STAT616 - HW #5

Elliot Smith

4/12/2018

Problem 1

Part a

- Sample SD = 2.188418
- $\sqrt{\bar{X}} = 2.325383$
- The Poisson distribution is described by one parameter, λ, which represents both the Mean and Variance (EX and VarX, respectively). So in the case of the Poisson distribution, taking the SD of a sample and taking the square root of the mean will both yield $\sqrt{\lambda}$

Part b

- Standard Error of Sample SD: 1.458007
- Standard Error of $\sqrt{\bar{X}}$: 0.4000194
- I prefer the estimate of $\sqrt{\bar{X}}$ because it is a smaller standard error, thus telling us that the estimate is more precise than the other estimate.

Part c

- Standard Error of Sample SD: 1.228888
- Standard Error of $\sqrt{\bar{X}}$: 0.375584
- I again prefer the estimate of $\sqrt{\bar{X}}$ because it is a smaller standard error, thus telling us that the estimate is more precise than the other estimate.

Part d

The parametric estimate standard errors are larger, implying a larger uncertainty in these estimates. Normally, I would prefer the method that result in the smaller standard error, however, I prefer the parametric method because it is representative of the true distribution and less of a sample-based estimate of the true distribution, and as such, is more close to the true distribution.

Problem 2

Part a

- Correlation Coefficient - avesat-pctsat: 0.1013043
- The correlation coefficient define the degree of relation between these two variables; this value of 0.1 is a weak, positive relationship between avesat and pctsat.
- Boot Strap Standard Error of Sample Correlation Coefficient: 0.06038997

Part b

- Correlation Coefficient - avesat-pctsat: 0.3265814
- The correlation coefficient define the degree of relation between these two variables; this value of 0.33 is a weak-to-moderate, positive relationship between avesat and pctsat.
- Boot Strap Standard Error of Sample Correlation Coefficient: 0.3118617

Part c

- Correlation Coefficient - avesat-pctsat: -0.1192082
- The correlation coefficient define the degree of relation between these two variables; this value of -0.12 is a weak, negative relationship between avesat and pctsat.
- Boot Strap Standard Error of Sample Correlation Coefficient: 0.1337926

Part d

The standard error for the correlation coefficient for the A schools is 0.3118617, while the standard error for the correlation coefficient for the C schools is 0.1337926 This shows that the data is much more spread for A schools as compared to C schools and thus our estimators are less precise for A schools as compared to C schools. The correlation coefficient for the A schools is 0.3265814, while the correlation coefficient for the C schools is -0.1192082. The shows that there is a weak-to-moderate, positive relationship between percent of students taking the SAT and the average score for A schools, while there is almost no relationship (though what is there is a weak, negative relationship) for the C schools.

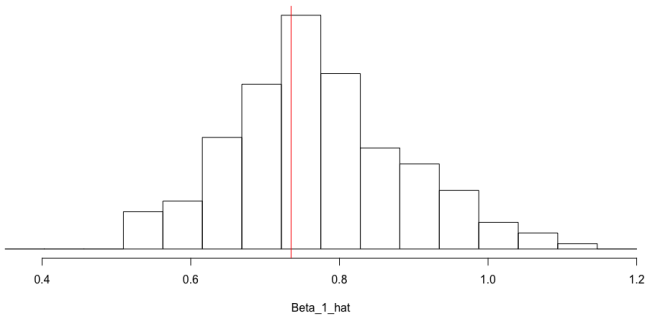
I think the correlation signs are different because in C schools, we may infer that most of the students are not good students, but the good students there will still take the exam. As a higher percent of students are encouraged to take the exam, we can infer that they will not be as good students as those that have already taken the exam at the C schools. At A schools, we may infer that there are a larger pool of strong students, so as a larger percent of students are encouraged to take the test, the average score will go up as more, academically strong students take the exam.

Part e

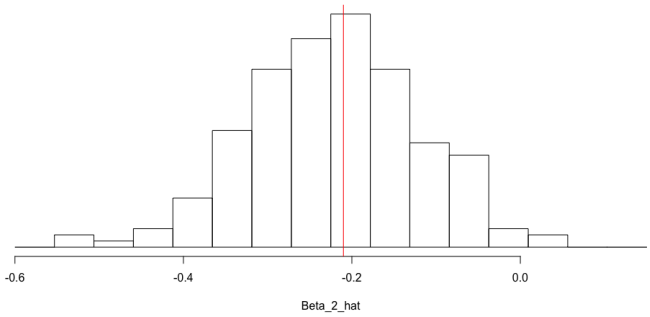
- 95% Confidence Interval - Median of Average Score between A schools and B schools: [52.25497, 52.86223]
- Because zero is not contained in this interval (and that we generated a satisfactory number of 5000 boot strap samples) we may conclude that there is a significant difference between the median, average SAT scores of A schools and B schools.

