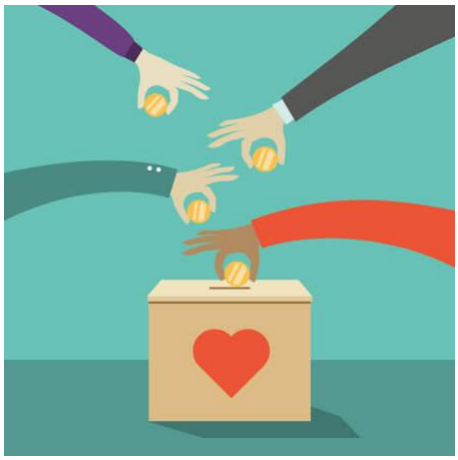
PROBABILITY MODELS
BANA 7031

FINAL PROJECT

ALUMNI DONATIONS



SUBMITTED BY:

SAGAR SAHOO (M13433382)
SANKIRNA JOSHI (M13263600)

MS BUSINESS ANALYTICS
UNIVERSITY OF CINCINNATI

1. ABSTRACT

Alumni donations are an important source of revenue for colleges and universities. If administrators could determine the factors that influence increases in the percentage of alumni who donate, they might be able to implement policies that could lead to increased revenues. The objective of the project is to apply the concepts of statistical inference to the data set. Various methods such as ECDF, parametric and nonparametric bootstrap, MLE and Bayesian were used to infer the population statistics of the sample data set.

2. INTRODUCTION

Research shows that students who are more satisfied with their contact with teachers are more likely to graduate. As a result, one might suspect that smaller class sizes and lower student-faculty ratios might lead to a higher percentage of satisfied graduates, which in turn might lead to increases in the percentage of alumni who donate. The dataset for this report has been sourced from the repository: "https://bgreenwell.github.io/uc-bana7052/data/alumni.csv ". It shows data for 48 national universities (America's Best Colleges, Year 2000 Edition). We dive into the details of the dataset to obtain certain key statistical inferences about the features of the dataset and perform various statistical test and show our findings in this report.

3. VARIABLE DEFINITION

- **School** - University name
- **percent_of_classes_under_20** - The Percentage of classes offered with fewer than 20 students
- **student_faculty_ratio** - The number of students enrolled divided by the total number of faculty
- **alumni_giving_rate** - The percentage of alumni that made a donation to the university
- **private** - If the school is private or not

4. DESCRIPTIVE ANALYSIS:

We use the **summary** command to have an overview of the variables in the alumni dataset. The results are presented in the table below.

Variable	Min.	1st Qu	Med	Mean	3rd Qu	Max.
percent_of_classes_under_20	29	44.75	59.5	55.72917	66.25	77
student_faculty_ratio	3	8	10.5	11.54167	13.5	23
alumni_giving_rate	7	18.75	29	29.27083	38.5	67

Table 1.

As a part of the analysis, we would not be using all the variables of the dataset. The variables **percent_of_classes_under_20, student_faculty_ratio** and **alumni_giving_rate** are continuous

variables where-as the **private** variable private is a binary/indicator variable. The Name of the school "**School**" won't be used as it is not numerical, and the variable doesn't impact the analysis.

We first plot the continuous variables to have an idea of the distribution. Below are the histograms of the variables:

```
par(mfrow=  c(3,1))
hist(alumni$percent_of_classes_under_20)
hist(alumni$student_faculty_ratio)
hist(alumni$alumni_giving_rate)
```

