Problem Set 5

Daniel Shapiro

9/29/2022

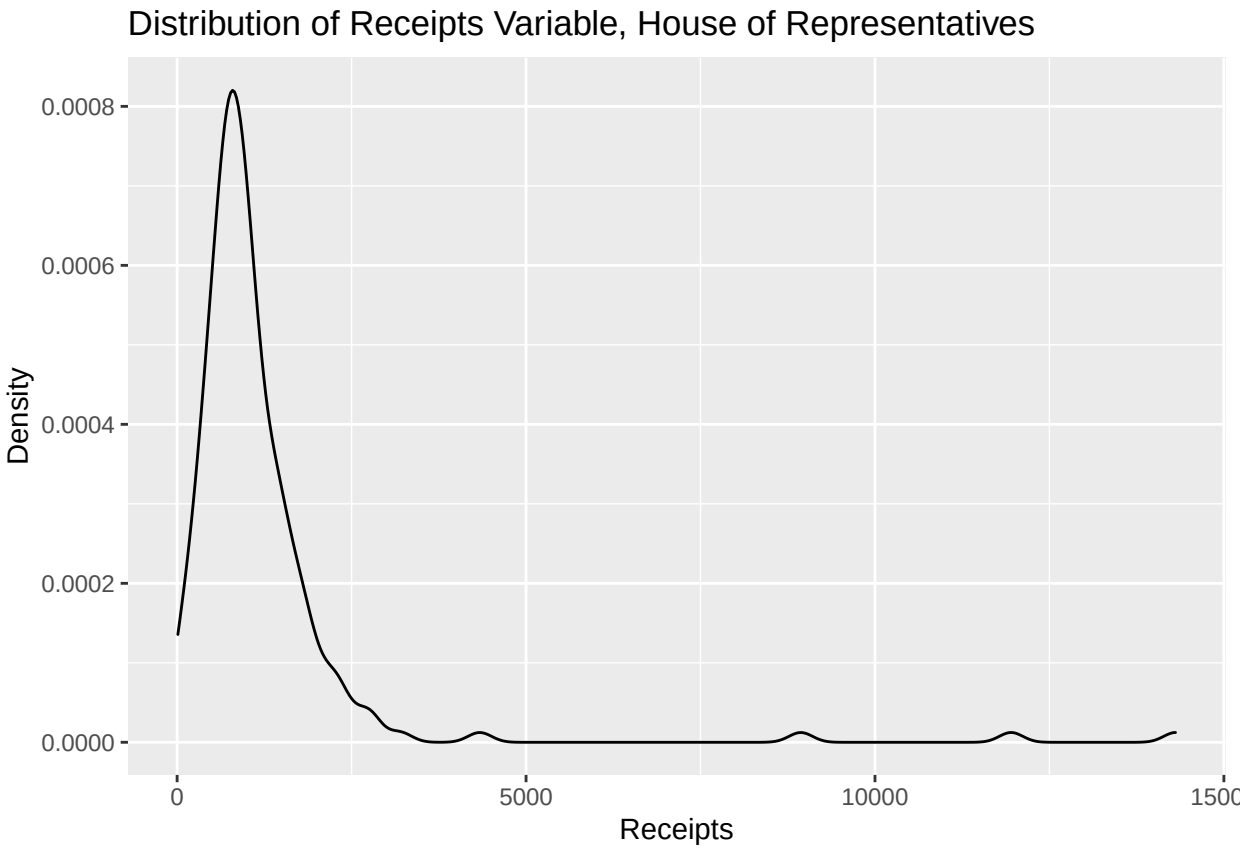
Question 1 Background:

For this problem we will use data on funds raised by a subset of incumbent Democrats in the US House of Representatives for the 2006 election cycle. Download the data, ***house.csv****. In the dataset, you will find the name of the representatives (****house$candidate****) and the funds they raised in thousands of dollars (****house$receipts****).*

```
house <- read.csv("house.csv")
```

1a) First, use `ggplot()` and `geom_density()` to plot the density of `house$receipts` and briefly describe the distribution of the receipts variable in the House of Representatives dataset.

```
ggplot(house, aes(x = receipts)) +
  geom_density() +
  labs(title = "Distribution of Receipts Variable, House of Representatives",
       x = "Receipts",
       y = "Density")
```



The distribution of the house$receipts variable is located for the most part between 0 and 2500; however, there are a few values that occur later on – one just below 5000, another around 8500, another around 12000, and another around 14000.

