$Deaths_CardiovascularDiseases / boot_data$Numbe
boot_data$Deaths_NeoplasmsRate <- boot_data$Deaths_Neoplasms / boot_data$Number_Of_Deaths
boot_data$Deaths_CardiovascularDiseasesPopulationRate <- boot_data$Deaths_CardiovascularDiseases / boot_
boot_data$Deaths_NeoplasmsPopulationRate <- boot_data$Deaths_Neoplasms / boot_data$Population
```

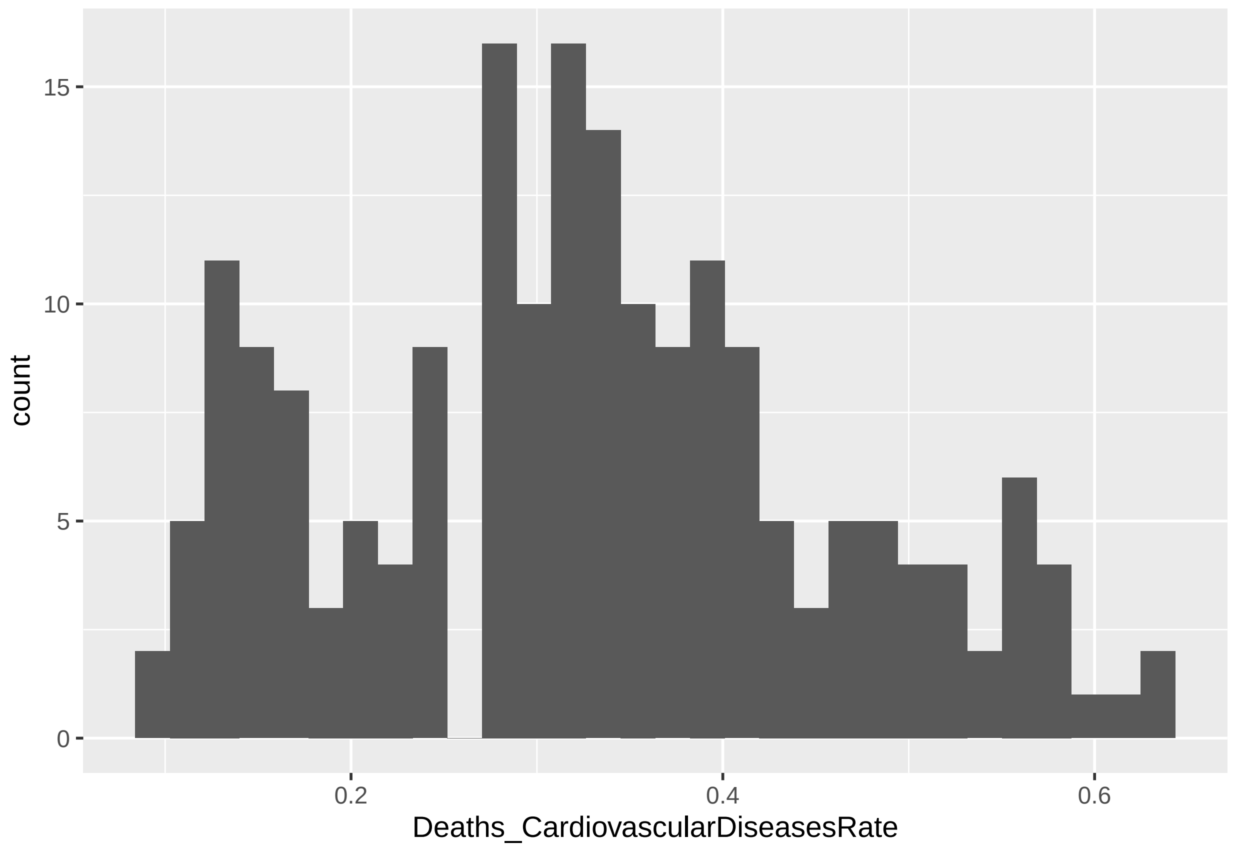
Now that we have all the data needed to move forward we can find population descriptive statistics that will be useful in comparing against our estimated population statistics from the bootstrapping method. We can also visualize the data to check that it follows a normal distribution using a histogram.

