Homework 05 - STAT440

Joseph Sepich (jps6444)

10/04/2020

```
set.seed(42)
```

Problem 1

```
# load dataset
scores <- data.matrix(read.csv('./data/score.csv'))
scores[0:10,]
```

```
##       HW1 HW2 HW3 HW4 HW5
##  [1,]  93  99  81  81  98
##  [2,]  76  94  97  85  98
##  [3,]  91  88  86  80  98
##  [4,]  66  87  76  85  82
##  [5,]  76  76  78  85  76
##  [6,]  64  74  87  77  88
##  [7,]  62  81  78  78  82
##  [8,]  71  85  82  75  68
##  [9,]  75  72  70  75  85
## [10,]  77  87  72  75  54
```

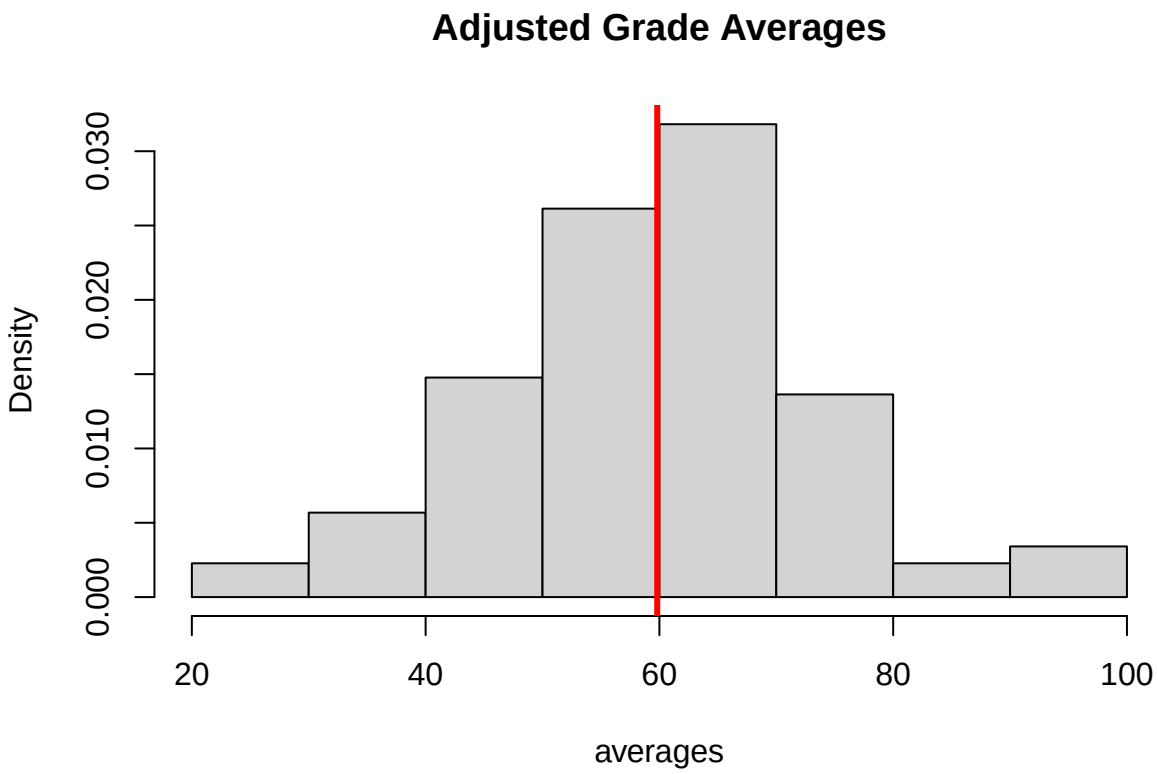
Adjust the scores by dropping the lowest.

```
adjusted_scores <- t(as.matrix(apply(scores, 1, function(x) x[-(match(min(x), x))])))
adjusted_scores[0:10,]
```

```
##       [,1] [,2] [,3] [,4]
##  [1,]   93   99   81   98
##  [2,]   94   97   85   98
##  [3,]   91   88   86   98
##  [4,]   87   76   85   82
##  [5,]   76   78   85   76
##  [6,]   74   87   77   88
##  [7,]   81   78   78   82
##  [8,]   71   85   82   75
##  [9,]   75   72   75   85
## [10,]   77   87   72   75
```