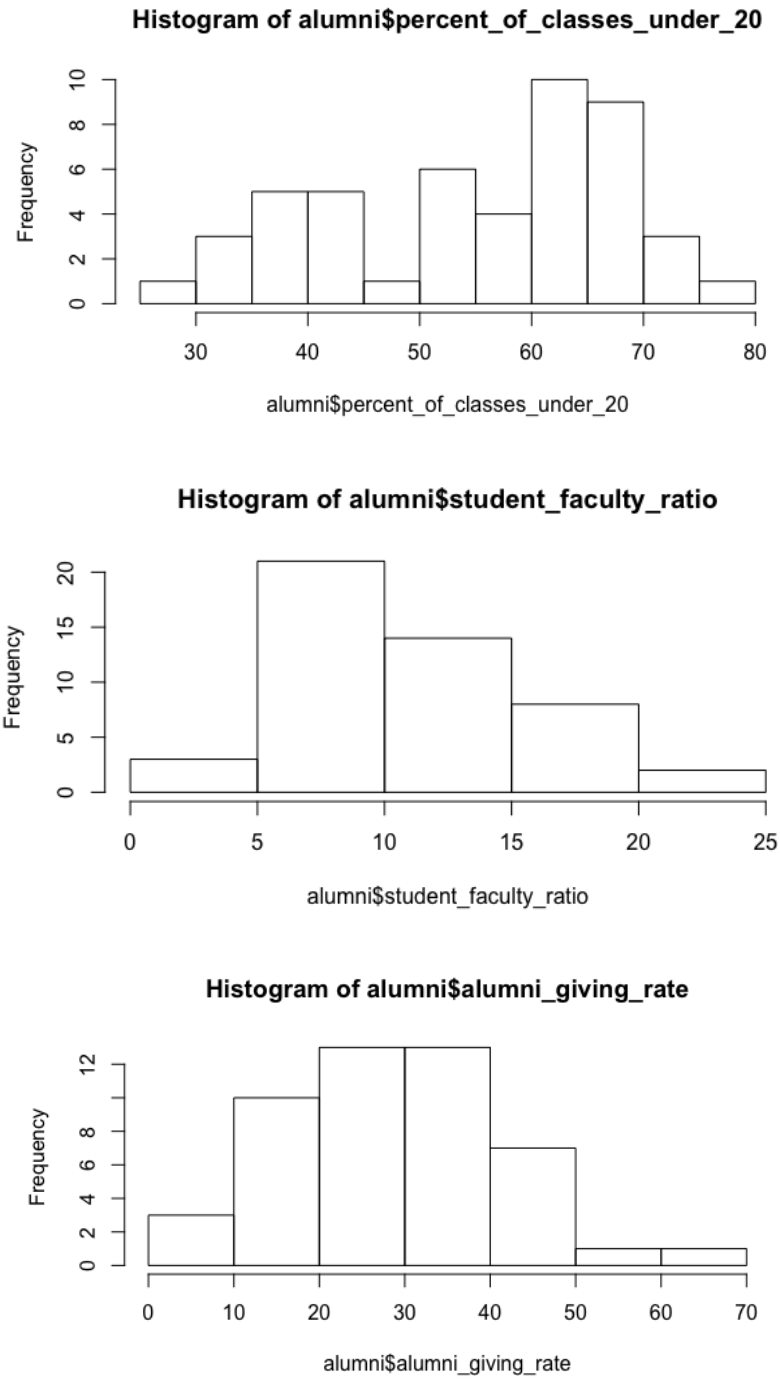


Figure 1.

From the above plots in Figure 1., we see that **student_faculty_ratio** follows normal distribution. The **percent_of_classes_under_20** and **alumni_giving_rate** variables are also approximately normal, but slightly skewed.

5. STATISTICAL INFERENCE:

We would follow the below approach to analyse our dataset as per the Project Objective:

5.1 Empirical CDF of alumni_giving_rate
5.2 Non-Parametric Bootstrap for estimating standard errors (SE) and confidence interval (CI) for correlation between student_faculty_ratio and alumni_giving_rate
5.3 Maximum Likelihood Estimator (MLE) for difference in means of the alumni_giving_rate of private vs non-private schools. And estimate standard errors and confidence intervals using parametric bootstrap for the difference in means of alumni_giving_rate
5.4 Hypothesis testing using Wald Test for difference in means of alumni_giving_rate:

$$Null\ Hypothesis, \qquad H_0: \mu_1 - \mu_2 = 0$$

$$Alternate\ Hypothesis, \qquad H_1: \mu_1 - \mu_2 \neq 0$$

5.5 Bayesian Analysis: Incorporate prior and posterior distribution mechanism for the variable student_faculty_ratio.

5.1 EMPIRICAL CDF

An empirical distribution function is the distribution function associated with the empirical measure of a sample. This cumulative distribution function is a step function that jumps up by 1/n at each of the n data points.
We use ecdf module in R to compute the ECDF and then plot a 95% Confidence Interval for the ECDF. The plot is shown in Figure 2 below.

```
Alpha=0.05
n=length(alumni$percent_of_classes_under_20)
Eps=sqrt(log(2/Alpha)/(2*n))
grid<-seq(25,85, length.out = 1000)
plot(alumni_giving_rate.ecdf,col="blue",main='Empirical CDF of Alumni Giving Rate')
lines(grid, pmin(alumni_giving_rate.ecdf(grid)+Eps,1),col="red")
lines(grid, pmax(alumni_giving_rate.ecdf(grid)-Eps,0),col="red")
```