1b) Find the mean and standard deviation of `house$receipts`, and the standard error of the mean. Write your own function for the standard deviation – do not used a canned `R` function.

```
mean(house$receipts)
```

```
## [1] 1190.31
```

```
deviation <- function(data){

  # Define the mean. I put this with an extra letter just in case it gets confused
  # with a function.

  meann <- mean(data)

  # Add an empty vector to do a for loop over

  vector <- c(rep(0, times = length(data)))

  for(i in 1:length(data)){
    vector[i] <- (data[i] - meann)^2
```

```
  }

  summ <- sum(vector)
  variance <- summ/(length(vector) - 1)
  deviation <- sqrt(variance)

deviation
}

deviation(house$receipts)
```

```
## [1] 1500.763
```

The mean is listed as 1190.3095183 and the standard deviation is 1500.7633085.

```
se <- deviation(house$receipts)/sqrt(length(house$receipts))
se
```

```
## [1] 108.5914
```

This 108.5914219 is the standard error of the mean.

1c) Find 95% confidence intervals for your estimate of the mean. Interpret this CI (in words).

The function to find the confidence interval can be expressed as: $\overline{X} \pm t_{\alpha/2,N-1} S_{\overline{x}}$

We already have $\overline{X}$; that's just the mean of the string. We also already have $S_{\overline{x}}$; that's just the standard error. So now we need to get α and the t score. Then we can plug everything in.

```
# We make this equal to .05 because we are looking for the 95% CI. Will
# have to divide by two later. All of this is just setup.

alpha <- .05

degrees_freedom <- length(house$receipts) - 1

# Here, we use the quantile function from the stats package. Honestly, I've never
# been super sure how the quantile function works. But I know I've used it
# in R before for this sort of thing, so I'm doing it again here.

ts <- qt(p = alpha/2, df = degrees_freedom, lower.tail = FALSE)

# Regardless, now we have the T score. I set lower.tail to FALSE because otherwise
# it came out flipped.

margin <- ts * se

lower <- mean(house$receipts) - margin
upper <- mean(house$receipts) + margin

c(lower, upper)
```

```
## [1]  976.1099 1404.5092
```

These two values, pictured above, are the bounds of our 95% confidence interval for the string house$receipts. This means that we can be 95% confident that the mean candidate receipt value lies between 976.1099 and 1404.5092.

1d) Confirm, through simulation, that if you redraw 1,000 samples from your data of the size of your data (with replacement), about 95% of them fall into your confidence interval.

```
# A favorite of mine -- rep_sample_n from the infer package.

sample <- house %>% infer::rep_sample_n(size = 191, replace = TRUE, reps = 1000) %>%
  select(-candidate)

# Map the mean() to the data. I didn't need the candidate column either.

samplemean <- sample %>%
  group_by(replicate) %>%
  nest() %>%
  mutate(mean = map(.x = data, .f = ~mean(.x$receipts, na.rm = TRUE))) %>%
  select(-data) %>%
  ungroup()

testdata <- samplemean %>%
  filter(mean > 976.1099) %>%
  filter(mean < 1404.5092)

nrow(testdata) / nrow(samplemean)
```

```
## [1] 0.955
```

Confirmed!

Question 2 Background:

Now, we will use the house data to examine how well different estimators recover the true population mean of funds raised. We will now assume that our sample represents the full population (e.g., the mean found above is the true μ). Our parade of estimators will involve drawing 1,000 random samples (without replacement) of size n from the population. Let $X = \{X_1, \ldots, X_N\}$ denote the vector of sample observations, so X_i denotes the i-th observation drawn in each sample with $i = 1, \ldots, N$. We consider the following estimators:

- The Sample Mean: $\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} X_i$
- The Sample Median: If your sample is ordered such that $X_1 \leq X_2 \leq X_3 ... \leq X_n$, the median is the midpoint value.
- The Deletion Estimator I: Order the sample observations such that $X_1 \leq X_2 \leq X_3 ... \leq X_n$. Discard the minimum value $min(X) = X_1$ from the sample, and compute the mean of the remaining observations.
- The Deletion Estimator II: Order the sample observations such that $X_1 \leq X_2 \leq X_3 ... \leq X_n$. Discard the two lowers values X_1, X_2 and the two highest values (X_{n-1}, X_n) from the sample, and compute the mean of the remaining observations.