Part a

```
averages <- rowMeans(adjusted_scores)
sample_avg <- mean(averages)
hist(averages, freq=FALSE, main = "Adjusted Grade Averages")
abline(v=sample_avg, col="red", lw=3)
```



Part b

You could express θ as an expectation of an indicator function: $E[1_{X>C}(x)] = P(X > C) = \theta$. This value can be approximated by sampling from X and you obtain a 0 or 1 depending on whether it is greater than C or not. You would divide the sum of samples (0 or 1) by N to get the expected value of the indicator, which is also the approximation of the probability $P(X > C)$. In this specific problem we are sampling from the student's average adjusted scores. If the average adjusted score we randomly select is greater than C the indicator is a 1, otherwise it is a zero.

Part c

```
C <- 70
n <- 10000
# sample from data (X)
samples <- sample(averages, n, replace=TRUE)
```

```r
# expectation of indicator
theta_hat <- mean(samples > C)
print(theta_hat)
```

```
## [1] 0.1961
```

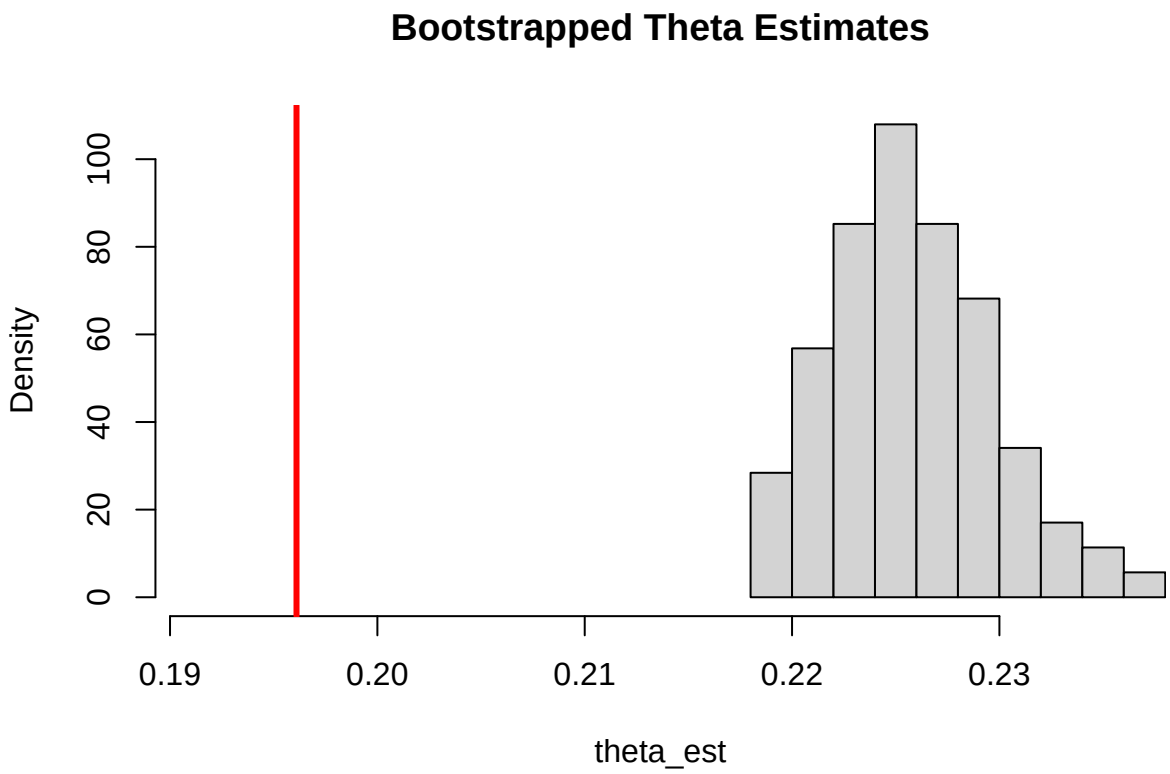
The Monte Carlo estimate states that $P(X > 70) \approx 0.1961$, so about 19.61% of students have an average adjust test score about 70%.