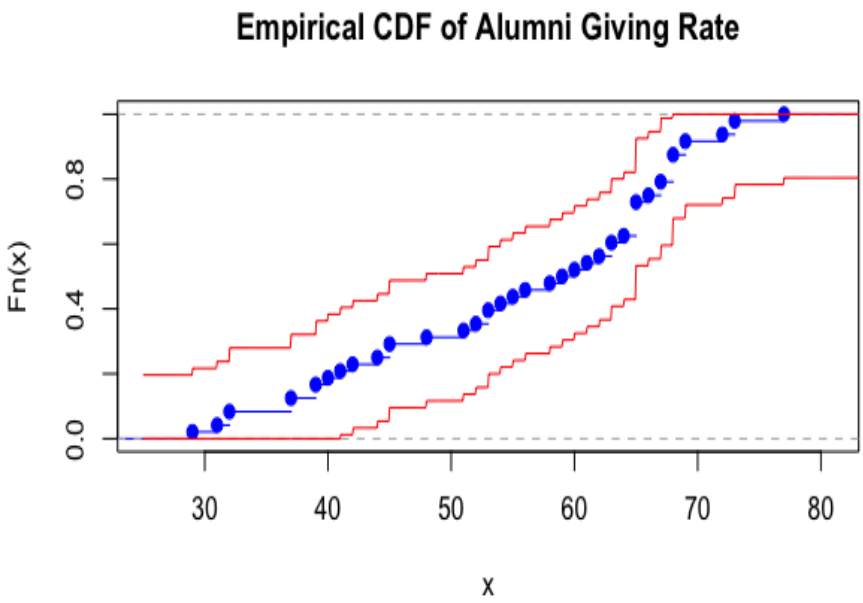


Figure 2.

5.2 NON-PARAMETRIC BOOTSRAP:

We would use non-parametric boot-strapping method to find the correlation between student_faculty_ratio and alumni_giving_rate. Using cor function, we got the correlation between them to be -0.7423975

```
cor.sample <- cor(alumni$student_faculty_ratio, alumni$alumni_giving_rate)
alumniBootSample <- alumni[,c("student_faculty_ratio", "alumni_giving_rate")]
N <- dim(alumniBootSample)[1]
cor.boot <- replicate(3000, cor(alumniBootSample[sample(1:N, size = N, replace = TRUE),])[1,2])
sd.cor.boot <- sqrt(var(cor.boot))
sd.cor.boot
```

```
## [1] 0.05267549
```

```
cor.sample
```

```
## [1] -0.7423975
```

```
hist(cor.boot,col="red", main='Histogram of Non Parametric Bootstrap samples')
```

