PSTAT 194CS Final Project

Jiaying Wu, Neala Rashidfarrukhi, Noa Rapoport

6/2/2022

Introduction

In this project, we will be looking at external and internal factors to analyze how these variables affect mental health using sampling methods. The data published by the CDC is a collaborative project in all of the states within the United States called the Behavioral Risk Factor Surveillance System (BRFSS). The BRFSS is "a system of ongoing health-related telephone surveys designed to collect data on health-related risk behaviors, chronic health conditions, and use of preventive services from the noninstitutionalized adult population (≥ 18 years) residing in the United States" (Overview: BRFSS 2020 1). This program collects data on health risk behaviors, chronic diseases, and conditions in order to use preventative health services related to the leading causes of disability and death in the United States. There are many variables assessed by the BRFSS during the telephone surveys, but the variables we will be looking into in this project are Physical Activity, BMI, and Education. The main question we will be posing is: How do different levels of physical activity, BMI, and Education have a direct impact on the mental health status of individuals? If we can determine that there is a positive correlation between these variables, then with scientific research there can be plausible solutions to help people who suffer from negative mental health due to lack of physical activity, lack of education, and an increased BMI.

Abstract

The first challenge we faced with this project was deciding what data to use. However, we all knew that we wanted to focus on information that was relevant and where our results would actually be important. We decided to analyze mental health because poor mental health is such a common and overlooked problem over people in general, but specifically college students. By analyzing whether mental health could be affected by education, education or BMI, we were able to calculate real and meaningful results that could be helpful for our fellow peers. In order to analyze whether there were correlations between the variables we decided to do several types of sampling. To analyze mental health and BMI we used cluster sampling, systemic sampling, and the accept/reject method to test whether BMI had any effect on mental health. Based on these methods, we found very little correlation between the two and concluded that BMI did not have a strong association with one's mental health levels. To look at the correlation between mental health and education, we used stratified sampling, bootstrap, and the inverse CDF method. Similarly to BMI, we found a pretty low correlation between the two variables. Lastly, we analyzed mental health and the frequency of physical activity using stratified sampling. Due to the distrubution of the frequency we were unable to predict the distribution of the data which is why we decided to use stratified sampling. Again, after looking at the correlation between the two variables we concluded that there was a very weak relationship between frequency of strength activities and mental health.

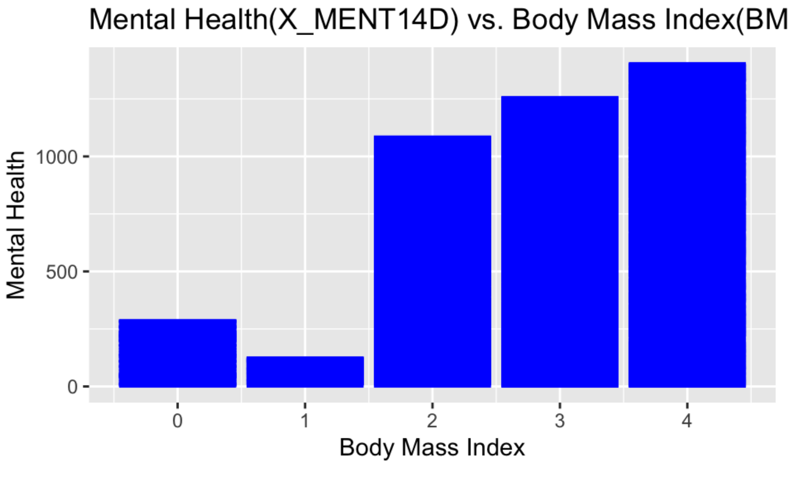
Methods and Results

Assoaciation between BMI and Mental Health Status

In order to further understand and analyze the relationship between Body Mass Index and Mental health Status we will be using two types of sampling methods: Cluster Sampling and Systemic Sampling. Cluster Sampling is when we divide data into smaller "clusters" and sample from this data. Systemic Sampling is a sampling method in which sample members from a larger population are selected according to a random starting point but with a fixed, periodic interval. This interval, called the sampling interval, is calculated by dividing the population size by the desired sample size. Both these methods are very similar to stratified sampling.

We hypothesize that there will be a weak positive assocation between BMI and Mental Health Status. We will test our hypothesis using 3 methods.
(1) Cluster Sampling
(2) Systemic Sampling
(3) Accept/Reject Method

We will look at Mental Health Status against BMI. From this we can see that it has a bell shaped curve which follows the normal distribution.

