CS544 Homework 5

Phillip Escandon - [escandon@bu.edu](mailto:escandon@bu.edu)

22 February, 2021

Table of Contents

## 1. Central Limit Theorem

Using the city of Boston earnings dataset complete 3 distributions and compare the standard deviation and mean of each

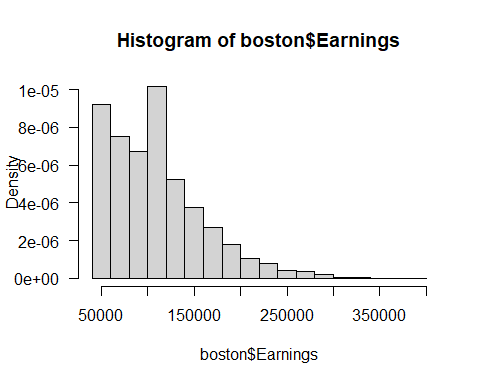
Retrieve the data

boston <- read.csv("http://people.bu.edu/kalathur/datasets/bostonCityEarnings.csv",colClasses = c("character","character","character","integer","character"))

### 1.a Show the histogram of the employee earnings. Use breaks from 40000 to 400000 in

steps of 20000 and show the corresponding tick labels on the x-axis

# Cannot get the xlim to work correctly  
s <- seq(40000,400000,20000)  
ee <- hist(boston$Earnings, breaks = s,xlim = c(40000,400000),prob = TRUE,las = 1 )



mean0 <- dollar(mean(boston$Earnings))  
sd0 <- dollar(sd(boston$Earnings))  
  
paste("Mean of the Boston Earnings dataset is $", mean0)

## [1] "Mean of the Boston Earnings dataset is $ $108,681"

paste("Standard Deviation from the mean is ", sd0)

## [1] "Standard Deviation from the mean is $50,474.70"

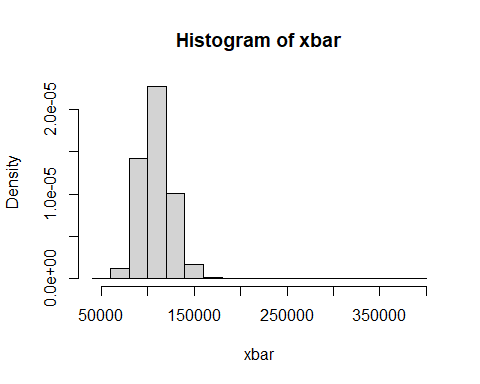
paste("I can infer that the earnings for Boston City public employees is skewed left and with < $100,000 salaries. The high salaries for some jobs are surprising to my high school kids.")

## [1] "I can infer that the earnings for Boston City public employees is skewed left and with < $100,000 salaries. The high salaries for some jobs are surprising to my high school kids."

### 1.b Draw 5000 samples of this data of size 10, show the histogram of the sample means.

Compute the mean of the sample means and the standard deviation of the sample means.

samples <- 5000  
sample.size <- 10  
  
xbar <- numeric(samples)  
  
for (i in 1: samples) {  
 xbar[i] <- mean(sample(boston$Earnings, sample.size, replace = FALSE))  
}  
  
hist(xbar, prob = TRUE,   
 breaks = s, xlim=c(40000,400000))

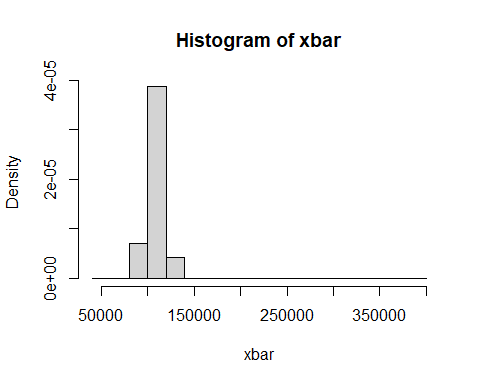


# Computing the mean and standard deviation  
mean10 <- dollar(mean(xbar))  
sd10 <- dollar(sd(xbar))

### 1.c Draw 5000 samples of this data of size 40, show the histogram of the sample means.

Compute the mean of the sample means and the standard deviation of the sample means.