Part d

The histogram looks similar to a normal distribution. Often grades of students fall upon a normal distribution and instructors usually use a normal distribution when grading on a curve. For this reason we will use a normal distribution for our parametric bootstrap.

Part e

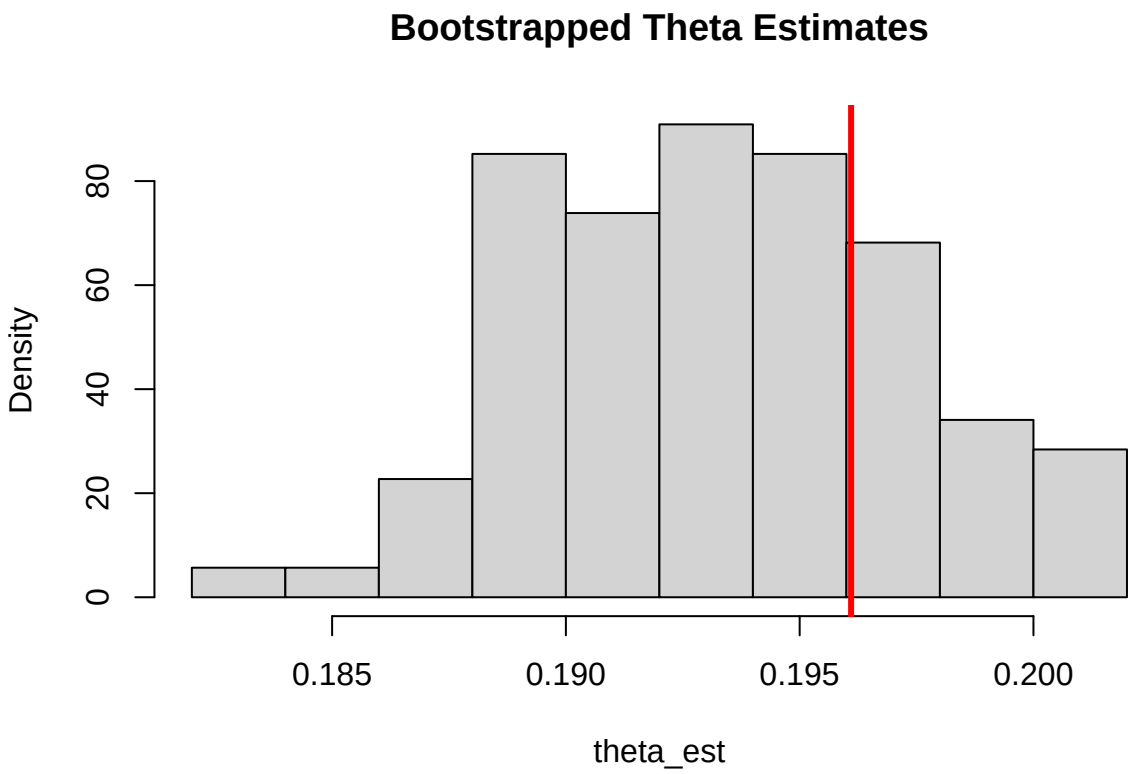
```r
k <- 10000
n <- length(averages)

# parameter estimates
mu_hat <- mean(averages)
sigma_hat <- sd(averages)

# parametric bootstrap from estimated distribution
boot_samples <- matrix(rnorm(k*n, mean=mu_hat, sd=sigma_hat), nrow=k)

# apply indicator function
boot_samples <- t(as.matrix(apply(boot_samples, c(1,2), function(x) if (x <= C) {0} else {1})))
theta_est <- rowMeans(boot_samples)

hist(theta_est, freq = FALSE, main="Bootstrapped Theta Estimates", xlim = c(theta_hat - 0.005, max(theta
abline(v=theta_hat, col="red", lw=3)
```



This results makes sense due to our choice of a normal distribution. The histogram of the actual test scores shows fewer students past 70 then you would expect if you traced a normal distribution over it.

