6611_HW1

Weixuan Liu

August 28, 2019

Problem 1

a

Ex. 1a

Normal($\mu = 125, \sigma = 8$)

```
set.seed(6611)
normal <- rnorm(n = 10000, mean = 125, sd = 8)
```

Poisson($\lambda = 1.5$)

```
poisson <- rpois(n=10000, lambda = 1.5)
```

Binomial($n = 5, p = 0.15$)

```
binomial <- rbinom(n = 10000, size = 5, prob = 0.15)
```

Ex. 1b

(1) For normal distribution, we know that $\mu = \mu = 125$, and (2) for poisson distribution, we know that $\mu = \lambda = 1.5$, (3) for binomial distribution, we know that $\mu = n \cdot p = 5 \cdot 0.15 = 0.75$. To verify our result, we will be running the following codes:

```
mean(normal)
```

```
## [1] 125.0299
```

```
mean(poisson)
```

```
## [1] 1.4881
```

```
mean(binomial)
```

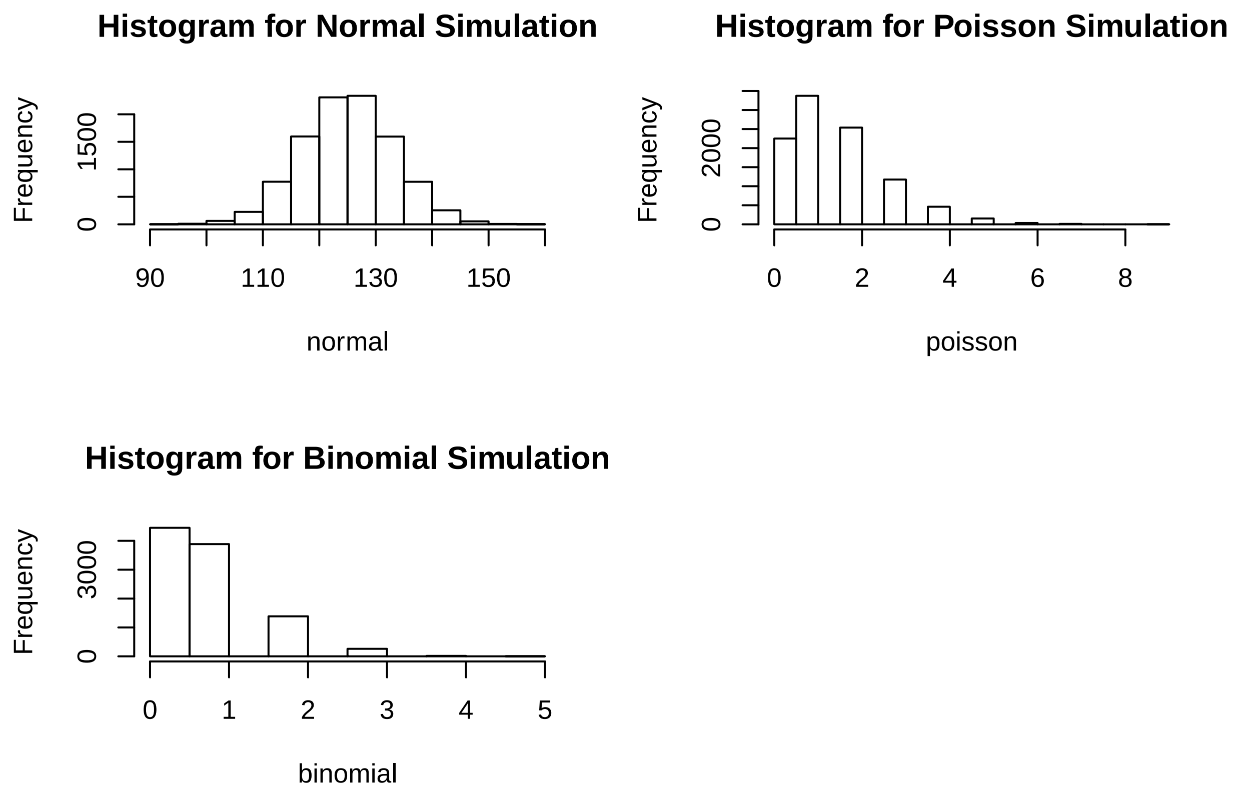
```
## [1] 0.7504
```

The result is not significantly deviated from the theoretical mean, therefore, we know that the theoretical mean is correct.