Problem 3

Plot for $\hat{\beta}_1^*$



Plot for $\hat{\beta}_2^*$



Algorithm for Bootstrap

Once we have generated our Beta estimates (as per the instructions in the Handout) we simply need to perform the simulation (sample the data with replacement) and repeat the prodedure a large number of times (in my case, 500). By performing this operation, we will generate many Beta estimates; we will use these Beta estimates to construct our Histogram and then plot the observed Beta estimates as a vertical line in their respective plots to get a contextual view of our Beta estimates as compared to our simulated results.

Code Appendix

```
########## Problem 1 ##########



## Load in the data
```

```r
data <- c(2, 5, 5, 5, 2, 4, 4, 9, 5, 10, 6, 8, 7, 5, 4, 5, 3, 5, 6, 6, 3, 5, 2, 7, 10, 7, 6)

## Set some data parameters

B <- 500
n <- length(data)

##### Part a #####

sd(data)
sqrt(mean(data))

# Poisson distribution has one parameter, theta, which represents both the EX and VarX



##### Part b #####



## Calculate x_bar

x_bar <- mean(data)

## Containers for statistics

sim_SD_1b <- numeric()
sim_sqrt_1b <- numeric()

## Perform the simulation

for (i in 1:B) {

    temp <- rpois(n = n, lambda = x_bar)
    sim_SD_1b[i] <- sd(temp)
    sim_sqrt_1b[i] <- sqrt(mean(temp))

}

## Harvest the standard errors of the results

se_SD_1b <- sqrt((1/(n - 1))*sum((sim_SD_1b - mean(sim_SD_1b))^2))
se_sqrt_1b <- sqrt((1/(n - 1))*sum((sim_sqrt_1b - mean(sim_sqrt_1b))^2))

# I prefer sqrt(mean(X)) method because it has a smaller standard error and thus is more accurate/preci



##### Part c #####



## Containers for statistics
```

```r
sim_SD_1c <- numeric()
sim_sqrt_1c <- numeric()

## Perform the simulation

for (i in 1:B) {

    temp <- sample(x = data, size = n, replace = TRUE)
    sim_SD_1c[i] <- sd(temp)
    sim_sqrt_1c[i] <- sqrt(mean(temp))

}

## Harvest the standard errors of the results

se_SD_1c <- sqrt((1/(n - 1))*sum((sim_SD_1c - mean(sim_SD_1c))^2))
se_sqrt_1c <- sqrt((1/(n - 1))*sum((sim_sqrt_1c - mean(sim_sqrt_1c))^2))

# I prefer sqrt(mean(X)) method because it has a smaller standard error and thus is more accurate/preci



##### Part d #####



# The parametric SEs are larger, implying a larger spread. Normally, I would prefer the method that
# results in the smaller standard error, however, I prefer the parametric method because it
# is representative of the true distribution, and less of a sample-based estimate of the true distribut
# and as such, is more close to the true distribution.




########## Problem 2 ##########




## Load in the data

data_2 <- read.csv("~/Documents/Rice_University/Spring_2018/STAT616/HW05/ncschools.csv")

## Remove schools where no SATs were taken

data_2_final <- data_2[data_2$pctsat != 0, ]
data_2_final <- data_2_final[data_2_final$category == "H", ]

## Remove data where percent was over 100%

data_2_final <- data_2_final[data_2_final$pct <= 1, ]
```

```r
## Set some data parameters

B <- 500
n <- dim(data_2_final)[1]



##### Part a #####



## Correlation between pctsat and avesat

cor(data_2_final$avesat, data_2_final$pctsat)

# There is approximately no relationship between pcstat and avesat, only 0.12

## Create a containers for the results

output_2a <- setNames(data.frame(matrix(NA, ncol = 2, nrow = n)), c("avesat", "pctsat"))
sim_corr_2a <- numeric()

## Perform the simulation

for (i in 1:B) {

    rand_inds <- sample(x = 1:n, size = n, replace = TRUE)

    for (j in 1:length(rand_inds)) {

        temp <- data_2_final[rand_inds[j], 8:9]
        output_2a[j, ] <- temp

    }

    sim_corr_2a[i] <- cor(output_2a$avesat, output_2a$pctsat)

}

## Harvest the standard errors of the results

se_cc_2a <- sqrt((1/(n - 1))*sum((sim_corr_2a - mean(sim_corr_2a))^2))



