WS07 - Law of Large Numbers and Central Limit Theorem

Shawn Kim

Directions: *Please upload a PDF to Gradescope that includes both your written responses and corresponding R code inputs/outputs (if requested) for each problem.*

Special Directions When responding to the explanation questions, it may be helpful to look back at the Monte Carlo with importance sampling section of the notes. Be sure to demonstrate the correct use of mathematical notation in your work. When showing your work, clearly show your reasoning by entering all necessary algebra/calculations as text or inserting a clear well-cropped image of your work using an R chunk.

Probem 1. Consider the function $g(x) = \dfrac{1}{\sin(\exp(\sqrt{x}) - 1)}$ whose numerical integral from $x = 0$ to $x = 9/16$ is defined in `R` with the name `gx`.

```
# NOTE: To define a function in R follows the general structure below,
# function_name <- function(x) {define function here, return function value}
# NOTE: You've seen this in WS08, but it can by tricky, so we are helping out
# again
gx = function(x) {
    gx = (sin(exp(sqrt(x)) - 1))^-1
    return(gx)
}
```

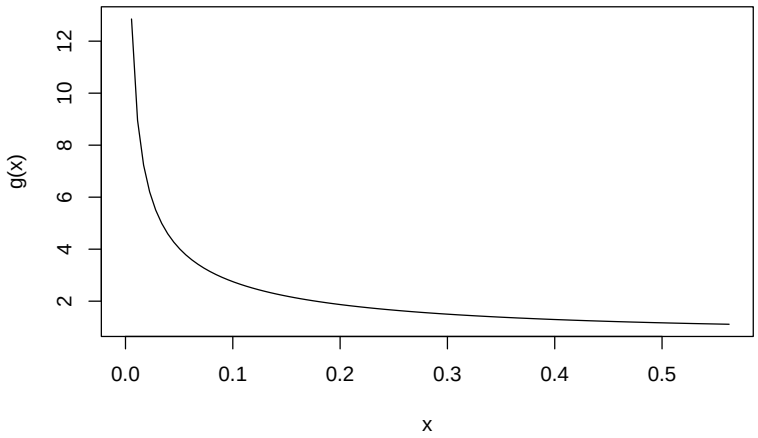
In this problem, we would like to estimate $\int_0^{9/16} g(x)\,dx$ using Monte Carlo simulations.
NOTE: In `R`*, we can estimate the integral using the* `integrate()` *command to get* $\int_0^{9/16} g(x)\,dx \approx 1.31075$.

```
integrate(gx, lower = 0, upper = 9/16)
```

```
## 1.31075 with absolute error < 1e-06
```

Problem 1 Part a) Below is a plot of g on the interval $(0, 9/16]$. Explain why we would expect that importance sampling would be useful for getting an accurate measurement of our definite integral $\int_0^{9/16} g(x)\,dx$?

```
curve(gx, from = 0, to = 9/16, xlab = "x", ylab = "g(x)")
```



It would be important to consider importance sampling for getting the definite integral for this plot because we can see from the graph that the highest g(x) values are when x is approaching 0 which would have more of an impact on the definite integral estimate. We should use an appropriate PDF to draw samples over X that will give more chance for the high values of g(x) to be considered (and then weigh sum of g(x) by the chance that x happened)

Problem 1 Part b) Use the **simple Monte Carlo** integration procedure to produce 1000 estimates of $\int_0^{9/16} g(x)\ dx$ based on a sample of 240 uniform random variables. Determine the mean and standard deviation of the 1000 estimates of $\int_0^{9/16} g(x)\ dx$.

```
set.seed(2022)
int_gx_simple240 = rep(0, 1000)  # vector to store each of the 1000 estimates of int_a^b g(x)

# NOTE: We will perform the MC integration procedure 1000 times and store each
# integral estimate in the k-th index of vector int_gx_simple240

# NOTE: The for loop (below) will repeat the MC process 1000 times

for (k in 1:1000) {
    xs = runif(240, 0, 9/16)  #uniformly distributed xvals
    gx_values = gx(xs)  #the yvals corresponding to the above xvals
    avg_gx = mean(gx_values)  #compute the mean of the yvals
    int_gx_simple240[k] = avg_gx * (9/16)  #the mean of yvals times interval length
}

# NOTE: Compute the mean of the vector int_gx_simple240 below
mean(int_gx_simple240)
```

```
## [1] 1.303507
```

```
# NOTE: Compute the standard deviation of the vector int_gx_simple240 below
sd(int_gx_simple240)
```