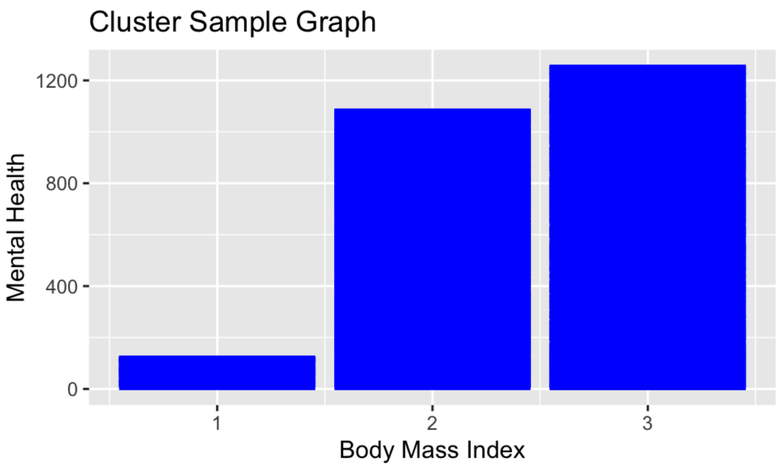


Cluster Sampling

We will use one-stage Cluster Sampling and look at BMI based off of 4 levels. 1 being underweight and 4 being obese.

```
##
##   1   2   3
##  74 649 790
```

We will look at a histogram of our sampled values. We can see that most people lie within a BMI of 2 or 3 which is normal weight or overweight.



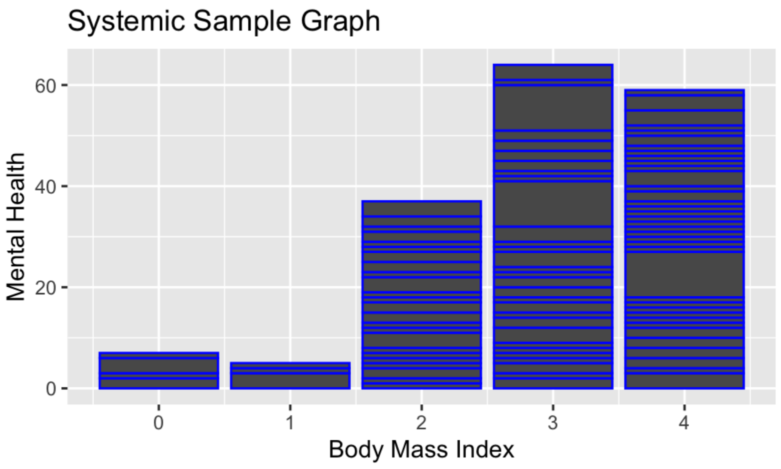
Let’s see if there is a correlation between BMI and Mental Health Status using the one-stage cluster sampling.

```
## [1] -0.03391276
```

We get a correlation of -0.0339 which is a weak neagtive correlation. Thus we cannot conclude that there is a correlation between BMI and Mental Health Status.

Systemic Sampling

Let’s use systemic sampling to sample from a large population, and divide the data into our desired sample size.



```
## [1] -0.01148764
```

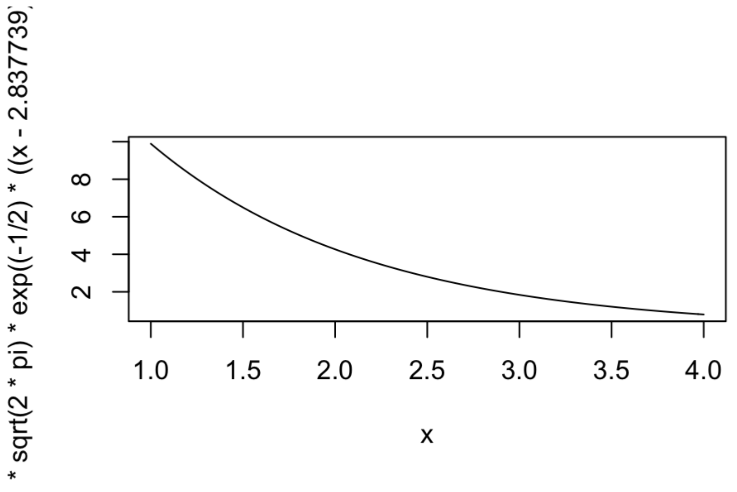
We can see again here with systemic sampling that there is a correlation of -0.11 between BMI and Mental Health Status. Thus we can conclude that there is a weak neagtive association between BMI and Mental Health Status.