2a) Write a function that takes data and returns a dataframe of values for the mean, median, and two deletion estimators (eg, it should take data and return a dataframe with 4 rows and 2 columns). Hint: start your function by defining a dataframe with an "estimator" and "result" column. Use `sort()` to sort your data. Make sure your function works by evaluating the following toy sample: $X = 2, 1, 100, 3, 4, 3, 2$. Your function should return the following:

Estimator	Sample Mean	Sample Median	Deletion Estimator I	Deletion Estimator II
$\hat{\mu}$	16.429	3	19	2.667

```
tablefunc <- function(data){

  estimator <- c("Sample Mean", "Sample Median", "Deletion Estimator I", "Deletion Estimator II")

# Empty vector time.

  result <- c(rep(0), times = 4)

  deletion <- data %>%
    sort(decreasing = FALSE)
  deletion1 <- deletion[-1] %>%
    mean()

# A bit round-about, but it does the job.

  deletionintermediate <- deletion[-c(1, 2)]

  deletionlast <- deletionintermediate %>%
    sort(decreasing = TRUE)

  deletion2 <- deletionlast[-c(1, 2)] %>%
    mean()

  for(i in 1:4){
    if(estimator[i] == "Sample Mean"){result[i] <- sum(data)/length(data)
    }
    else if(estimator[i] == "Sample Median"){result[i] <- median(data)
    }
    else if(estimator[i] == "Deletion Estimator I"){result[i] <- deletion1}
    else(result[i] <- deletion2)

  }
  data.frame(estimator, result)
  }

vector <- as.numeric(c(2, 1, 100, 3, 4, 3, 2))
tablefunc(vector)
```

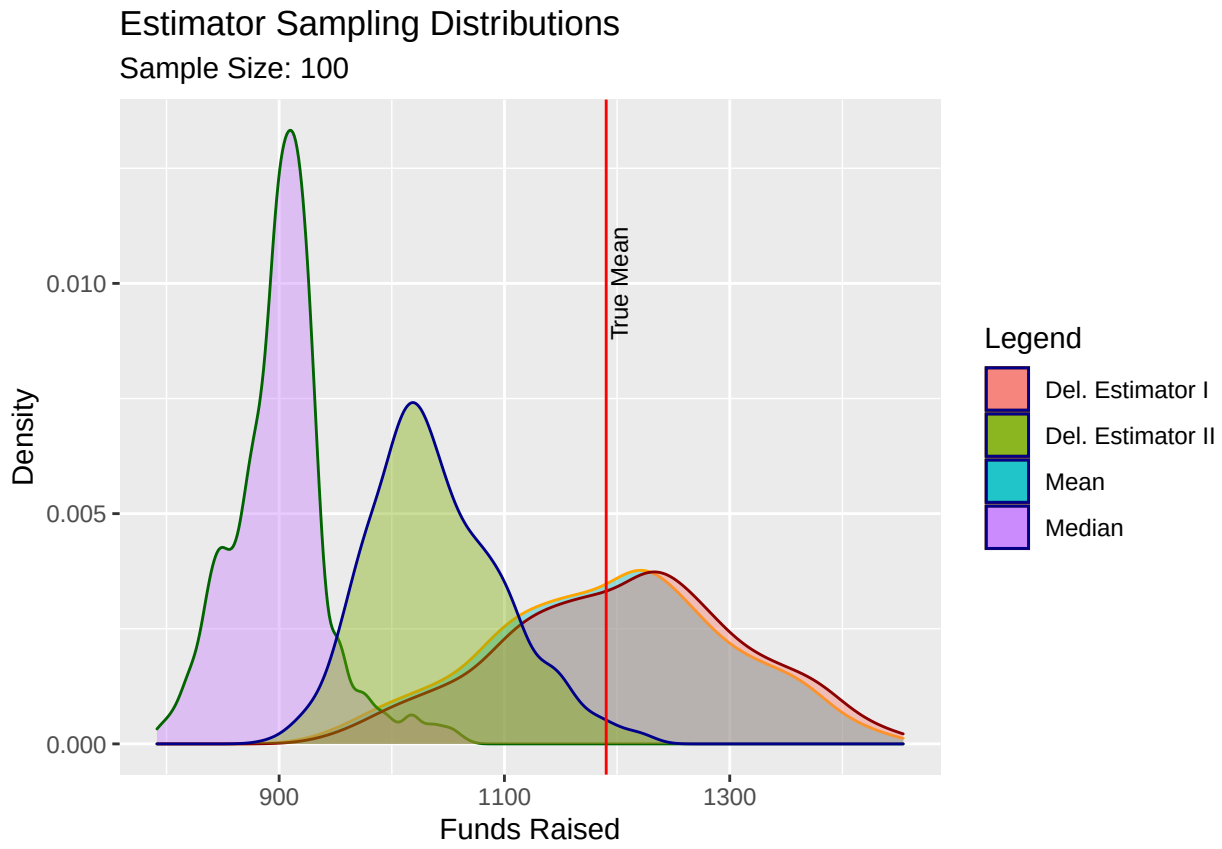
```
##                estimator    result
## 1           Sample Mean 16.428571
## 2         Sample Median  3.000000
## 3  Deletion Estimator I 19.000000
## 4 Deletion Estimator II  2.666667
```