Part f

```
k <- 10000
n <- length(averages)

# non-parametric sampling procedure
boot_samples <- matrix(sample(averages, k*n, replace=TRUE), nrow=k)

# apply indicator function
boot_samples <- t(as.matrix(apply(boot_samples, c(1,2), function(x) if (x <= C) {0} else {1})))
theta_est <- rowMeans(boot_samples)

hist(theta_est, freq = FALSE, main="Bootstrapped Theta Estimates")
abline(v=theta_hat, col="red", lw=3)
```



Problem 2

Part a

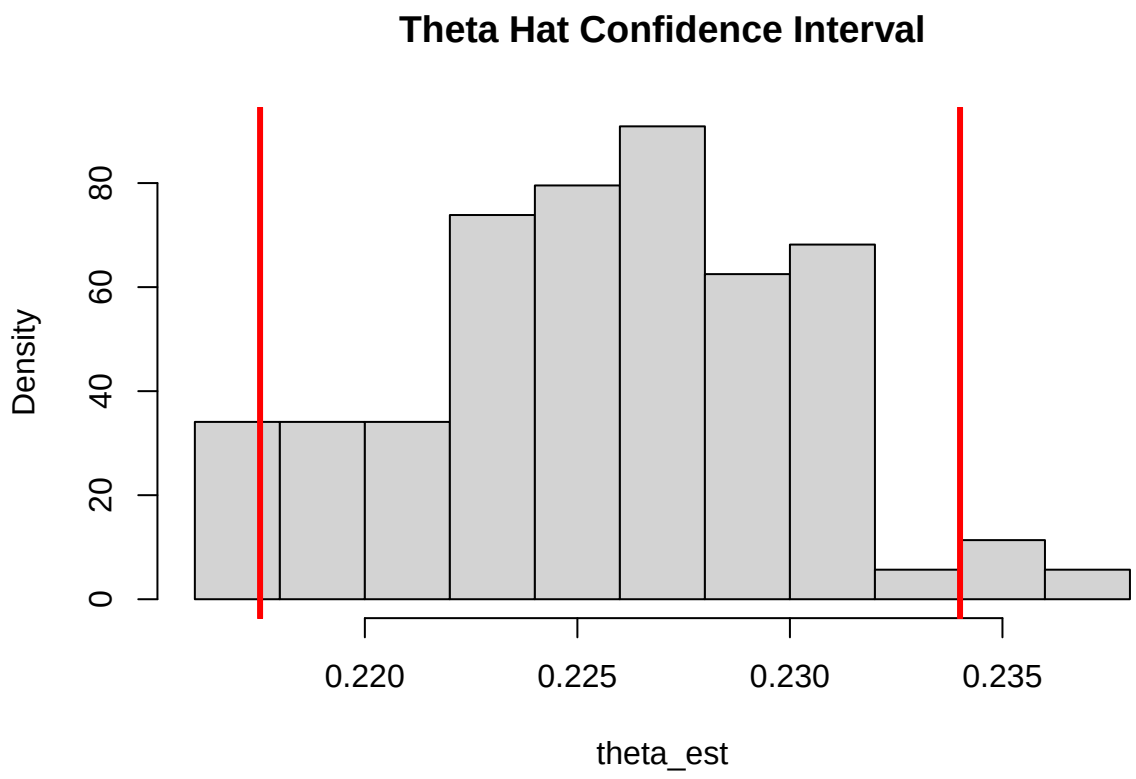
Again to use Monte Carlo methods to approximate $P(l \leq \hat{\theta} \leq u)$ we can use an indicator function $1_{\hat{\theta} \in [l,u]}(\hat{\theta}_i)$. We need to find the values l and u that satisfy $E[1_{\hat{\theta} \in [l,u]}(\hat{\theta}_i)] = 1 - \alpha$. This gives us the following sum $\frac{1}{M} \Sigma_{i=1}^{M} 1_{\hat{\theta} \in [l,u]} \hat{\theta}_i$. We want to find an l and u that makes the value of this sum equal to $1 - \alpha$. All we are doing with this sum is finding the percentage of our bootstrapped estimates that lie in the interval between the lower bound and upper bound. For a 95% confidence interval we would want to find bounds where this sum evaluates to 0.95.