Ex. 1c

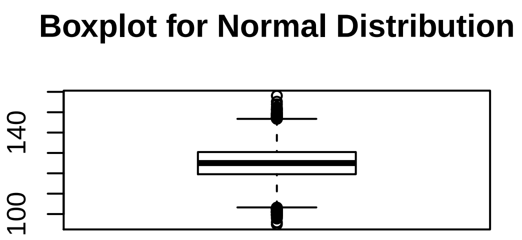
We first show the histogram for these 3 distributions

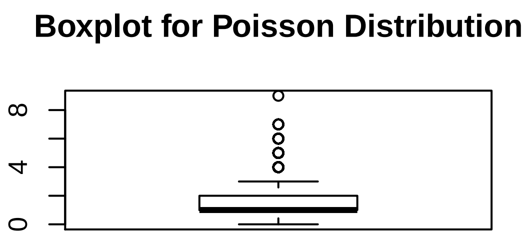
```
par(mfrow =c(2,2))
hist(normal, main = "Histogram for Normal Simulation")
hist(poisson, main = "Histogram for Poisson Simulation")
hist(binomial, main = "Histogram for Binomial Simulation")
```

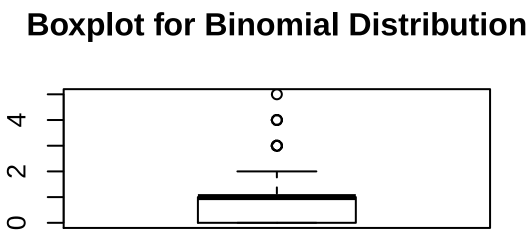


The following is the boxplot for these 3 distributions

```r
par(mfrow =c(2,2))
boxplot(normal ,main = "Boxplot for Normal Distribution")
boxplot(poisson ,main = "Boxplot for Poisson Distribution")
boxplot(binomial ,main = "Boxplot for Binomial Distribution")
```