samples <- 5000  
sample.size <- 40  
  
xbar <- numeric(samples)  
  
for (i in 1: samples) {  
 xbar[i] <- mean(sample(boston$Earnings, sample.size, replace = FALSE))  
}  
options(digits = 4)  
hist(xbar, prob = TRUE,   
 breaks = s, xlim=c(40000,400000))



mean40 <- dollar(mean(xbar))  
sd40 <- dollar(sd(xbar))

### 1.d Compare of means and standard deviations of the above three distributions.

paste(" The mean and SD of the original dataset are ", mean0," and ", sd0)

## [1] " The mean and SD of the original dataset are $108,681 and $50,474.70"

paste(" The mean and SD of the 10 sample sized dataset are ", mean10," and ",sd10)

## [1] " The mean and SD of the 10 sample sized dataset are $108,626 and $16,071.44"

paste(" The mean and SD of the 40 sample sized dataset are ", mean40," and ", sd40)

## [1] " The mean and SD of the 40 sample sized dataset are $108,695 and $7,983.06"

paste(" The mean became clear quickly with a small data set and the standard deviation changed quite a bit from the orgiinal. To me this would point to outliers and could be investigated wtih a boxplot.")

## [1] " The mean became clear quickly with a small data set and the standard deviation changed quite a bit from the orgiinal. To me this would point to outliers and could be investigated wtih a boxplot."

#boxplot(boston$Earnings,horizontal = TRUE ))

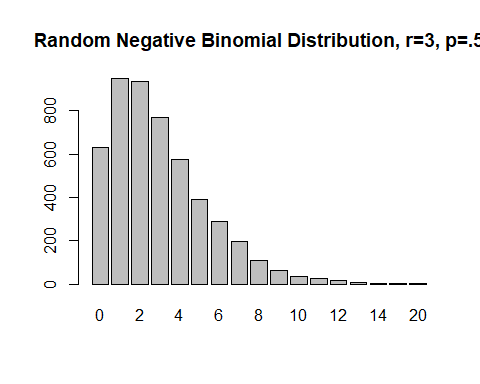
## 2. Central Limit Theorem - Negative Binomial Distribution

Suppose the input data follows the negative binomial distribution with the parameters size = 3 and prob = .5

### 2.a Generate 5000 random values from this distribution. Show the barplot with the

proportions of the distinct values of this distribution.

# Probability  
p <- .5  
# size  
s <- 3  
  
# Test plot   
#x <- 0:30  
#plot(dnbinom(x, size = s, prob = p))  
  
  
x <- rnbinom(5000, size = s, prob = p)  
barplot(table(x))  
title("Random Negative Binomial Distribution, r=3, p=.5")



mean(x)

## [1] 2.998

sd(x)

## [1] 2.449

### 2.b With samples sizes of 10, 20, 30, and 40, draw 1000 samples from the data

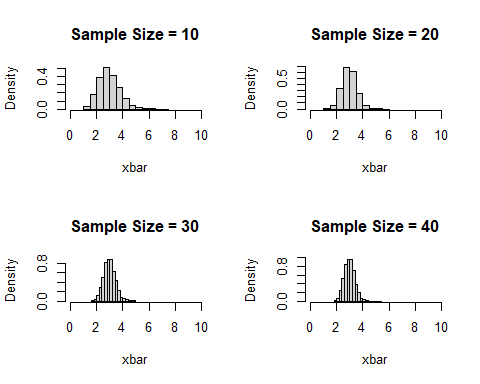
generated in a). Use sample() function with replace as FALSE. Show the histograms of the densities of the sample means. Use a 2 x 2 layout.