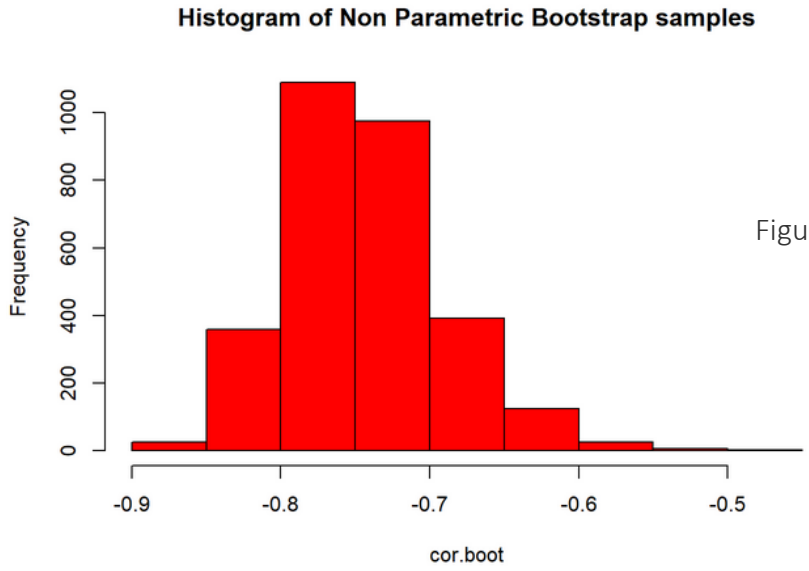


Figure 3.

```
normal.ci<-c(cor.sample-2*sd.cor.boot, cor.sample+2*sd.cor.boot)
normal.ci
```

```
## [1] -0.8477484 -0.6370465
```

```
pivotal.ci<-c(2*cor.sample-quantile(cor.boot,0.975),
              2*cor.sample-quantile(cor.boot,0.025))
pivotal.ci
```

```
##      97.5%       2.5%
## -0.8593199 -0.6540103
```

```
quantile.ci<-quantile(cor.boot, c(0.025, 0.975))
quantile.ci
```

```
##       2.5%      97.5%
## -0.8307847 -0.6254750
```

At 95%, we got the Bootstrap SE to be 0.05245229. The normal, pivotal and quantile CI's were as well found out (as shown above).

5.3 MAXIMUM LIKELIHOOD ESTIMATION

Maximum likelihood estimation (MLE) is a method of estimating the parameters of a probability distribution by maximizing a likelihood function, so that under the assumed statistical model the observed data is most probable.

We would use MLE to find the difference in Mean of Alumni Giving Rate for Private and Non Private Schools.

The dataset was divided into 2 parts based on private=1 and private = 0 (as shown in snippet). Once we calculated the mean of the sub-groups we then found our MLE for difference of means.
mu_hat = 19.78788
sigma_hat = 2.544257

```
private_1=alumni$alumni_giving_rate[alumni$private=="1"]
private_0=alumni$alumni_giving_rate[alumni$private=="0"]
n.private_1=length(private_1)
mu.private_1=mean(private_1)
sigma.private_1<-sd(private_1)
mu.private_0=mean(private_0)
sigma.private_0<-sd(private_0)
n.private_0=length(private_0)

sigma_hat<-sqrt(var(private_1)/n.private_1+var(private_0)/n.private_0)
sigma_hat
```

```
## [1] 2.544257
```

```
mu_hat=mu.private_1-mu.private_0
mu_hat
```

```
## [1] 19.78788
```

```
theta.hat<-function(x, y){
  mean(x)-mean(y)
}

boot.theta.hat<-replicate(3200, theta.hat(rnorm(n.private_1, mean = mu.private_1, sd = sigma.private_
1),
                                          rnorm(n.private_0, mean = mu.private_0, sd = sigma.private_
0)))
se<-sd(boot.theta.hat)
hist(boot.theta.hat,main = 'Histogram plot of paramteric Boot. samples')
```

For parametric bootstrapping, the "theta" function for difference of mean was replicated 3200 times using replicate function.
The Standard Error was found out to be 2.490274 and the CI to be (14.80733 24.76843)

The Boot-strapped theta values were plotted as below in Figure 4:

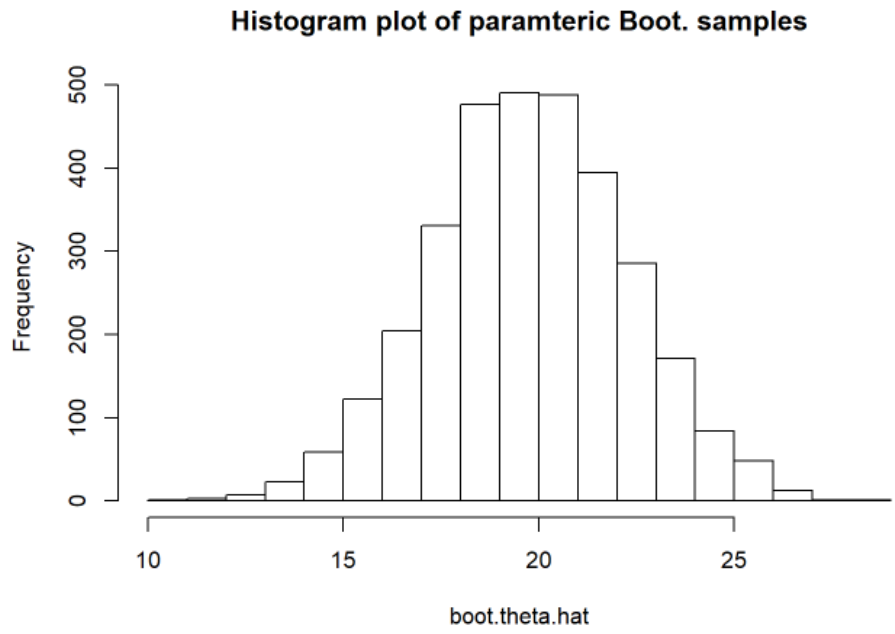


Figure 4.