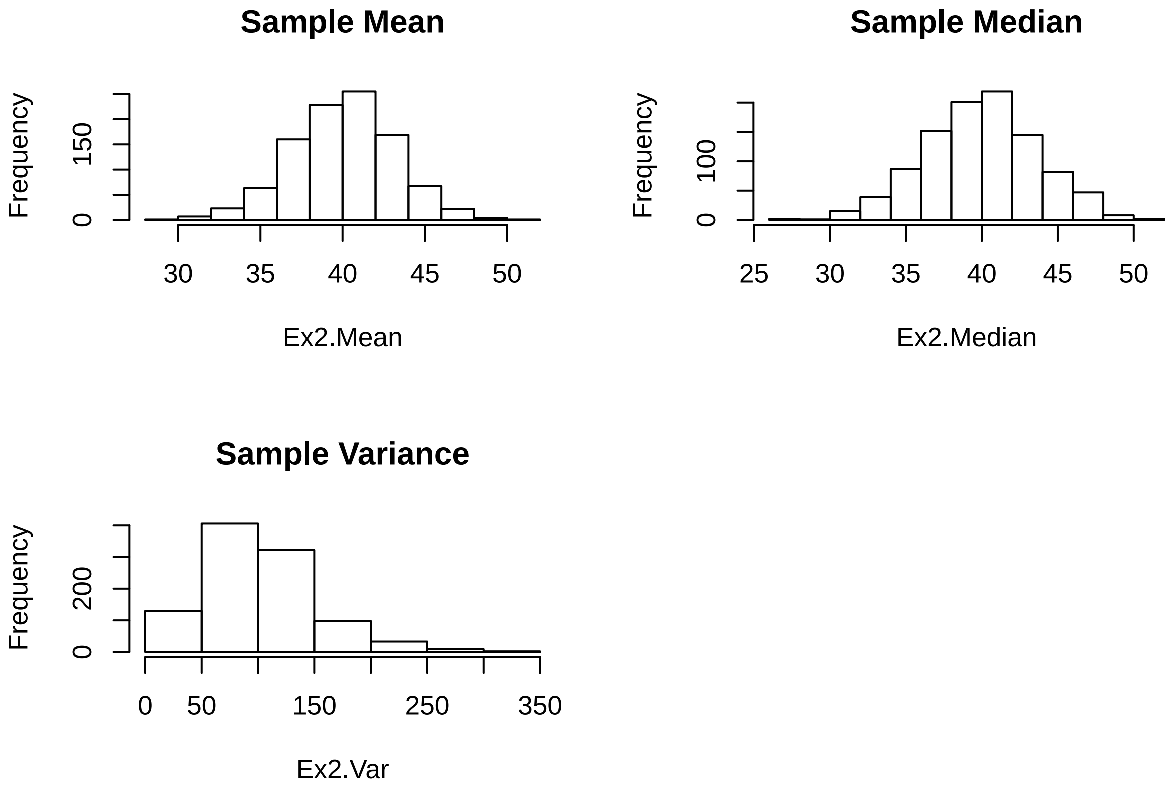






Ex. 2a

```r
set.seed(6611)
Ex2.Mean <- rep(NA, 1000)
Ex2.Median <- rep(NA, 1000)
Ex2.Var <- rep(NA, 1000)
###### Normal Sample Mean

for(i in 1:1000)
{

 new.samp <- rnorm(n = 10, mean = 40, sd = 10 )

 ###### Normal Sample Mean

 temp <- mean(new.samp)
 Ex2.Mean[i] <- temp

###### Normal Sample Median

 temp1 <- median(new.samp)
 Ex2.Median[i] <- temp1
###### Normal Sample Variance

 temp2 <- var(new.samp)
 Ex2.Var[i] <- temp2
}

par(mfrow =c(2,2))
hist(Ex2.Mean, main = "Sample Mean")
hist(Ex2.Median, main = "Sample Median")
hist(Ex2.Var, main = "Sample Variance")
```