par(mfrow = c(2,2))  
p <- .5  
nbiSample <- 5000  
# size  
s <- 3  
x <- rnbinom(nbiSample, size = s, prob = p)  
xbar <- numeric(nbiSample)  
  
# x has been setup in previous code segment  
for (size in c(10, 20, 30, 40)) {  
 for (i in 1:nbiSample) {  
 xbar[i] <- mean(sample(x, size, replace = FALSE))  
 }  
  
# Create a hist for each, arbitrarily kept the breaks at 15  
 hist(xbar, prob = TRUE,   
 breaks = 15, xlim=c(0,10),  
 main = paste("Sample Size =", size))  
   
 cat("Sample Size = ", size, " Mean = ", mean(xbar),  
 " SD = ", sd(xbar), "\n")  
}

## Sample Size = 10 Mean = 2.994 SD = 0.7881

## Sample Size = 20 Mean = 2.99 SD = 0.5403

## Sample Size = 30 Mean = 3 SD = 0.4351



## Sample Size = 40 Mean = 2.995 SD = 0.3887

par(mfrow = c(1,1))  
  
#10 / sqrt(c(10,20,30,40))

### 2.c Compare of means and standard deviations of the data from a) with the four

sequences generated in b).

paste("With the four sample sizes changing, the mean was very accurate and the standard deviation was increasingly less as the sample size grew. Changing the nbiSample var from 5000 to 500 still resulted in values very close when the same sample size of 40 was used. ")

## [1] "With the four sample sizes changing, the mean was very accurate and the standard deviation was increasingly less as the sample size grew. Changing the nbiSample var from 5000 to 500 still resulted in values very close when the same sample size of 40 was used. "

## 3. Sampling

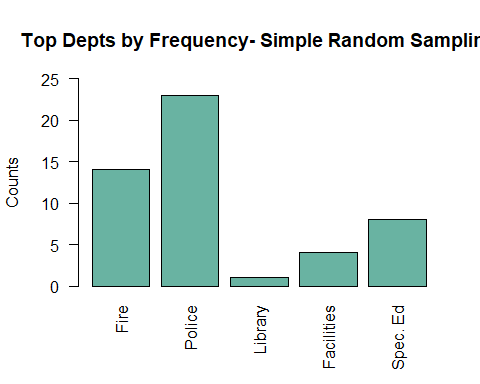
Create a subset of the dataset from #1 with only the top 5 departments based on the number of employees working in that department.  
The top 5 should be computed using R code. Use the %in% operator.  
Sampling size: 50 for each  
Random seed: Last 4 of your BU id : 5121

# Trying the %>% operator  
# Created a sorted var, send it to table for a count, sort it by decreasing values and get the names  
sortedDept <- boston$Department %>% table() %>% sort(decreasing = TRUE) %>% names()  
  
# Get the top 5 names  
top5DepartmentNames <- sortedDept[1:5]  
  
# Get a subset of the top 5 departments in a dataframe using the %in% operator  
topDepartments <- subset(boston,boston$Department %in% top5DepartmentNames)

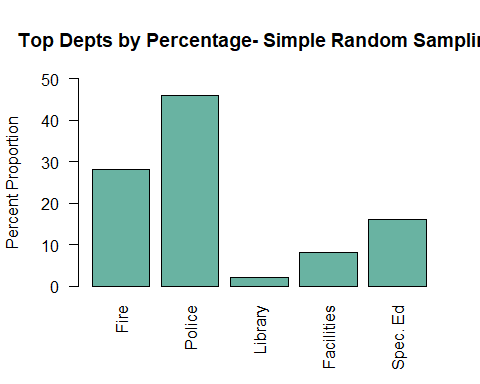
### 3.a Show the sample drawn using simple random sampling without replacement.

Show the frequencies for the selected departments. Show the percentages of these with respect to sample size.