2b) Using the function you defined above, simulate the sampling distribution for all the estimators from part a) for the receipts variable. Your sample size in this case should be $n = 100$, and you should conduct 1,000 simulations (hint: create an empty dataframe, then use `rbind()` to append the results of each iteration of your simulation). Create a single density plot using `ggplot()` and `geom_density()`, including a vertical line at the true population mean, and with each line in a different color. Axes and legends should be appropriately titled.

```
# I don't do this exactly how the problem instructs, but I prefer my way.

sampling <- house %>% infer::rep_sample_n(size = 100, replace = FALSE, reps = 1000)

# Each time, I'm going to pull the desired coefficient from my nested estimator
# dataframes. So for "mean", I'll pull the mean. I can use the handy pluck() function.

samplestats <- sampling %>%
  group_by(replicate) %>%
  nest() %>%
    mutate(stats = map(.x = data, .f = ~tablefunc(.x$receipts))) %>%

# I mapped my function over every sample. Issue is, now I have dataframes
# mapped within my column. So I need to do a series of "plucking" to solve that.

  mutate(statistics = map(stats, ~pluck(.x, var = 2))) %>%
  mutate(mean = map(statistics, ~pluck(.x, 1))) %>%
  mutate(median = map(statistics, ~pluck(.x, 2))) %>%
  mutate(deletion1 = map(statistics, ~pluck(.x, 3))) %>%
  mutate(deletion2 = map(statistics, ~pluck(.x, 4))) %>%

# Don't need my huge nested columns anymore, so I'm going to get rid of those.

  select(-c(data, stats, statistics))

# Last, I need to make my columns numeric.

samplestats$mean <- as.numeric(samplestats$mean)
samplestats$median <- as.numeric(samplestats$median)
samplestats$deletion1 <- as.numeric(samplestats$deletion1)
samplestats$deletion2 <- as.numeric(samplestats$deletion2)

# Now we're good to start setting up the plot!
```

```
# Had to do some tricky manual labor here to get the legend with the right colors.

fills <- c("Del. Estimator II" = "lightblue", "Del. Estimator I" = "red", "Mean" = "yellow", "Median" =

# Have to make a dataframe for geom_text().

xint <- mean(house$receipts)

xint <- data.frame(xint)

ggplot(samplestats) +
  geom_density(aes(x = mean, fill = "Mean"), color = "orange", alpha = 0.4) +
```

```
geom_density(aes(x = median, fill = "Median"), color = "darkgreen", alpha = 0.4) +
geom_density(aes(x = deletion1, fill = "Del. Estimator I"), color = "darkred", alpha = 0.4) +
geom_density(aes(x = deletion2, fill = "Del. Estimator II"), color = "darkblue", alpha = 0.4) +
geom_vline(xintercept = xint$xint, color = "red") +
 geom_text(data = xint, x = xint$xint, y = 0.010, label = "True Mean", size = 3, angle = 90, vjust =
labs(x = "Funds Raised",
     y = "Density",
     title = "Estimator Sampling Distributions",
     subtitle = "Sample Size: 100",
     fill = "Legend")
```



2c) Interpret your results above in terms of bias.