Part b

We can use the quantile function in R to find a $1 - \alpha$ confidence interval by first obtaining the bootstrapped estimates for $\hat{\theta}$ and finding the $\alpha/2$ and $1 - \alpha/2$ quantiles for that vector of bootstrapped estimates of $\hat{\theta}$. Getting these two quantiles gives us the lower and upper bounds for an interval in the middle that contains $1 - \alpha$ percent of the bootstrapped estimates, which is exactly the value we stated we wanted in the previous part.

Part c

```
alpha <- 0.05

# copied and pasted procedure
k <- 10000
n <- length(averages)

# parameter estimates
mu_hat <- mean(averages)
sigma_hat <- sd(averages)

# parametric bootstrap from estimated distribution
boot_samples <- matrix(rnorm(k*n, mean=mu_hat, sd=sigma_hat), nrow=k)

# apply indicator function
boot_samples <- t(as.matrix(apply(boot_samples, c(1,2), function(x) if (x <= C) {0} else {1})))
theta_est <- rowMeans(boot_samples)

bounds <- quantile(theta_est,probs=c(alpha/2,(1-alpha/2)))
hist(theta_est, freq = FALSE, main="Theta Hat Confidence Interval")
abline(v=bounds, col="red", lw=3)
```

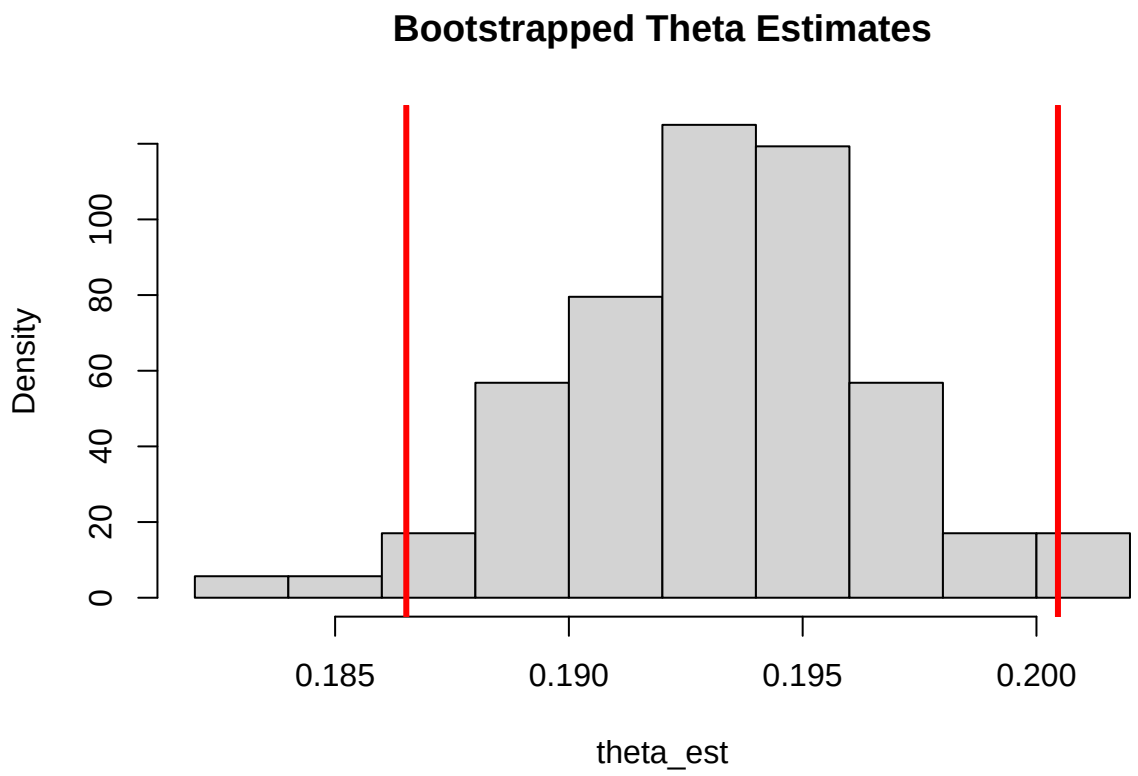


```r
print(bounds)
```

```
##      2.5%     97.5%
## 0.2175350 0.2340025
```