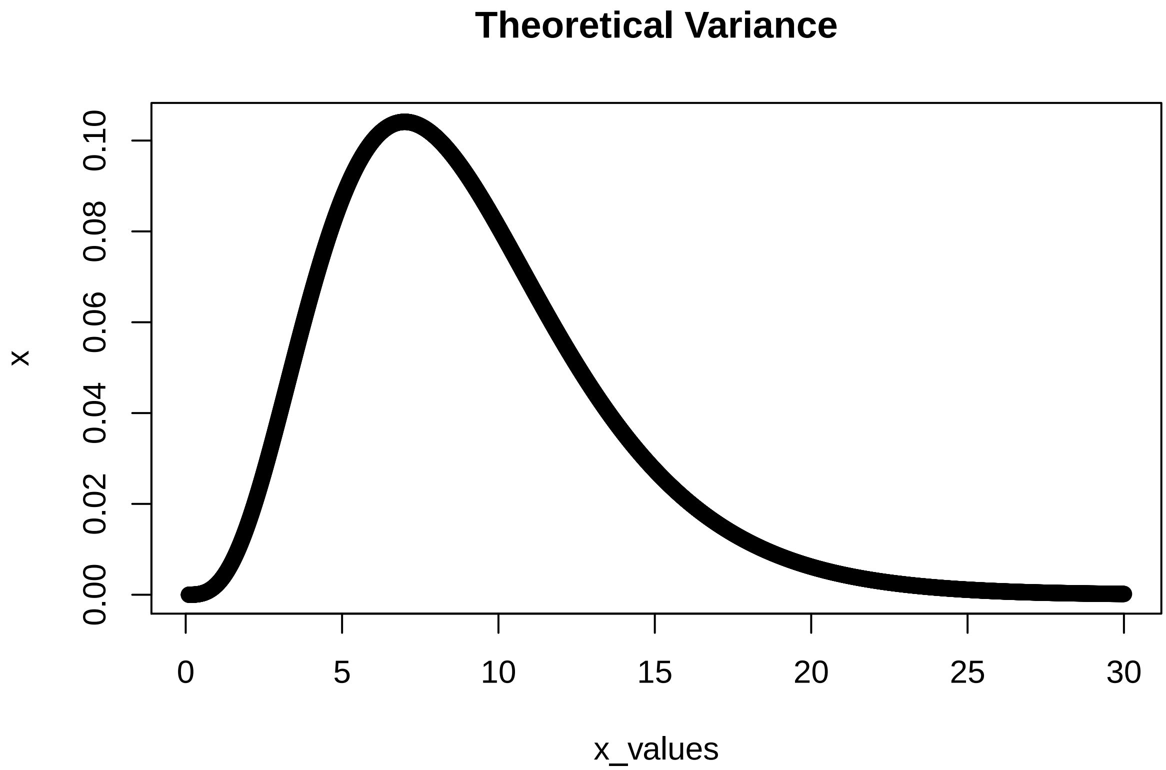


Ex. 2b

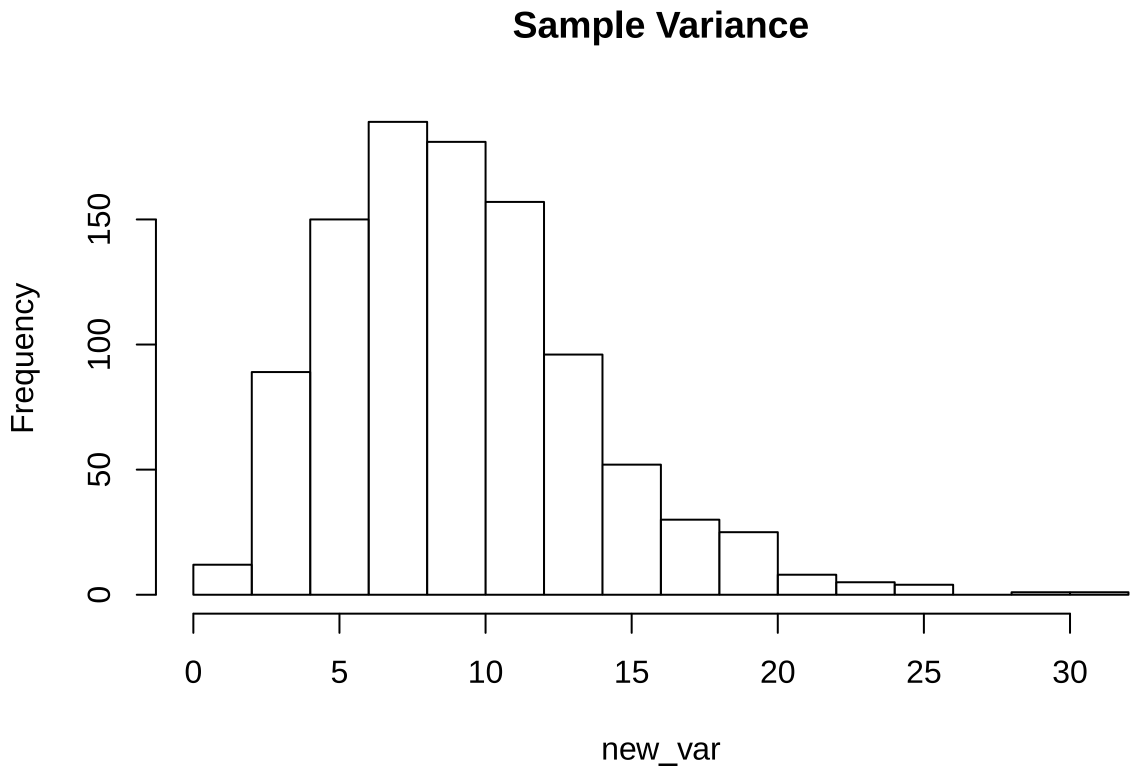
Based on theory, they are all normally distributed, for mean, the parameter values should be normal($\mu = \bar{X}_{10}, \sigma^2 = \frac{\sigma^2}{10}$). For median, $\mu =$ median and $\sigma^2 = 1.253^2 \frac{\sigma^2}{10}$.

Ex. 2c

```
x_values <- seq(0.1, 30, 0.01)
x <- dchisq(x_values,df = 9)
plot(x_values,x, main = "Theoretical Variance")
```



```
new_var <- Ex2.Var * (9/100)
hist(new_var, main = "Sample Variance")
```



Comparing these two plots, we see that sample and theoretical variance are extremely similar.