5.4 HYPOTHESIS TESTING USING WALD TEST

In statistics, the Wald test assesses constraints on statistical parameters based on the weighted distance between the unrestricted estimate and its hypothesized value under the null hypothesis, where the weight is the precision of the estimate.

$$\textit{The size } \alpha \textbf{ Wald test } \textit{is: reject } H_0 \textit{ when } |W| > z_{\alpha/2} \textit{ where}$$

$$W = \frac{\widehat{\theta} - \theta_0}{\widehat{\mathrm{se}}}.$$

$\textit{Null Hypothesis, } H_0$: There is no difference in mean alumni giving rate for Private vs Non-private Schools.

Using Wald test we found that W statistic is 7.743326 and p-value is 9.769963e-15

```
z.stat <- mu_hat/se
p.value=2*(1-pnorm(abs(z.stat)))
z.stat
```

```
## [1] 7.743326
```

```
p.value
```

```
## [1] 9.769963e-15
```

At α=0.05, this implies that we have enough evidence to **reject** the null that there is no difference in mean alumni giving rate for Private and Non- private Schools.

5.5 BAYESIAN ANALYSIS:

Bayesian analysis is a statistical paradigm that answers research questions about unknown parameters using probability statements.Here, we use Bayesian analysis to find the distribution of population mean of student_faculty_ratio.

We assume 2 different prior distributions i.e with mean/standard deviation close to actual values and with values far from actual values (as shown in snippet). Then we find the posterior distribution and plot it.

Case 1: Prior Distribution with Mean as 10 and Standard Deviation as 1

```
mu.student_faculty_ratio <- mean(alumni$student_faculty_ratio)
sd.student_faculty_ratio <- sd(alumni$student_faculty_ratio)
mu.prior_1 <- 10
sd.prior_1 <- 1
Ib1 <- 1/sd.prior_1
Ix <- length(alumni$student_faculty_ratio)/(sd.student_faculty_ratio)^2
mu.posterior1 <- (mu.prior_1*Ib1 + (mu.student_faculty_ratio)*Ix)/(Ib1+Ix)
sd.posterior1 <- 1/(Ib1+Ix)
mu.posterior1
```

```
## [1] 11.03453
```

```
sd.posterior1
```

```
## [1] 0.3289542
```

```
posterior1 <- rnorm(100,mean=mu.posterior1,sd=sd.posterior1)
hist(posterior1, col=rgb(1,0,0,0.5),
     main="Overlapping Histogram for Posterior student_faculty_ratio",
     xlab="Posterior student_faculty_ratio", legend=T)
```

The Posterior Distribution was plotted below in Figure 6. The Mean was found to be 11.03453 and Standard Deviation to be 0.3289542