```
# Find cardiovascular disease rate statistics
boot_means <- boot_data %>% summarize(mean_cardiovasculardiseasesrate = mean(Deaths_CardiovascularDiseas
                                      mean_neoplasmsrate = mean(Deaths_NeoplasmsRate),
                                      sd_cardiovasculardiseasesrate = sd(Deaths_CardiovascularDiseasesRa
                                      sd_neoplasmsrate = sd(Deaths_NeoplasmsRate))
```

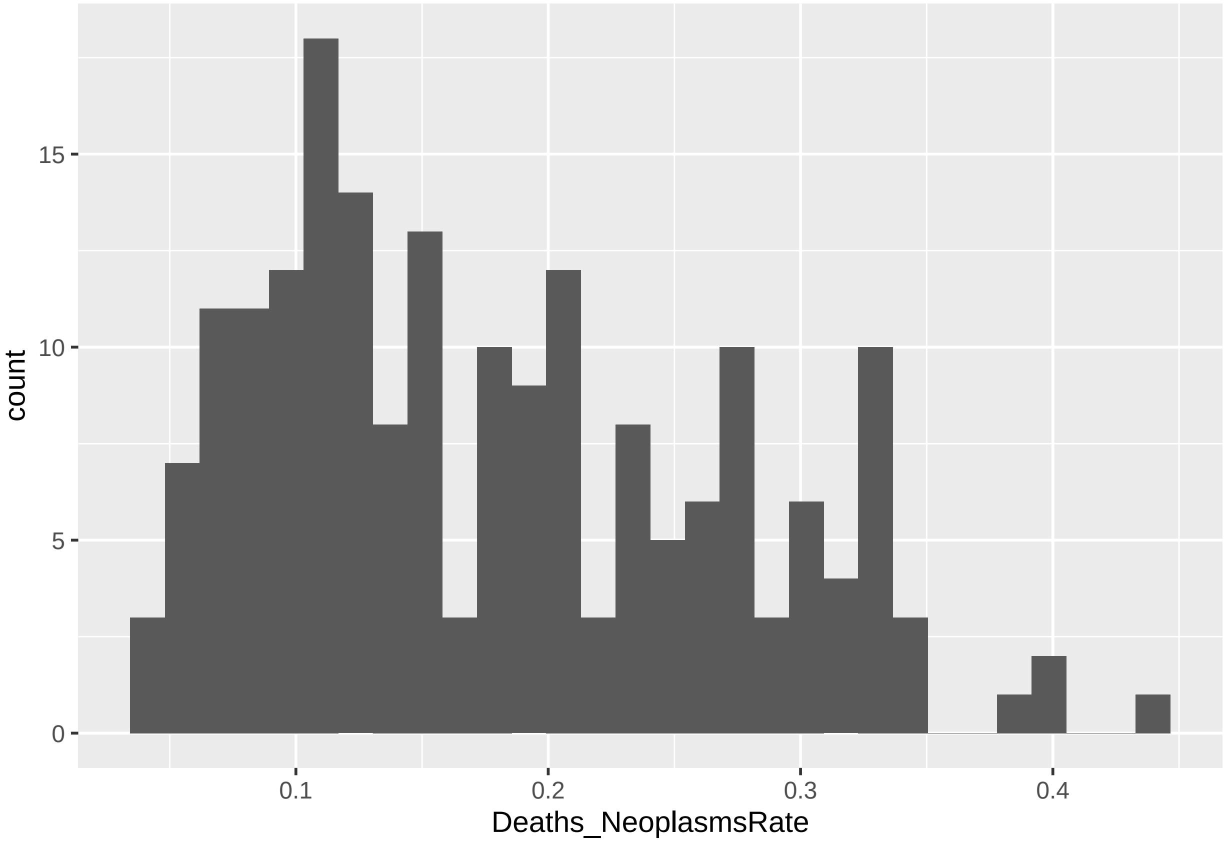
We find that the average rate of deaths caused by cardiovascular diseases across the entire population is 0.3273, or 32.73%, with a standard deviation of 0.1309, or 13.09%.

We find that the average rate of deaths caused by neoplasms across the entire population is 0.1792, or 17.92%, with a standard deviation of 0.0894, or 8.94%.