# seed from BU id  
set.seed(5121)  
s <- srswor(50, nrow(topDepartments))  
  
sample.2 <- topDepartments[s != 0, ]  
#head(sample.2)  
  
# Frequency table  
t1 <- table(sample.2$Department)  
#todo: Change naming convention on xaxis  
barplot(t1, main = "Top Depts by Frequency- Simple Random Sampling", ylab = "Counts",  
 ylim=range(pretty(c(0, t1))),  
 las = 2,  
 names.arg=c("Fire","Police","Library","Facilities","Spec. Ed"),   
 col="#69b3a2")



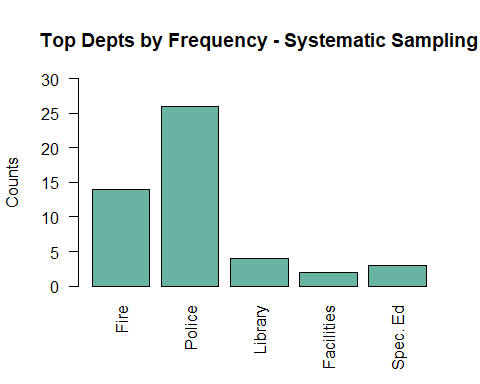
# Proportion Table  
t2<- prop.table(table(sample.2$Department)) \* 100  
#todo: Change naming convention on xaxis  
# Renaming the prop table using the t var from above.   
barplot(t2, main = "Top Depts by Percentage- Simple Random Sampling", ylab = "Percent Proportion",  
 ylim=range(pretty(c(0, t2))),  
 las = 2,  
 names.arg=c("Fire","Police","Library","Facilities","Spec. Ed"),   
 col="#69b3a2")



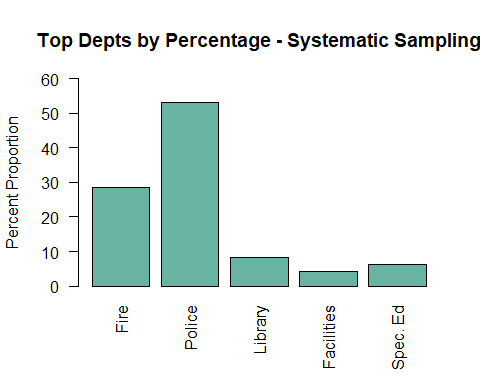
### 3.b Show the sample drawn using systematic sampling.

Show the frequencies for the selected departments. Show the percentages of these with respect to sample size.

set.seed(5121)  
# sample size 50  
N <- nrow(topDepartments)  
n <- 50  
  
k <- ceiling(N / n)  
#k  
  
r <- sample(k, 1)  
#r  
  
# select every kth item  
  
s <- seq(r, by = k, length = n)  
  
sample.3 <- topDepartments[s, ]  
#head(sample.3)  
  
  
# Cut and paste code from above. dangerous..  
# Frequency table  
t3 <- table(sample.3$Department)  
#todo: Change naming convention on xaxis  
barplot(t3, main = "Top Depts by Frequency - Systematic Sampling", ylab = "Counts",  
 ylim=range(pretty(c(0, t3))),  
 las = 2,  
 names.arg=c("Fire","Police","Library","Facilities","Spec. Ed"),   
 col="#69b3a2")



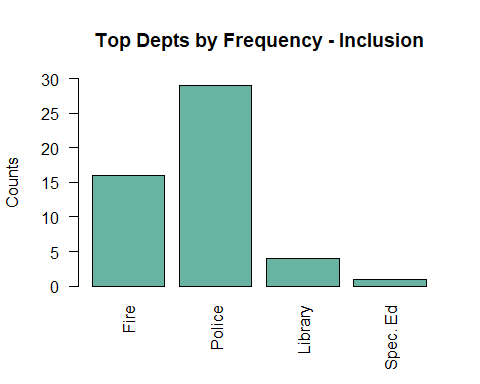
# Proportion Table  
t4<- prop.table(table(sample.3$Department)) \* 100  
#todo: Change naming convention on xaxis  
# Renaming the prop table using the t var from above.   
barplot(t4, main = "Top Depts by Percentage - Systematic Sampling", ylab = "Percent Proportion",  
 ylim=range(pretty(c(0, t4))),  
 las = 2,  
 names.arg=c("Fire","Police","Library","Facilities","Spec. Ed"),   
 col="#69b3a2")