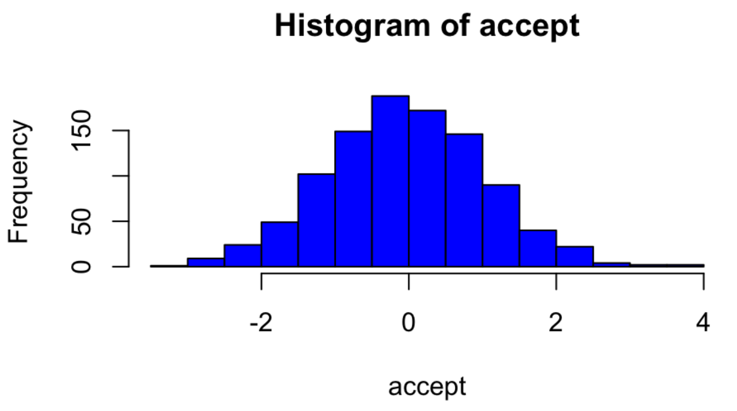
Accept/Reject Method

```
## [1] 2.837739
```

```
## [1] 1.188876
```

We know that BMI follows a normal distribution so we will fit it to a normal PDF where we calculated $\sigma = 1.188876$ and $\mu = 2.837739$.





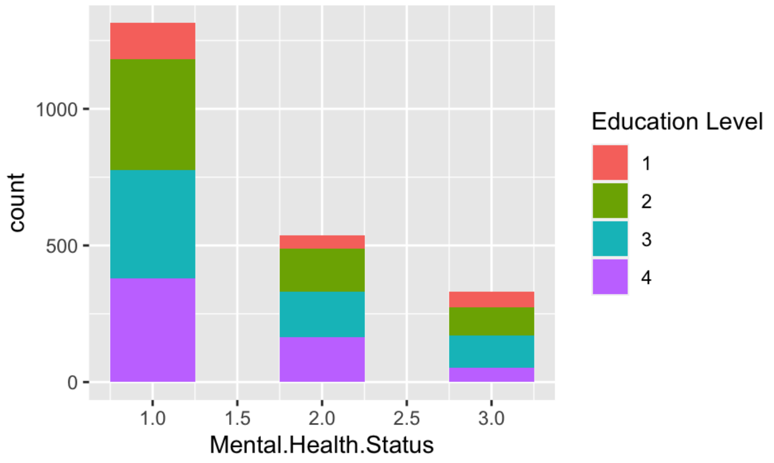
Overall, we can see that the histgram of our simulated values fits our original curve so we can say that the accept/reject method worked.

Association between Mental Health and Education

We want to see the relationship between education levels and mental health status, so we select these two columns.

Education	Mental.Health.Status
3	1

Education	Mental.Health.Status
1	1
4	1
4	1
2	3
1	3



We notice that there are different education levels and each level occupies unequal proportion of mental health status, so we consider using stratified sampling to obtain our sample. Our strata is 4 education levels. To verify our idea, we will compare the variance and standard error of each sampling method. Here we will generate 3 types of samples can compare their performance based on their variances.

1. Sample data using the `sample_n()` function, bootstrap.
2. Sample data from each stratum using the `sample_n()` function.
3. Sample data from each stratum using discrete inverse cdf method. We consider using discrete inverse cdf method because we can determine the empirical pmf of the `Mental.Health.Status` from the population so that we can easily find its inverse cdf.

Stratified Sampling

In the table below we obtain our sample size for each stratum.

Education	n	proportion	Number.of.Sample
1	239	0.11	240
2	664	0.304	663
3	684	0.313	683
4	595	0.273	596

Discrete Inverse CDF Method

In our original data, we notice that there are three status of mental health, and each status occupies different proportions. We use this result to construct the empirical pmf of mental health status.

Mental.Health.Status	n	proportion
1	1314	0.602
2	538	0.247
3	330	0.151

Suppose we want to generate random samples from a discrete random variable X with probability mass function

x	p(x)
1	0.6
2	0.25
3	0.15

The the cdf $F(x)$ is

x	$F(x)$
<0	0
$0 \leq x < 1$	0.6
$1 \leq x < 2$	0.85
$2 \leq x < 3$	1

The inverse CDF $F^{-1}(u)$ is:

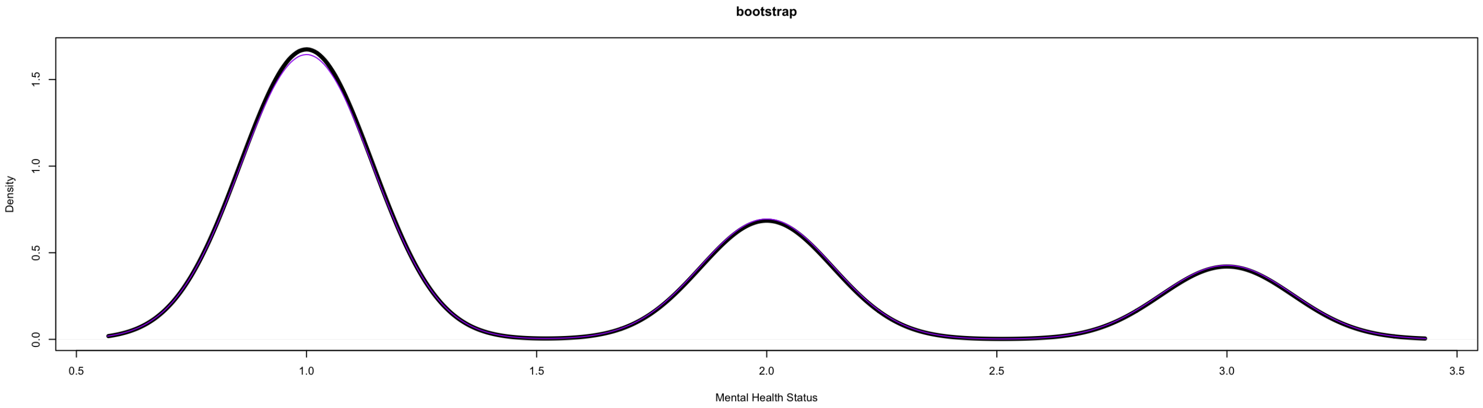
u	$F^{-1}(u)$
$u \leq 0.6$	1
$0.6 < u \leq 0.85$	2
$0.85 < u \leq 1$	3

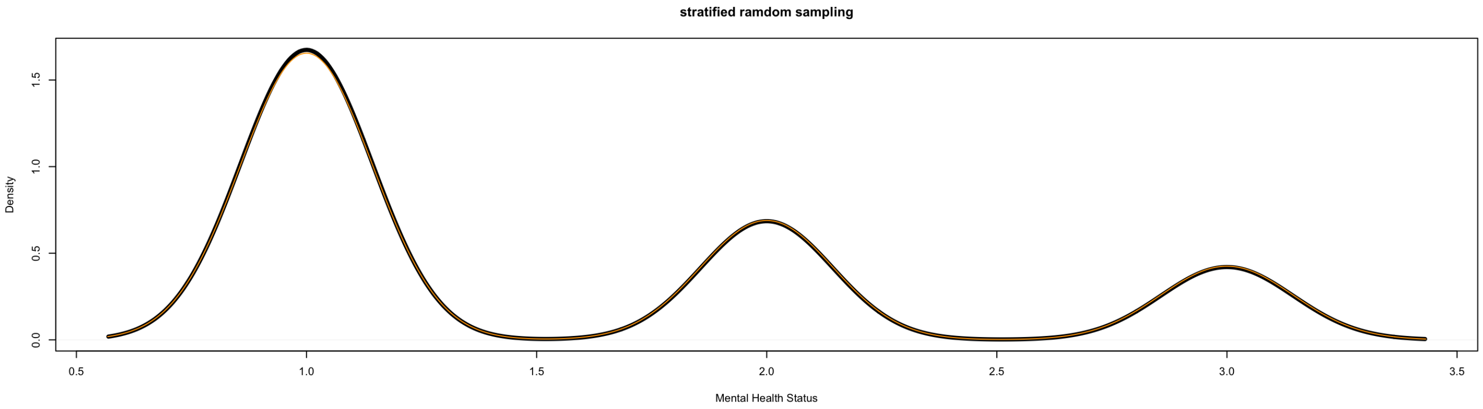
Now we check the `Mental.Health.Status_sim` proportion from stratified sampling with inverse CDF, we can see that the proportion is correspondent with our empirical pmf.