```
# Display a histogram of the cardiovascular diseases rate
ggplot(boot_data, aes(x = Deaths_CardiovascularDiseasesRate)) +
geom_histogram()
```



```
# Display a histogram of the neoplasms rate
ggplot(boot_data, aes(x = Deaths_NeoplasmsRate)) +
geom_histogram()
```



Visualizing the histograms for both the cardiovascular disease death rate and the neoplasms death rate it looks like both may follow a normal distribution. Further hypothesis tests would have to be done to validate that they do follow normal distributions. For our purposes, the visualization is good enough.

Introduction

Methods

The bootstrapping method is fairly simple to understand but results in a variety of useful applications. Bootstrapping is a method of resampling that uses random sampling with replacement to mirror the sampling process and allows us to make inferences about the entire population based on just a sample from the full population. In order to utilize the bootstrapping method, the first step is to create n number of new random samples from the current sample by allowing replacement. This means that every observation in the sample has the same probability of showing up in the new sample with every replacement. There are a variety of different types of bootstrapping that determine how to create the new samples. In the Monte Carlo method for case resampling, the new samples must have the exact size as the original sample. Bayesian bootstrap creates new samples by re-weighting the original sample. The smooth bootstrap adds random noise to each observation. The parametric bootstrap fits a parametric model to the original sample, where the new observations are pulled from the model. These are just the surface of what bootstrapping methods are available. Once the new sample(s) are created, you can analyze each sample like you would the original sample to gather information about how it performs. Finally, you can make assumptions about the entire population based on the results of the multiple samples. In our modeling, we will be using the Monte Carlo method for case resampling as it's the easiest to visualize and understand what is happening during the process. We will also use $n = 1000$ to give a good balance of the bootstrapping method and performance, although it can be scaled up to hundreds of thousands of samples.

Statistical Modeling

Estimating Population Mean and Standard Deviation The first experiment we will apply the bootstrapping method is finding the confidence intervals for the population mean and standard deviation. In order to accomplish this task we need to follow the steps of the bootstrapping method. Since we currently have an entire population worth of data, we will take a sample of the population, $n = 20$. In this example we will also be looking at only the cardiovascular diseases death rate.

```
# Create our initial sample
boot_initial_sample <- sample(1:nrow(boot_data), 20, replace = FALSE)
```

Now that we have our original sample we can use bootstrapping case resampling with the Monte Carlo method (new samples are the exact size as the original sample) to find 1,000 new samples using replacement. For each new sample created we will save the mean and standard deviation.

```
# Create vectors to store new sample means and standard deviations
boot_estimated_means <- rep()
boot_estimated_sds <- rep()
# Create 1,000 new samples and save the means and standard deviations
for (x in 1:1000) {
  boot_new_sample <- sample(boot_initial_sample, 20, replace = TRUE)
  boot_estimated_means <- append(boot_estimated_means,
                                 pull(summarize(boot_data[boot_new_sample,], mean(Deaths_CardiovascularD
  boot_estimated_sds <- append(boot_estimated_sds,
                               pull(summarize(boot_data[boot_new_sample,], sd(Deaths_CardiovascularDisea
}
```