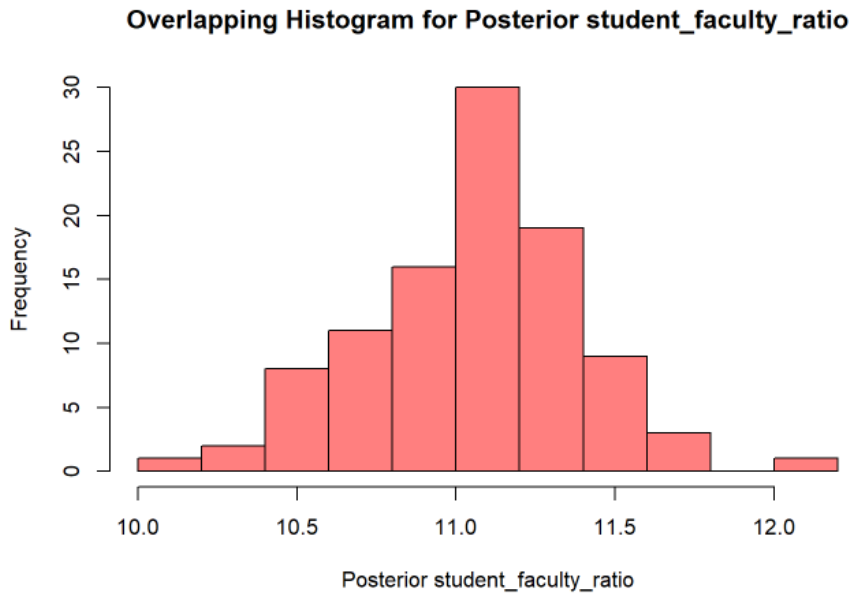


Figure 6.

Case 2: Prior Distribution with Mean as 20 and Standard Deviation as 100

```
mu.student_faculty_ratio <- mean(alumni$student_faculty_ratio)
sd.student_faculty_ratio <- sd(alumni$student_faculty_ratio)
mu.prior_2 <- 20
sd.prior_2 <- 100
Ib2 <- 1/sd.prior_2
Ix2 <- length(alumni$student_faculty_ratio)/(sd.student_faculty_ratio)^2
mu.posterior2 <- (mu.prior_2*Ib2 + (mu.student_faculty_ratio)*Ix2)/(Ib2+Ix2)
sd.posterior2 <- 1/(Ib2+Ix2)
mu.posterior2
```

```
## [1] 11.58293
```

```
sd.posterior2
```

```
## [1] 0.4878199
```

```
posterior2 <- rnorm(100,mean=mu.posterior2,sd=sd.posterior2)
hist(posterior2, col=rgb(0,0,1,0.5))
```

The Mean was found to be 11.58293 and Standard Deviation to be 0.4878199. Despite choosing a SD far from actual value, the posterior mechanism estimated value close to actual standard deviation. The Posterior Distribution was plotted below in Figure 7.

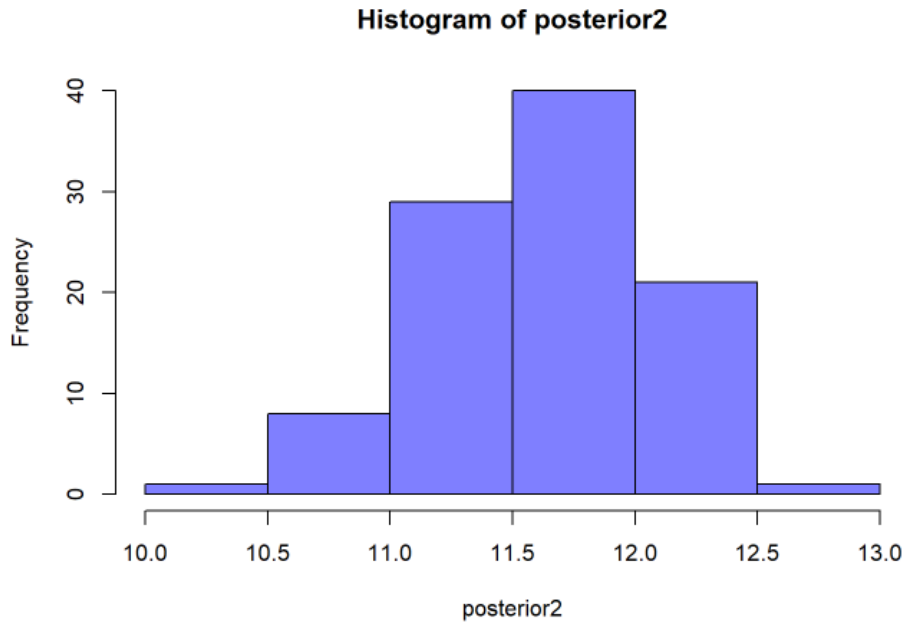


Figure 7.

6 CONCLUSION

- We implemented Empirical CDF Technique for Alumni Giving Rate.
- Thereafter, we used parametric boot strapping to estimate the correlation between Alumni Giving rate and Student Faculty Ratio.
- Using Wald test Statistic, we inferred that there is difference in alumni giving rate between private and non-private US Schools.
- And we used Bayesian analysis to find the distribution of population mean of student_faculty_ratio.