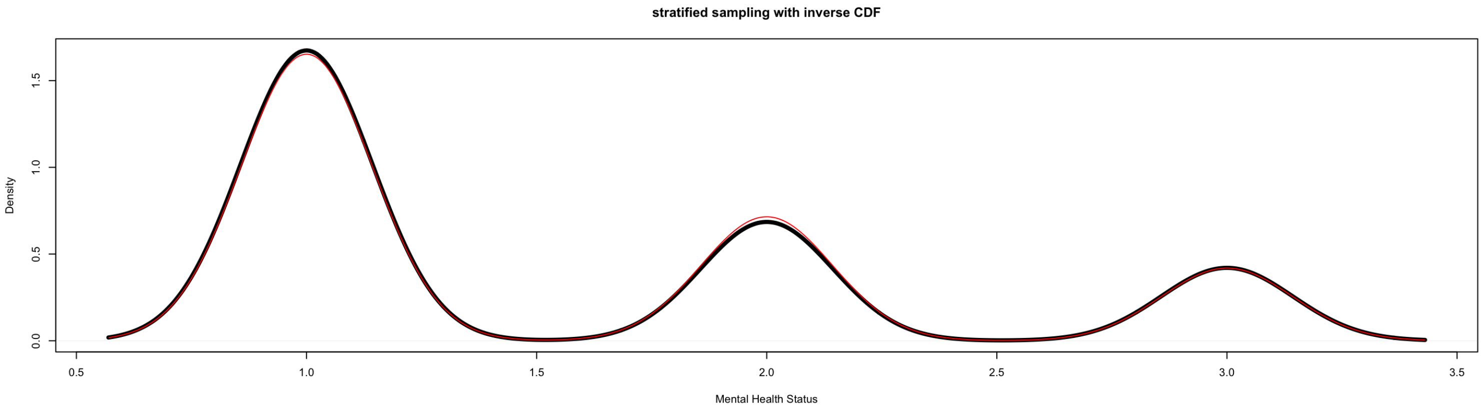
Mental.Health.Status_sim	n	proportion
1	1294	0.593

Mental.Health.Status_sim	n	proportion
2	560	0.257
3	328	0.15

We plot three samples along with the population, we can see that all of them are close to the true population, but the stratified sample using discrete inverse cdf method has the least variance and standard error. So our sampling is good.







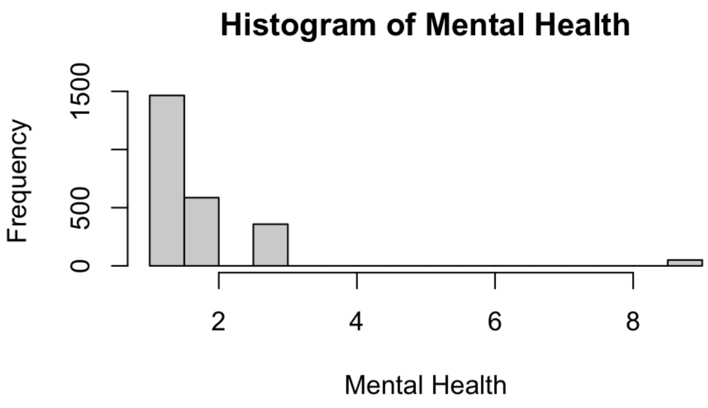
```
##                                      Variance StandardError
## original data                       0.5503223  0.0003399804
## random sample                       0.5555411  0.0003415886
## stratified ramdom sampling          0.5541856  0.0003411716
## stratified sampling with inverse CDF 0.5476108  0.0003391417
```

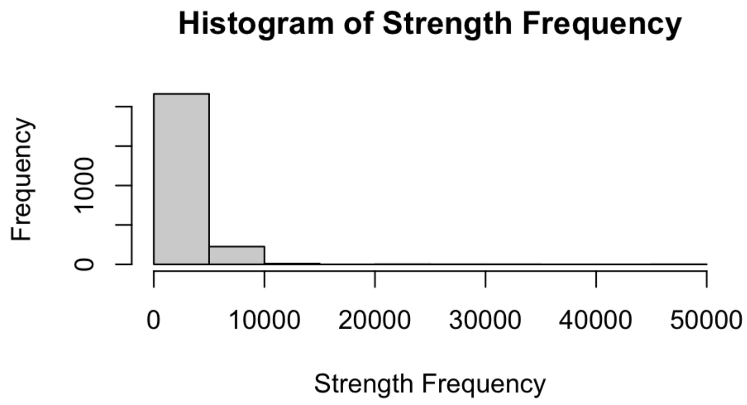
```
## [1] -0.04449628
```

The correlation turns out to be -0.044. This value is fairly low, which indicates that there is a weak association (if any) between education level and mental health status.

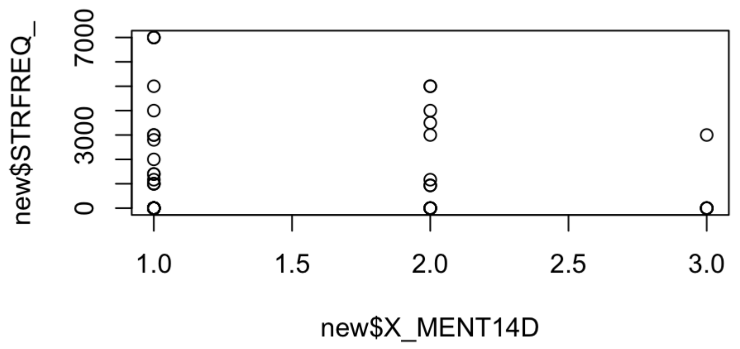
Assoaciation between Mental Health and Physical Activities

For this section, we wanted to answer the question: What is the relationship between physical and mental health? In order to answer this question, we must look at two different variables in our data, Mental Health, and Physical health. In the mental health table, a 1 signifies having 0 days with bad mental, a 2 signifies 1-13 days when mental health is not good, a 3 signifies more than 14 days with bad mental health, and a 9 signifies missing values. The second variable we will be examining is strength frequency which is the variable for strength activity frequency per week. In order to analyze our data and see how they are related we will graph a frequency graph to see the the approximate distribution of mental health, and activity frequency.