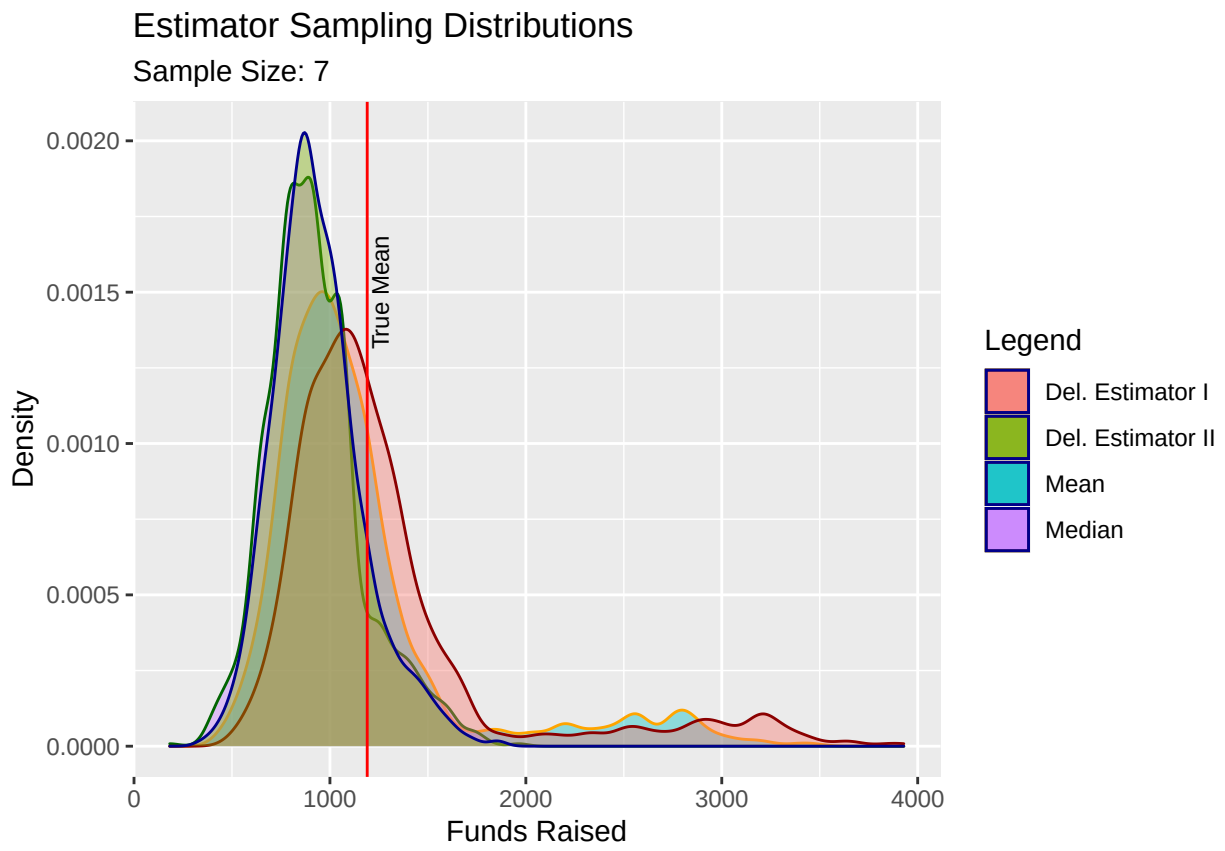
In this case, the sample mean is an unbiased estimator of the population mean. As we can see by our geom_density() graph, it doesn't systematically overestimate or underestimate the population mean. The first deletion estimator is also relatively unbiased, as the density plot for the estimator appears to be relatively closely centered around these two numbers.

The sample median and Deletion Estimator II, however, are more biased. The population median is , and the density plot for the sample median appears to generally be centered around that point. However, this is significantly skewed left of the population mean. Deletion Estimator II, also, is biased; it is generally far to the left of the mean.