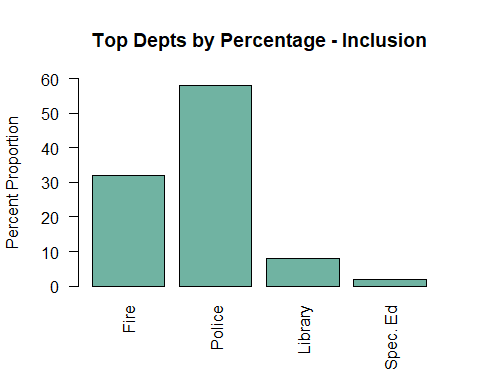
### 3.c Calculate the inclusion probabilities using the Earnings variable.

Using these values, show the sample drawn using systematic sampling with unequal probabilities. Show the frequencies for the selected departments. Show the percentages of these with respect to sample size.

# min i  
#### 3.6. Unequal Probabilities  
#inclusionprobabilities()  
  
set.seed(5121)  
# UPsystematic  
  
pik <- inclusionprobabilities(  
 topDepartments$Earnings, 50)  
#length(pik)  
  
#sum(pik)  
  
s <- UPsystematic(pik)  
  
sample.4 <- topDepartments[s != 0, ]  
#head(sample.4)  
  
  
  
  
# Cut and paste code from above. dangerous..  
# Frequency table  
t5 <- table(sample.4$Department)  
#todo: Change naming convention on xaxis  
barplot(t5, main = "Top Depts by Frequency - Inclusion", ylab = "Counts",  
 ylim=range(pretty(c(0, t5))),  
 las = 2,  
 names.arg=c("Fire","Police","Library","Spec. Ed"),   
 col="#69b3a2")



# Proportion Table  
t6<- prop.table(table(sample.4$Department)) \* 100  
  
#todo: Change naming convention on xaxis  
# Renaming the prop table using the t var from above.   
barplot(t6, main = "Top Depts by Percentage - Inclusion", ylab = "Percent Proportion",  
 ylim=range(pretty(c(0, t6))),  
 las = 2,  
 names.arg=c("Fire","Police","Library","Spec. Ed"),   
 col="#70b3a2")



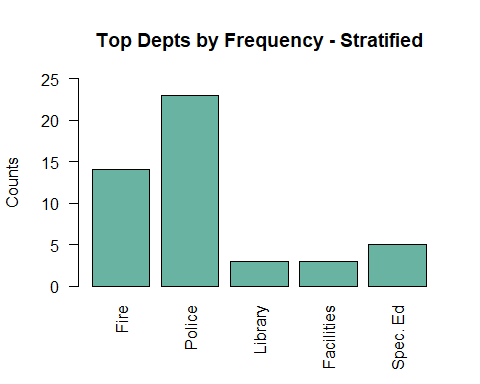
### 3.d Stratified Sampling: Order the data using the Department variable.

Draw a stratified sample using proportional sizes based on the Department variable. Show the frequencies for the selected departments. Show the percentages of these with respect to sample size.