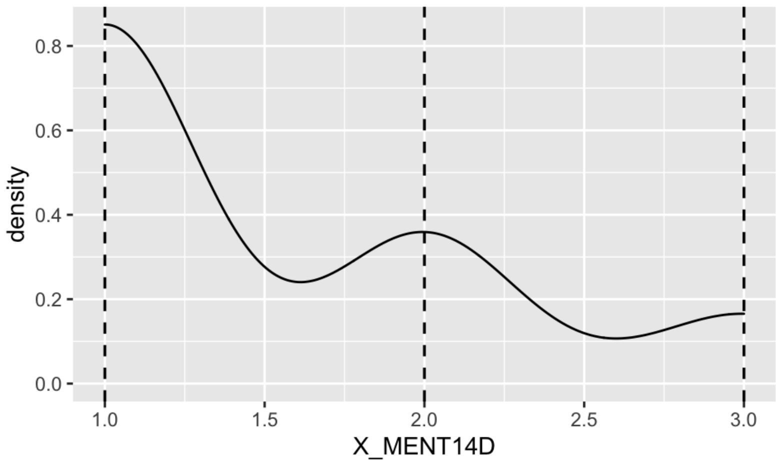


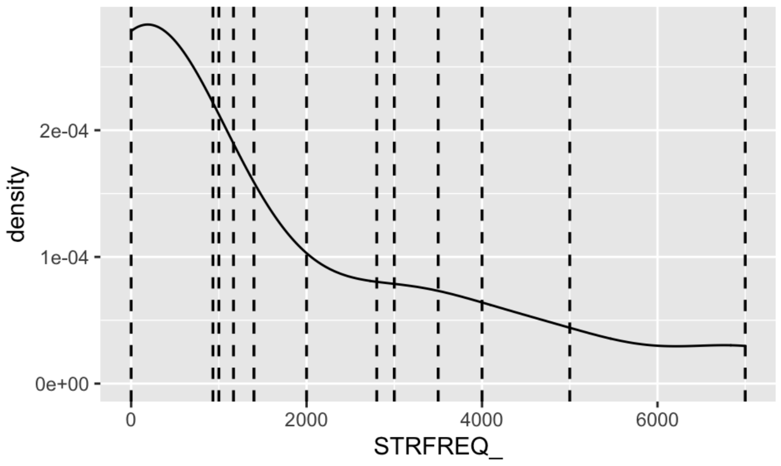


The next goal is to generate samples through stratified sampling and see if there is a correlation between the frequency of activity per week, and mental health. We split up the stratas by using the different mental health categories: 1, 2 and 3.



```
## [1] -0.104934
```





The correlation coefficient and plot shows that there is very little correlation between mental health, and the frequency that one does strength activities during a week. I believe that the reason for such low correlation is due to the fact, that many people just stated that they had very low frequency of strength activity per week, which thus resulted in very low correlation between good mental health, and high strenght activity.

Discussion

Overall, even though the questions we posed were highly relevant to our current times, and our data was accurate, we were a little disappointed with the results we obtained because we were unable to make many strong conclusions and find correlations between the mental health and each variable. I think that using different data could have been more effective in answering our original question: how is mental health affected by external and internal factors? For example, when we were looking at the relationship between whether someone had good mental health and physical activity, frequency of strength activities may not have been the best data to answer this question. Another issue that probably affected our results was the way in which the mental health data was categorized with just 1, 2, and 3 to describe the mental health of someone. It is also worth noting that it is hard to only account for mental health by looking at the number of days where one had "poor mental health" as that could mean very different things for different people. One conclusion that surprised us very little was that education and poor mental health had little correlation. A big misconception is that people with higher levels of education are often happier, but clearly our results show that this is not the case, and that no matter the educational level of someone, they can still have poor mental health. I think this project could definitely be extended by using some different data to look at the correlations.