```
# Display some estimated means
head(boot_estimated_means)
```

```
## [1] 0.3648342 0.2750236 0.3001378 0.2725383 0.3361595 0.3141000
```

```
# Display some estimated standard deviations
head(boot_estimated_sds)
```

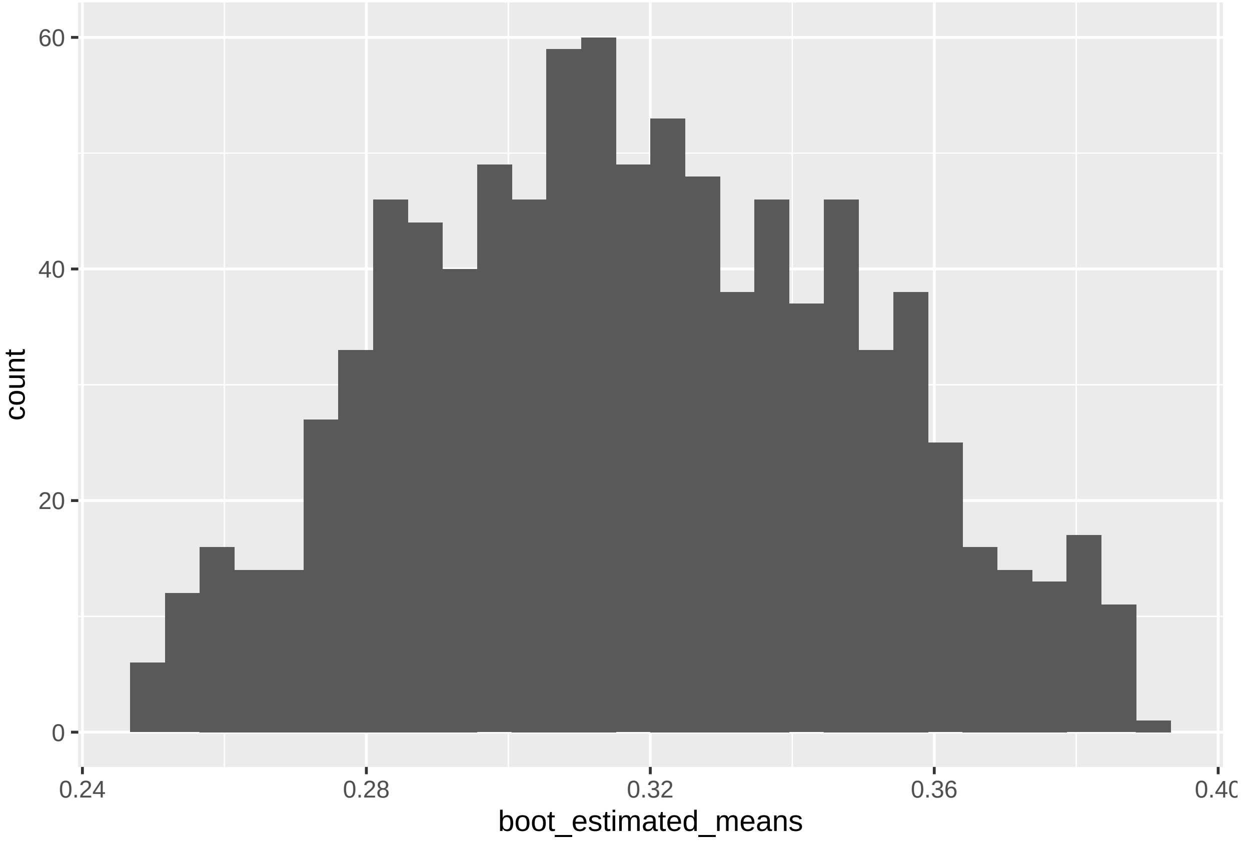
```
## [1] 0.1565556 0.1287556 0.1825296 0.1081098 0.1638658 0.1405736
```

Finally, now that we have the mean and standard deviations for all 1,000 samples creating using bootstrapping, we can trim the top and bottom 2.5% to find our 95% confidence interval for the population mean and standard deviation.

```
# Sort the estimated means from smallest to largest
boot_estimated_means <- sort(boot_estimated_means)
# Sort the estimated standard deviations from smallest to largest
boot_estimated_sds <- sort(boot_estimated_sds)
# Trim the top and bottom 2.5%
start = length(boot_estimated_means) * 0.025
end = length(boot_estimated_means) * 0.975
boot_estimated_means <- boot_estimated_means[start:end]
boot_estimated_sds <- boot_estimated_sds[start:end]
```

Now, by retrieving the first and last element in both the mean and standard deviation vectors we will find our 95% estimated population intervals. The 95% confidence interval for the population mean is [0.2471, 0.3888] or [24.71%, 38.88%]. The 95% confidence interval for the population standard deviation is [0.1166, 0.1879] or [11.66%, 18.79%].

```
# Display a histogram of the cardiovascular diseases deaths estimated population
# means rate
ggplot() +
geom_histogram(aes(boot_estimated_means))
```



Constructing Confidence Intervals on Regression Parameters The next experiment we will apply the bootstrapping method to is finding the confidence intervals for the regression parameters in a simple linear regression model to determine how variable a model is. Again, we will follow the steps of the bootstrapping method to achieve these results. To make things easier, we will use the sample initial sample created in the first experiment. For the model we will see if the deaths rate of neoplasms can be used to predict the deaths rate of cardiovascular diseases.

Again, we will use the Monte Carlo method to create $n = 1000$ new samples that are the exact size of the initial sample. For each one of these samples we will fit a simple linear regression model and save the coefficient and regression parameter.