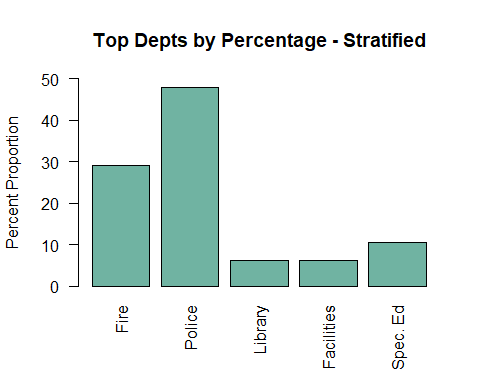
sortedDepts <- sort(table(topDepartments$Department),decreasing = TRUE)  
set.seed(5121)  
# Then do the stratified sampling  
# order and include all columns  
orderedDepts<- topDepartments[order(topDepartments$Department),]  
  
freq <- table(orderedDepts$Department)  
# Samples = 50  
st.sizes <- 50 \* freq / sum(freq)  
#st.sizes  
  
#as.vector(st.sizes)  
  
#as.vector(t(st.sizes))  
  
st.sizes <- as.vector(t(st.sizes))  
st.sizes <- st.sizes[st.sizes != 0]  
  
#st.sizes  
  
st.3 <- strata(orderedDepts,   
 stratanames = c("Department"),  
 size = st.sizes, method = "srswor",  
 description = TRUE)

## Stratum 1   
##   
## Population total and number of selected units: 1672 14.38   
## Stratum 2   
##   
## Population total and number of selected units: 2732 23.5   
## Stratum 3   
##   
## Population total and number of selected units: 384 3.302   
## Stratum 4   
##   
## Population total and number of selected units: 415 3.569   
## Stratum 5   
##   
## Population total and number of selected units: 611 5.255   
## Number of strata 5   
## Total number of selected units 50

#st.3  
  
st.sample3 <- getdata(orderedDepts, st.3)  
  
#st.sample3  
  
freq2 <- table(st.sample3$Department)  
  
  
# Frequency Plot  
t7 <- table(st.sample3$Department)  
#todo: Change naming convention on xaxis  
barplot(t7, main = "Top Depts by Frequency - Stratified", ylab = "Counts",  
 ylim=range(pretty(c(0, t7))),  
 las = 2,  
 names.arg=c("Fire","Police","Library","Facilities","Spec. Ed"),   
 col="#69b3a2")



# Proportion Plot  
t8<- prop.table(table(st.sample3$Department)) \* 100  
barplot(t8, main = "Top Depts by Percentage - Stratified", ylab = "Percent Proportion",  
 ylim=range(pretty(c(0, t8))),  
 las = 2,  
 names.arg=c("Fire","Police","Library","Facilities","Spec. Ed"),   
 col="#70b3a2")



### 3.e Compare the means of Earnings variable for these four samples against the mean for

the data.

# Original Dataset  
paste("The mean for Earning variable in the original dataset is", dollar( mean(boston$Earnings)))

## [1] "The mean for Earning variable in the original dataset is $108,681"

# Simple Random Sampling  
paste("The mean for Earning variable in the Simple Random dataset is", dollar( mean(sample.2$Earnings)))

## [1] "The mean for Earning variable in the Simple Random dataset is $133,382"

# Systematic Sampling  
paste("The mean for Earning variable in the Systematic Random dataset is", dollar( mean(sample.3$Earnings,na.rm=TRUE)))

## [1] "The mean for Earning variable in the Systematic Random dataset is $145,045"

# Inclusive Sampling  
paste("The mean for Earning variable in the Inclusive Random dataset is", dollar( mean(sample.4$Earnings)))

## [1] "The mean for Earning variable in the Inclusive Random dataset is $164,963"

# Stratified Sampling  
paste("The mean for Earning variable in the Stratified Random dataset is", dollar( mean(st.sample3$Earnings)))

## [1] "The mean for Earning variable in the Stratified Random dataset is $131,764"

"It appears that a simple random sample works best for this particular problem, but am questioning why the mean is not closer to the original dataset mean as it worked in out example data sets."

## [1] "It appears that a simple random sample works best for this particular problem, but am questioning why the mean is not closer to the original dataset mean as it worked in out example data sets."