Appendix A: Resource used

Here is a list of all resources we used in order to create this project. (https://www.statology.org/stratified-sampling-r/ (https://www.statology.org/stratified-sampling-r/)) (https://faculty.math.illinois.edu/~r-ash/Stat/StatLec1-5.pdf (https://faculty.math.illinois.edu/~r-ash/Stat/StatLec1-5.pdf)) (https://utw11041.utweb.utexas.edu/ORMM/computation/unit/rvadd/continuous_dist/beta.html# (https://utw11041.utweb.utexas.edu/ORMM/computation/unit/rvadd/continuous_dist/beta.html#):~:text=A%20linear%20transformation%20of%20a,time%20to%20accomplis (https://www.math.arizona.edu/~jwatkins/f-transform.pdf (https://www.math.arizona.edu/~jwatkins/f-transform.pdf)) (https://towardsdatascience.com/beta-distribution-intuition-examples-and-derivation-cf00f4db57af (https://towardsdatascience.com/beta-distribution-intuition-examples-and-derivation-cf00f4db57af))

Appendix B: All code for this report

```
library(knitr)
library(tidyverse)
library(dplyr)
library(pander)
library(boot)
library(tinytex)
data1 <- read.csv('brfss(2).csv')
# head(data1)
data1[is.na(data1)] = 0
ggplot(data1, aes(x=data1$BMI, y=data1$X_MENT14D)) +geom_histogram(stat='identity', col='Blue') +
  xlab("Body Mass Index") + ylab("Mental Health") +ggtitle("Mental Health(X_MENT14D) vs. Body Mass Index(BMI)")
set.seed(1)
clusters <- sample(unique(data1$BMI, data1$X_MENT14D), size=100, replace=F)

cluster_sample <- data1[data1$BMI %in% clusters, ]

table(cluster_sample$BMI)

ggplot(cluster_sample, aes(x=BMI, y=X_MENT14D)) +geom_histogram(stat='identity', col='Blue') +
  xlab("Body Mass Index") + ylab("Mental Health") + ggtitle("Cluster Sample Graph")
cor(cluster_sample[["BMI"]], cluster_sample[["X_MENT14D"]])
#define function to obtain systematic sample
set.seed(1)

obtain_sys = function(N,n){
  k = ceiling(N/n)
  r = sample(1:k, 1)
  seq(r, r + k*(n-1), k)
}

#obtain systematic sample
sys_sample_df = data1[obtain_sys(nrow(data1), 100), ]

#view first six rows of data frame
# head(sys_sample_df)
ggplot(sys_sample_df, aes(x=sys_sample_df$BMI, y=sys_sample_df$X_MENT14D)) +geom_histogram(stat='identity', col=
'Blue') +
  xlab("Body Mass Index") + ylab("Mental Health") +ggtitle("Systemic Sample Graph")
sample1<- sys_sample_df[-c(98:nrow(sys_sample_df)),]
cor(sample1[["BMI"]], sample1[["X_MENT14D"]])
mean(data1$BMI)
var(data1$BMI)
curve(1/1.188876*sqrt(2*pi)*exp((-1/2)*((x-2.837739)/1.188876))^2, 1, 4)
abline(a = .425167, b =0)

set.seed(1)

X = rnorm(4500, 0, 1) # we will reject some of these values
U = rnorm(4500, 0 ,1) # want this many because we want at least 1000 to random variables to fit our distribution


# targest distribution
pi_x <- function(x) { #accept-reject formula
  new_x = (1/1.188876*sqrt(2*pi)*exp((-1/2)*((x-2.837739)/1.188876))^2)
  return(new_x)
}

count = 1
accept = c() #empty vector that we will fill with the accepted random values from X

# Keep cycling until our sample size reaches 1000
while(count <= 4500 & length(accept) < 1000){
  test_u = U[count]
  test_x = pi_x(X[count])/(.27067*dunif(X[count],0,1))
  if (test_u <= test_x){
    accept = rbind(accept, X[count])
    count = count + 1
  }
  count = count + 1
}

hist(accept, col="Blue")
data <- read.csv("brfss.csv")
edu <- data %>% drop_na() %>%
  # filter(age != "Unsure/refused/missing" &Education != '9') %>%
   filter(Education != '9') %>%
  select(Education, X_MENT14D) %>%
  rename(Mental.Health.Status = X_MENT14D) %>%
  filter(Mental.Health.Status != '9')
  # mutate(Education = as.factor(Education))
head(edu) %>% pander(justify = "center")
ggplot(edu, aes(x = Mental.Health.Status, fill = as.factor(Education))) +
  geom_histogram(binwidth = 0.5) +
  guides(fill = guide_legend(title = "Education Level"))
strata <- edu %>%
  group_by(Education) %>%
  count()
strata <- strata %>%