Part d

```r
# copy and pasted procedure
k <- 10000
n <- length(averages)

# non-parametric sampling procedure
boot_samples <- matrix(sample(averages, k*n, replace=TRUE), nrow=k)

# apply indicator function
boot_samples <- t(as.matrix(apply(boot_samples, c(1,2), function(x) if (x <= C) {0} else {1})))
theta_est <- rowMeans(boot_samples)

bounds <- quantile(theta_est,probs=c(alpha/2,(1-alpha/2)))
hist(theta_est, freq = FALSE, main="Bootstrapped Theta Estimates")
abline(v=bounds, col="red", lw=3)
```



```r
print(bounds)
```

```
##      2.5%     97.5%
## 0.1865225 0.2004600
```

Problem 3

We want to test to see if exactly half to students have at least an average score of $C = 70$.

Part a

Null Hypothesis (H_0) would be $P(X \geq C)_0 = \theta_0 = 0.5$. The alternative hypothesis (H_A) would be $p(X \geq C)_A = \theta_A \neq 0.5$

Part b

Here we can use a test statistic $T := |\hat{\theta} - \theta_0|$, which makes our observed test statistic $T_{obs} := |\hat{\theta}_{obs} - \theta_0|$. This would give us an "extreme set" of any value of that falls within the set $[T_{obs}, \infty)$.

Part c

The p-value is an estimate of how often we expect to see values of $T := |\hat{\theta} - \theta_0|$ in our extreme set given a distribution with our null hypothesis. Translated to probability our p-value is $P(T \geq T_{obs})$. We can see that we can translate this probability into an expectation for estimation with Monte Carlo: $E[1_{T \in [T_{obs}, \infty]}(T_i)] = P(T \geq T_{obs})$. Therefore all we have to do is use a bootstrap to obtain various test statistics T, then apply the indicator function to them to determine if they are in the extreme set. We then obtain our p-value by finding the sample mean of this indicator function result.