Ex. 3a

```
set.seed(6611)
height <- rnorm(100, mean = 70, sd = sqrt(15))
median(height)
```

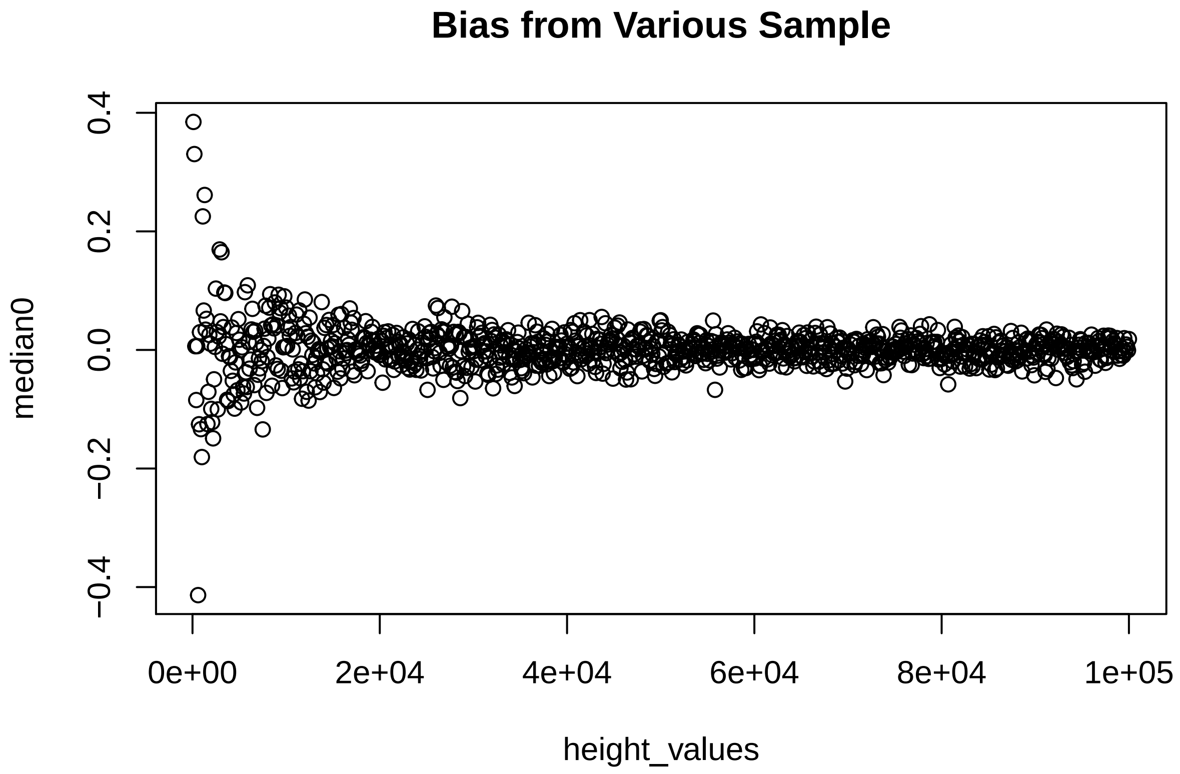
```
## [1] 69.93528
```

```
bias <- median(height)- 70
```

Therefore, we see that the bias of median estimate with respect to population mean is 0.43286 in the simulated case.

Ex. 3b

```
height_values <- seq(100, 100000, 100)
median0 <- rep(NA, length(height_values))
for (i in 1:length(height_values))
{
  median0[i] <- median(rnorm(height_values[i], mean = 70, sd = sqrt(15)))
  median0[i] <- median0[i] - 70
}
plot(height_values, median0, main = "Bias from Various Sample")
```



We see from the plot that as the sample size increase, the bias for median with respect to population mean is closer and closer to 0. Which means, for general sense, if the underlying distribution of height is normal, then we see that the median estimator for population is unbiased with respect to the population mean.

Ex. 3c

```
height_var <- rep(NA, length(height_values))
vari <- function(x)
{
  temp <- rep(NA, length(x))
  for (i in 1: length(x))
  {
    temp[i] <-(x[i]- median(x))^2
  }
  temp1 <- sum(temp)/length(x)
  return(temp1)
}
for (i in 1:length(height_values))
{
  height_sample <- rnorm(100*i,mean = 70, sd = sqrt(15))
  height_var[i] <- vari(height_sample)
}
```

From the previous question, we know that as sample size increase, the median estimator is closer to the mean estimator. Therefore, from the formula of variance with respect to median, we are able to conclude that as sample size increase, this variance is closer to n times the variance with respect to mean estimator.