```
# Create vectors to store new intercepts and regression parameters
boot_estimated_intercepts <- rep()
boot_estimated_regressionparameters <- rep()
# Create 1,000 new samples and save the means and standard deviations
for (x in 1:1000) {
  boot_new_reg_sample <- sample(boot_initial_sample, 20, replace = TRUE)
  boot_new_lm <- lm(Deaths_CardiovascularDiseasesRate ~ Deaths_NeoplasmsRate,
                    boot_data[boot_new_sample,])
  boot_estimated_intercepts <- append(boot_estimated_intercepts,
                                      boot_new_lm$coefficients[1])
  boot_estimated_regressionparameters <- append(boot_estimated_regressionparameters,
                                                boot_new_lm$coefficients[2])
}
```

Lastly, using the intercepts and regression parameters found we can construct our confidence intervals.

```
# Sort the estimated means from smallest to largest
boot_estimated_intercepts <- sort(boot_estimated_intercepts)
# Sort the estimated standard deviations from smallest to largest
boot_estimated_regressionparameters <- sort(boot_estimated_regressionparameters)
# Trim the top and bottom 2.5%
start = length(boot_estimated_intercepts) * 0.025
end = length(boot_estimated_intercepts) * 0.975
boot_estimated_intercepts <- boot_estimated_intercepts[start:end]
boot_estimated_regressionparameters <- boot_estimated_regressionparameters[start:end]
```

Now, by retrieving the first and last element in both the intercept and regression parameter vectors we will find our 95% estimated population intervals. The 95% confidence interval for the intercept is [0.1383, 0.1383]. The 95% confidence interval for the regression parameter is [1.1164, 1.1164].

Conlusion

Based on the results from the first experiment, estimating the population mean and standard deviation using the bootstrapping method, we can see that the true population mean and standard deviation either fall within the 95% confidence interval or are close to it. In this case the true population mean was 32.73% and our confidence interval was [24.71%, 38.88%]. Due to the nature of bootstrapping, the results may not always be perfect. Since there is replacement, while unlikely, it's entirely possible that the smallest value could be chosen to fill every single new sample. This would result in a much lower estimated population mean than the true population mean. Similarly, we bootstrapped for the estimate population standard deviation and ended up with the 95% confidence interval [11.66%, 18.79%] while the true population standard deviation is 13.09%.

In the second experiment we simulated 1,000 simple linear regression models using new samples created using the bootstrapping method to create a confidence interval for the regression intercept and parameter to ultimately see if there is much variability in the model. Our 95% confidence interval for the intercept is [0.1383, 0.1383]. Our 95% confidence interval for the regression parameter is [1.1164, 1.1164]. We can see that the intervals are either the same number or extremely small. This tells us that the model has very little variability, or it is not easily influenced by changing values. We can create the same simple linear regression model using the full population data to see that the intercept and parameter are very close or identical to the bootstrapping intervals. Increasing the size of our initial sample may increase the variability of the model.

```
boot_lm <- lm(Deaths_CardiovascularDiseasesRate ~ Deaths_NeoplasmsRate,
              boot_data)
```

Fitting a model to the full dataset gives us a intercept of 0.231 and a parameter value of 0.5373.

References

2022. https://data.worldbank.org/indicator/SP.POP.TOTL.

"Bootstrapping." 2022. *CORP-MIDS1 (MDS)*. https://www.mastersindatascience.org/learning/machine-learning-algorithms/bootstrapping/.

Chavez, Ivan. 2022. https://www.kaggle.com/datasets/ivanchvez/causes-of-death-our-world-in-data.

Efron, B. 1979. "Bootstrap Methods: Another Look at the Jackknife." *The Annals of Statistics* 7 (1): 1–26. http://www.jstor.org/stable/2958830.

Efron, Bradley. 1981. "Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap and Other Methods." *Biometrika* 68 (3): 589–99. http://www.jstor.org/stable/2335441.

Hossain, Mohammad. 2000. "Bootstrapping – an Introduction and Its Applications in Statistics." *Bangladesh Journal of Scientific Research* 18 (January): 75–88.

Sillabutra, Jutatip, Prasong Kitidamrongsuk, Chukiat Viwatwongkasem, Chareena Ujeh, Siam Sae-tang, and Khanokporn Donjdee. 2016. "Bootstrapping with r to Make Generalized Inference for Regression Model." *Procedia Computer Science* 86: 228–31. https://doi.org/https://doi.org/10.1016/j.procs.2016.05.103.

Totty, Njesa, James Molyneux, and Claudio Fuentes. 2021. "The Importance of Discussing Assumptions When Teaching Bootstrapping." arXiv. https://doi.org/10.48550/ARXIV.2112.07737.