##### Part b #####



## Update the dataset

data_2b <- data_2_final[data_2_final$grade == "A", ]
```

```r
## Set some data parameters

B <- 500
n <- dim(data_2b)[1]

## Correlation between pctsat and avesat

cor(data_2b$avesat, data_2b$pctsat)

# There is a marginally positive relationship between pcstat and avesat of 0.29

## Create a containers for the results

output_2b <- setNames(data.frame(matrix(NA, ncol = 2, nrow = n)), c("avesat", "pctsat"))
sim_corr_2b <- numeric()

## Perform the simulation

for (i in 1:B) {

    rand_inds <- sample(x = 1:n, size = n, replace = TRUE)

    for (j in 1:length(rand_inds)) {

        temp <- data_2b[rand_inds[j], 8:9]
        output_2b[j, ] <- temp

    }

    sim_corr_2b[i] <- cor(output_2b$avesat, output_2b$pctsat)

}

## Harvest the standard errors of the results

se_cc_2b <- sqrt((1/(n - 1))*sum((sim_corr_2b - mean(sim_corr_2b))^2))



##### Part c #####



## Update the dataset

data_2c <- data_2_final[data_2_final$grade == "C", ]

## Set some data parameters

B <- 500
n <- dim(data_2c)[1]

## Correlation between pctsat and avesat
```

```r
cor(data_2c$avesat, data_2c$pctsat)

# There is a marginally negative relationship between pcstat and avesat of -0.12

## Create a containers for the results

output_2c <- setNames(data.frame(matrix(NA, ncol = 2, nrow = n)), c("avesat", "pctsat"))
sim_corr_2c <- numeric()

## Perform the simulation

for (i in 1:B) {

    rand_inds <- sample(x = 1:n, size = n, replace = TRUE)

    for (j in 1:length(rand_inds)) {

        temp <- data_2c[rand_inds[j], 8:9]
        output_2c[j, ] <- temp

    }

    sim_corr_2c[i] <- cor(output_2c$avesat, output_2c$pctsat)

}

## Harvest the standard errors of the results

se_cc_2c <- sqrt((1/(n - 1))*sum((sim_corr_2c - mean(sim_corr_2c))^2))



##### Part d #####



# The standard error for the correlation coefficient for the A schools is 0.3069, while the
# standard error for the correlation coefficient for the B schools is 0.1336. This shows that
# the data is much more spread for A schools as compared to C schools. The correlation
# coefficient for the A schools is 0.2923, while the correlation coefficient for the C schools
# is -0.1192. The shows that there is a marginally positive relationship between percent of
# students taking the SAT and the average score, while there is almost no relationship (though
# what is there is marginally negative) for the C schools.

# I think the correlation signs are different because in C schools, we may infer that most
# of the students are not good students, but the good students there will still take the exam.
# As a higher percent of students are encouraged to take the exam, we can infer that they will
# not be as good students as those that have already taken the exam at poor schools. At A schools,
# we may infer that there are a larger pool of strong students, so as schools have more test-takers,
# we can hypothesize that they will be stronger students.
```

```
##### Part e #####



## Set some data parameters

B <- 5000

## Extract the B schools

data_2e <- data_2_final[data_2_final$grade == "B", ]

## Containers for difference in medians

sim_med_diff_2e <- numeric()

## Perform the simulation for the A and C schools

for (i in 1:B) {

    temp_a <- median(sample(x = data_2b$avesat, size = length(data_2b$avesat), replace = TRUE))
    temp_b <- median(sample(x = data_2e$avesat, size = length(data_2e$avesat), replace = TRUE))

    sim_med_diff_2e[i] <- temp_a - temp_b

}

## Build the confidence intervals

t.test(sim_med_diff_2e)

# As we can see from the confidence intervals supplied by t.test, there is a significant difference
# in the medians between A schools and B schools.
```