  mutate(proportion = round(n/sum(strata['n']),3) ) %>%
  mutate(Number.of.Sample = round(proportion*nrow(edu)))
strata %>% pander(justify = "center")
proportion <- edu %>%
  group_by(Mental.Health.Status) %>%
  count()
proportion <- proportion %>%
  mutate(proportion = round(n/sum(proportion['n']),3) )
proportion %>% pander(justify = "center")
set.seed(5)
```

```r
# 1. Sample 200 data from the population using the `sample_n()` function.
sample2.1 <- sample_n(edu, nrow(edu), replace = TRUE)

# 2. Sample data from each stratum  using the `sample_n()` function.
sample2.2 <- NULL
for (i in 1:nrow(strata)){
  sample2.2 <- rbind(sample2.2, sample_n(edu %>% filter(Education == strata[['Education']][i]), strata[['Number.of.Sample']][i], replace = TRUE))
}

# 3. Sample data from each stratum  using discrete inverse cdf method.
inverse_cdf<-function(u){
  if(u<=0.6){
    return(1)
  }
  else if(0.6<u && u<=0.85){
    return(2)
  }

  else{
    return(3)
  }
}

sample2.3 <- NULL
for (i in 1:nrow(strata)){
  u <- runif(strata[['Number.of.Sample']][i]) # Generate random number from U[0,1]
  Mental.Health.Status_sim <- sapply(u, inverse_cdf)
  Education <- replicate(strata[['Number.of.Sample']][i],strata[['Education']][i])
  tmp <- data.frame(Education, Mental.Health.Status_sim)
  sample2.3 <- rbind(sample2.3, tmp)
}
proportion2 <- sample2.3 %>%
  group_by(Mental.Health.Status_sim) %>%
  count()
proportion2 <- proportion2 %>%
  mutate(proportion = round(n/sum(proportion2['n']),3) )
proportion2 %>% pander()
# Visualize
par( mfrow= c(3,1) )
plot(density(edu[['Mental.Health.Status']]), lwd=4,  xlab='Mental Health Status', main = "bootstrap")
lines(density(sample2.1[['Mental.Health.Status']]), col = 'purple')
plot(density(edu[['Mental.Health.Status']]), lwd=4,  xlab='Mental Health Status', main = "stratified ramdom sampling")
lines(density(sample2.2[['Mental.Health.Status']]), col = 'orange')
plot(density(edu[['Mental.Health.Status']]), lwd=4,  xlab='Mental Health Status', main = "stratified sampling with inverse CDF")
lines(density(sample2.3[['Mental.Health.Status_sim']]), col = 'red')
# legend("topright", legend=c('orignal data', 'bootstrap', 'stratified ramdom sampling', 'stratified sampling with inverse CDF'), fill =c("black","purple", "orange", "red"))
df <- data.frame(Variance=c(var(edu[['Mental.Health.Status']]),
                            var(sample2.1[['Mental.Health.Status']]),
                            var(sample2.2[['Mental.Health.Status']]),
                            var(sample2.3[['Mental.Health.Status_sim']])),
                 StandardError = c( sd(edu[['Mental.Health.Status']])/nrow(edu),
                                    sd(sample2.1[['Mental.Health.Status']])/nrow(sample2.1),
                                    sd(sample2.2[['Mental.Health.Status']])/nrow(sample2.2),
                                    sd(sample2.3[['Mental.Health.Status_sim']])/nrow(sample2.3) ))
rownames(df) <- c('original data', 'random sample', 'stratified ramdom sampling', 'stratified sampling with inverse CDF')
df
cor(sample2.3[['Education']], sample2.3[['Mental.Health.Status_sim']])
hist(data$X_MENT14D, main = "Histogram of Mental Health", xlab = "Mental Health")
hist(data$STRFREQ_, main = "Histogram of Strength Frequency", xlab = "Strength Frequency")
new_data <- select(data, 'X_MENT14D','STRFREQ_')
new_data <- na.omit(new_data)
new_data <- subset(new_data, X_MENT14D < 9)
new <- new_data %>%
  group_by('X_MENT14D') %>%
  sample_n(50)
plot(new$X_MENT14D, new$STRFREQ_)
cor(new$X_MENT14D, new$STRFREQ_)
a <- ggplot(new, aes(x = X_MENT14D))
a + geom_density() +
  geom_vline(aes(xintercept = (X_MENT14D)),
             linetype = "dashed", size = 0.6)
b <- ggplot(new, aes(x = STRFREQ_))
b + geom_density() +
  geom_vline(aes(xintercept = (STRFREQ_)),
             linetype = "dashed", size = 0.6)
```