2d) Repeat the simulation above using samples of size $n = 7$. How do the results compare?

```
sampling2d <- house %>% infer::rep_sample_n(size = 7, replace = FALSE, reps = 1000)

samplestats2d <- sampling2d %>%
  group_by(replicate) %>%
  nest() %>%
    mutate(stats = map(.x = data, .f = ~tablefunc(.x$receipts))) %>%
  mutate(statistics = map(stats, ~pluck(.x, var = 2))) %>%
  mutate(mean = map(statistics, ~pluck(.x, 1))) %>%
  mutate(median = map(statistics, ~pluck(.x, 2))) %>%
  mutate(deletion1 = map(statistics, ~pluck(.x, 3))) %>%
  mutate(deletion2 = map(statistics, ~pluck(.x, 4))) %>%
  select(-c(data, stats, statistics))

samplestats2d$mean <- as.numeric(samplestats2d$mean)
samplestats2d$median <- as.numeric(samplestats2d$median)
samplestats2d$deletion1 <- as.numeric(samplestats2d$deletion1)
samplestats2d$deletion2 <- as.numeric(samplestats2d$deletion2)
```

```
fills <- c("Del. Estimator II" = "lightblue", "Del. Estimator I" = "red", "Mean" = "yellow", "Median" =

xint <- mean(house$receipts)

xint <- data.frame(xint)

ggplot(samplestats2d) +
  geom_density(aes(x = mean, fill = "Mean"), color = "orange", alpha = 0.4) +
  geom_density(aes(x = median, fill = "Median"), color = "darkgreen", alpha = 0.4) +
  geom_density(aes(x = deletion1, fill = "Del. Estimator I"), color = "darkred", alpha = 0.4) +
  geom_density(aes(x = deletion2, fill = "Del. Estimator II"), color = "darkblue", alpha = 0.4) +
  geom_vline(xintercept = xint$xint, color = "red") +
   geom_text(data = xint, x = xint$xint, y = 0.0015, label = "True Mean", size = 3, angle = 90, vjust =
  labs(x = "Funds Raised",
       y = "Density",
       title = "Estimator Sampling Distributions",
       subtitle = "Sample Size: 7",
       fill = "Legend")
```



Just looking at the graph here, we see that there is a rather large range for some of these values – much bigger than in the previous sample, when we were only sampling 100 values. This makes sense, because there's a lot more room for "unrepresentative" (not a technical term) data when we are looking at an n = 7 than an n = 100. When n = 7, there are just naturally going to be more outliers.

2e) Compare your estimators in terms of efficiency, using the sampling distributions. Explain your results.

Efficiency is measured by variance. If the variance is lower, the estimator is more efficient. So let's just first take our variance to figure it out mathematically. I'm going to use the one with size = 100: the sample with size = 7 would have the same *relative* statistics – so the proportions would be the same; the numbers would change a bit, but not relative to one another.

```
meanvar <- var(samplestats$mean)
medianvar <- var(samplestats$median)
del1var <- var(samplestats$deletion1)
del2var <- var(samplestats$deletion2)
```

According to these data, the mean and Deletion Estimator I have higher variances than the median and Deletion Estimator II. When size = 100, the mean has variance 10873.2709048, the median has variance 1758.8109976, Deletion Estimator I has variance 11088.8672633, and Deletion Estimator II has variance 3521.814399. So this means that the median and Deletion Estimator II are more efficient than the other two.

2f) Using mean squared error of your sampling distributions, identify which of the estimators is the most unbiased and efficient.

The mean squared error function can be defined as $MSE[\hat{\theta}] = Bias(\hat{\theta})^2 + V(\hat{\theta})$.

Bias, meanwhile, can be found out if $E[\hat{\theta}] - E[\theta] = 0$. I highly doubt that any of our samples will be entirely unbiased, but we can at least see *how* biased and unbiased our samples are. Let's use size = 100 again.

```
biasmean <- (mean(samplestats$mean) - mean(house$receipts))
biasmedian <- (mean(samplestats$median) - mean(house$receipts))

# Now, I have to use the functions that I used within my "tablefunc"
#function to find the values for the full population of the deletion estimators.

delet <- house$receipts %>%
    sort(decreasing = FALSE)

delet1 <- delet[-1] %>%
    mean()

deletintermediate <- delet[-c(1, 2)]

deletlast <- deletintermediate %>%
    sort(decreasing = TRUE)

delet2 <- deletlast[-c(1, 2)] %>%
    mean()


biasdel1 <- (mean(samplestats$deletion1) - mean(house$receipts))
biasdel2 <- (mean(samplestats$deletion2) - mean(house$receipts))
```

Now, we can go ahead and find the MSE of all four.

```
msemean <- biasmean^2 + meanvar
msemedian <- biasmedian^2 + medianvar
msedel1 <- biasdel1^2 + del1var
msedel2 <- biasdel2^2 + del2var
```

We get:

The mean squared error of the sample mean is 10888.9933073. The mean squared error of the sample median is 85245.1775483. The mean squared error of the sample first deletion estimator is 11339.1230864. The mean squared error of the sample second deletion estimator is 26232.437179.

As we can see, the estimator with the lowest mean squared error – meaning the most efficient and unbiased – is the mean.