Ex. 3d

```r
mean_val <- rep(NA, 10000)
median_val <- rep(NA, 10000)
for (i in 1:10000)
{
  temp <- rnorm(1000, mean = 70, sd = sqrt(15))
  mean_val[i] <- mean(temp)
  median_val[i] <- median(temp)
}
var(mean_val)
```

```
## [1] 0.01531306
```

```r
var(median_val)
```

```
## [1] 0.02383408
```

We see from this simulation example that the variance for the mean estimator is smaller than the variance for the median estimator. Therefore, since the simulation size is large enough, we have enough evidence to conclude that the mean estimator is more efficient.

Ex. 3e

Generally, if we have a statistics $T = r(X)$(sufficient statistics), and its associated expectation is denoted by $E_\theta(T)$, where θ is the variance of the distribution we specify. Then the variance of this statistics with respect to θ satisfies:

$$Var_\theta(T) \geq \frac{[E'_\theta(T)]^2}{nI(\theta)}$$

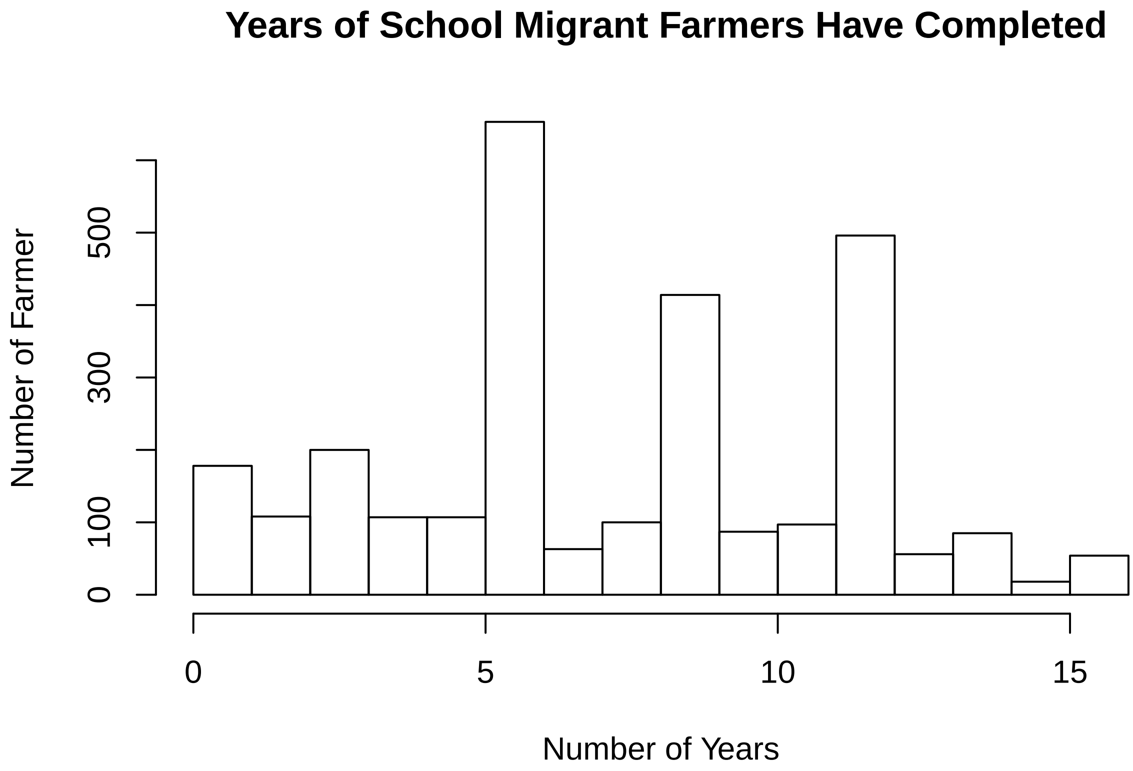
Where $I(\theta)$ is the fisher information (expectation of the second derivative of the log-likelihood, or the first derivative of the score function). Then from this result we see that if we have an Estimator T, then its variance must safisfy this inequality, when equality holds, we can say that this estimator is the most efficient one. If we only restrict our attention to the unbiased estimators, then by definition, we see that $E_\theta(T) = \theta$, and the first derivative of it is 1. Therefore, the inequality becomes:

$$Var_\theta(T) \geq \frac{1}{nI(\theta)}$$

Then if the variance of an estimator is exactly equal to the lower bound, we can say this estimator is the most efficient one. Referring back to this example, we see that both mean and median are the unbiased estimators for the population mean but with different efficiency.