Part d

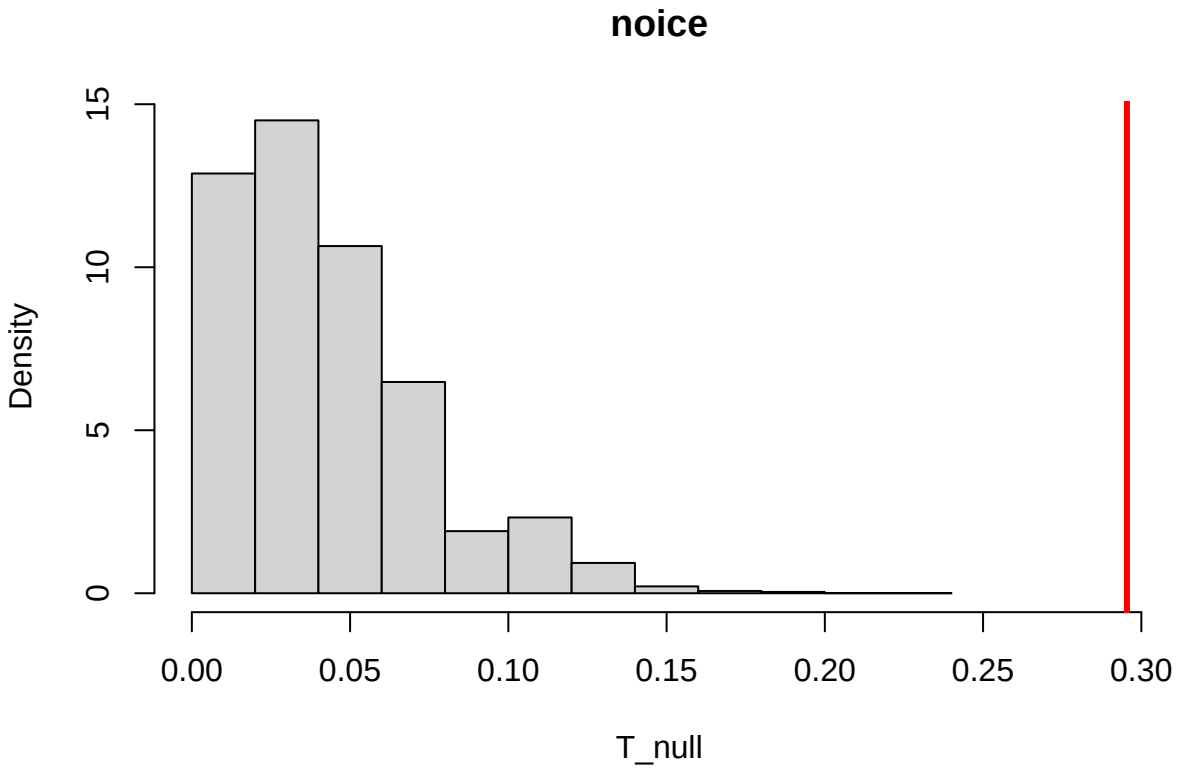
Since our null hypothesis gives us a parameter θ_0 that represents a proportion, we can sample from the binomial distribution with parameters n and p under the null where n is our sample size and p is the value p_0 dictated by the null hypothesis.

Part e

TODO

```r
k <- 10000
n <- length(averages)
p_null <- 0.5

# sample theta hat from null
theta_null <- rbinom(k, n, p_null) / n
T_null <- abs(theta_null - p_null)

# observed T
```

```r
# apply indicator to averages
# expectation of indicator
theta_hat <- mean(averages >= C)
T_obs <- abs(theta_hat - p_null)

hist(T_null, freq = FALSE, main="noice", xlim = c(0, T_obs))
abline(v=T_obs, col='red', lw=3)
```



Part f

```r
p_val <- mean(T_null >= T_obs)
print(p_val)
```

```
## [1] 0
```

This p-value represents the percentage of test statistics under the null hypothesis that were as extreme as the observed or more extreme. Since none were as extreme as the observed test statistic we get a p-value of zero. This gives us enough evidence to reject the null hypothesis that half of the students have an average test score at least 70 or higher.

Part g

TODO

To sample under the null hypothesis we can perform a transformation in the sampling procedure. When doing the non-parametric bootstrap on the averages we expect to get a certain sample mean usually of $\overline{X}$. If we subtract this sample mean from the samples we take we would expect to get 0 on average: $E[X - \overline{X}] = \frac{1}{n}\Sigma(X_i - \overline{X}) = \overline{X} - \frac{n}{n}\overline{X} = 0$. Therefore we can bootstrap under the null by sampling from our observations with replace, but for each sample we also add the null mean and subtract the sample mean: $y_i = x_i - \overline{X} + \mu_0$.