Ex. 5a,b

```r
NAWS <- read.csv("C:/BIOS 6611/NAWS2014.csv")
hist(NAWS$A09,xlab = "Number of Years", ylab = "Number of Farmer"
     , main = " Years of School Migrant Farmers Have Completed" )
```



Ex. 5c

Only reporting the average doesn't tell the whole story, the reason is that we have to obtain more information about how this is distributed. For instance, in this example, we don't know without plotting that the frequency near 8 is relatively low.

Ex. 5d

```
NAWS$category_edu <- "00 - 05"
NAWS[NAWS$A09 >= 6 & NAWS$A09 <= 8, ]$category_edu <- "06 - 08"
NAWS[NAWS$A09 >= 9 & NAWS$A09 <= 11, ]$category_edu <- "09 - 11"
NAWS[NAWS$A09 >= 12, ]$category_edu <- "12+"
```

Ex. 5e

```
table(NAWS$category_edu)
```

```
##
## 00 - 05 06 - 08 09 - 11     12+
##     700     816     598     709
```

```
table(NAWS$category_edu)/length(NAWS$category_edu)
```

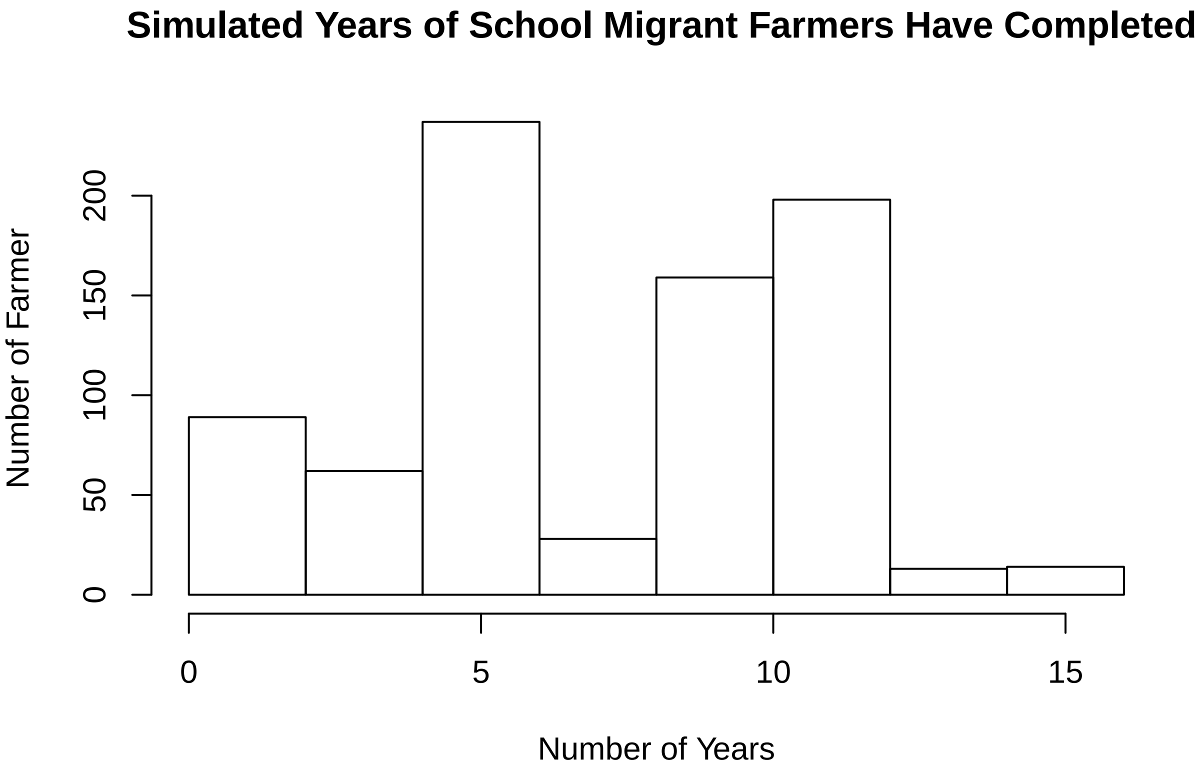
```
##
##   00 - 05   06 - 08   09 - 11       12+
## 0.2479632 0.2890542 0.2118314 0.2511513
```

Ex. 5f

```r
set.seed(6611)
mockdata <- data.frame(subject = 1:800, random = runif(n = 800))
mockdata$educ_cat <- "00 - 05"
mockdata[mockdata$random >= 0.249
         & mockdata$random <= 0.248+0.289, ]$educ_cat <- "06 - 08"
mockdata[mockdata$random >= 0.248+0.290
         & mockdata$random <= 0.248+0.289+0.212, ]$educ_cat <- "09 - 11"
mockdata[mockdata$random >= 0.248+0.289 + 0.213
         & mockdata$random <= 0.248+0.289+0.212+0.251, ]$educ_cat <- "12+"
mockdata$educ_years <- 0
mockdata[mockdata$educ_cat == "00 - 05", ]$educ_years <- runif(n = 197, min = 0, max = 6)
mockdata[mockdata$educ_cat == "06 - 08", ]$educ_years <- runif(n = 219, min = 6, max = 9)
mockdata[mockdata$educ_cat == "09 - 11", ]$educ_years <- runif(n = 171, min = 9, max = 12)
mockdata[mockdata$educ_cat == "12+", ]$educ_years <- runif(n = 213, min = 12, max = 17)
mockdata$edu_stop_yn <- rbinom(n = 800, size = 1, prob = 0.8)
mockdata[mockdata$educ_cat == "06 - 08" & mockdata$edu_stop_yn == 1,]$educ_years <- 6
mockdata[mockdata$educ_cat == "09 - 11" & mockdata$edu_stop_yn == 1,]$educ_years <- 9
mockdata[mockdata$educ_cat == "12+" & mockdata$edu_stop_yn == 1,]$educ_years <- 12
mockdata$educ_years <- as.integer(mockdata$educ_years)
```

Ex. 5g

```r
hist(mockdata$educ_years,xlab = "Number of Years", ylab = "Number of Farmer"
     , main = " Simulated Years of School Migrant Farmers Have Completed" )
```



From the resulting distribution demonstrating in the plot, we see that they looks very similar to the one it based on.

Problem 2

In the paper, the *Inductive Inference* means that the researchers will decide which hypothesis is better based on the observed result (what they see). The advantage of this method is to enlarge the scope of the knowledge and gnerate new hypothesis, but it cannot reflect or conclude the truth that is coming from the nature. "Deductive Inference" means someone are able to base on the fact that the hypothesis is true to anticipate what they would be observing. The advantage of this is that if we have knowledge about the hypothesis, and the hypothesis is true, then there is no way that the prediction based on it is false, which also means the result is certain and objective. However, it also constraint the scope of the knowledge only to the hypothesis, and it is not possible to generate new hypothesis like what *Inductive Inference* did.

The *P-value Fallacy* originated from the conflict between Long-run perspective and Short-run perspective. The Long-run perspective is a *Deductive Inference*, but Short-run is *Inductive Inference.* The author also gives an example to illustrate this. A same problem can be done with the same data but different methods to reach different p-value. Long-run and Short-run will typically generate different results, but if interpreting them using the same number, it will causes the problem of that *P-value Fallacy.* To illustrate this idea, the author also gives explanation from different angle, in a particular research, if subjects involved are interchangeable or identifiable will impact the final results, and in fact, they can be going to completely different directions. This also states that based on what is observed to determine the hypothesis's plausibility is not the same as Long-run experiment becuase it requires only this one single experiment, and the knowledge of those are based on this one single experiment. In general, if 2 researchers view a same problem from different perspectives, then they would have different repeated results in mind, which causes the difference, and also, this informs us of the true meaning of p-value is not evidence-based, it should gives a determination of reject the null hypothesis or not, but not choosing the hypothesis.

Different from p-value, *Confidence Interval* takes the “size of observed effect” into account. Like what author mentioned in the paper, size effect can impact the p-value a lot, small effect with large sample size can have the equivalent outcome as the large effect with small sample size. It can shift our attention from the p-value to the “range of effect” that is defined by the *Confidence Interval.* However, compared with p-values, although the direction is correct, it still shares some common problems with p-value. For instance, *Confidence Interval* doesn’t have the power to connect evidence outside with evidence from one single experiment. Furthermore, as mentioned in the second paragraph, a same experiment data would produced different results simply because of the methodology preference from different researchers, *Confidence Interval* doesn’t solve that problem neither.