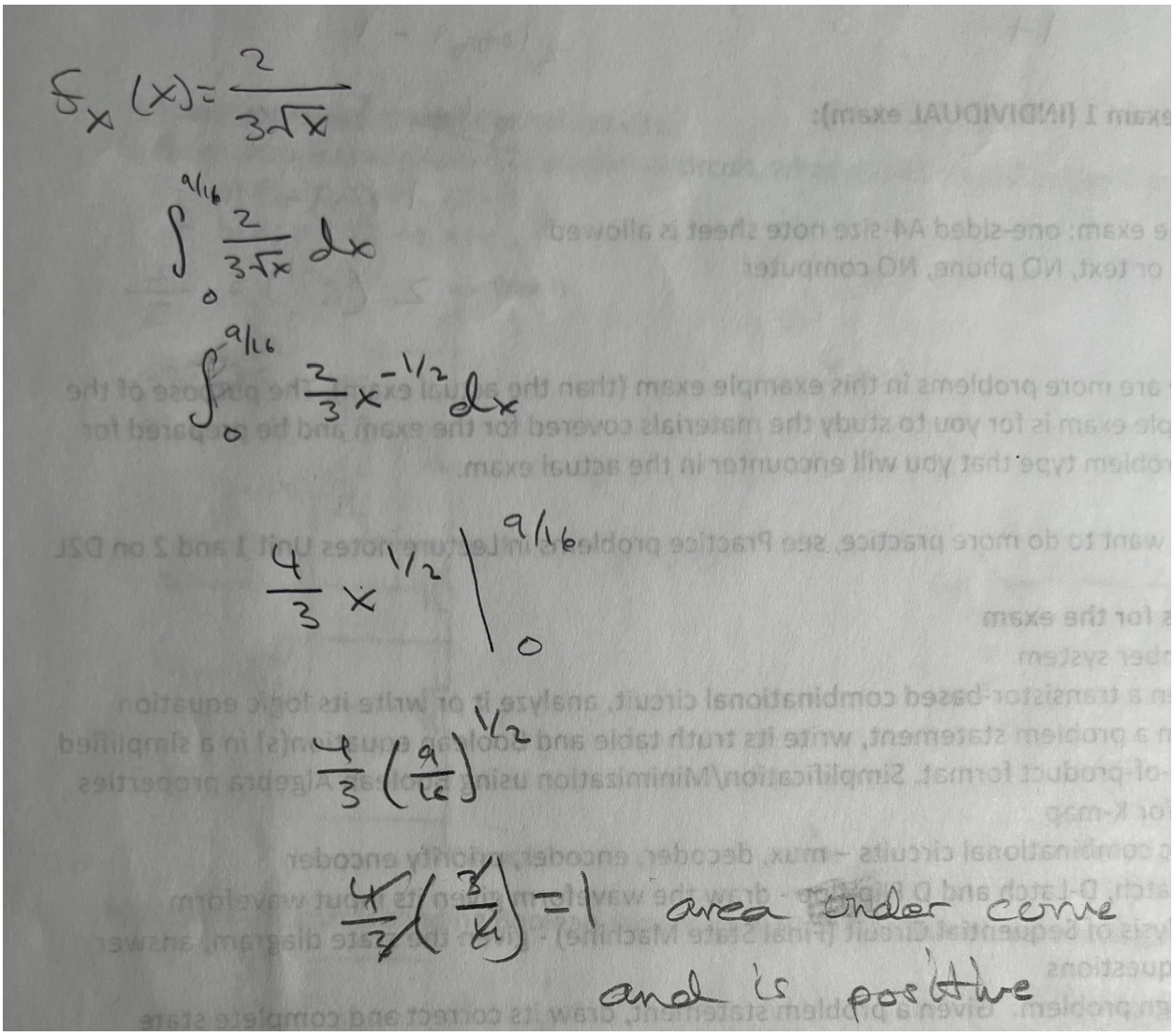
```
## [1] 0.1785548
```

Problem 1 Part c) Before we can perform **Monte Carlo integration with importance sampling**, we need a density function that 1) has the most probability where $g(x)$ is rapidly changing, and 2) allows us to easily determine a weight function (which will be explored in part (e))

Verify that $f_X(x)$ (below) is a valid density function.

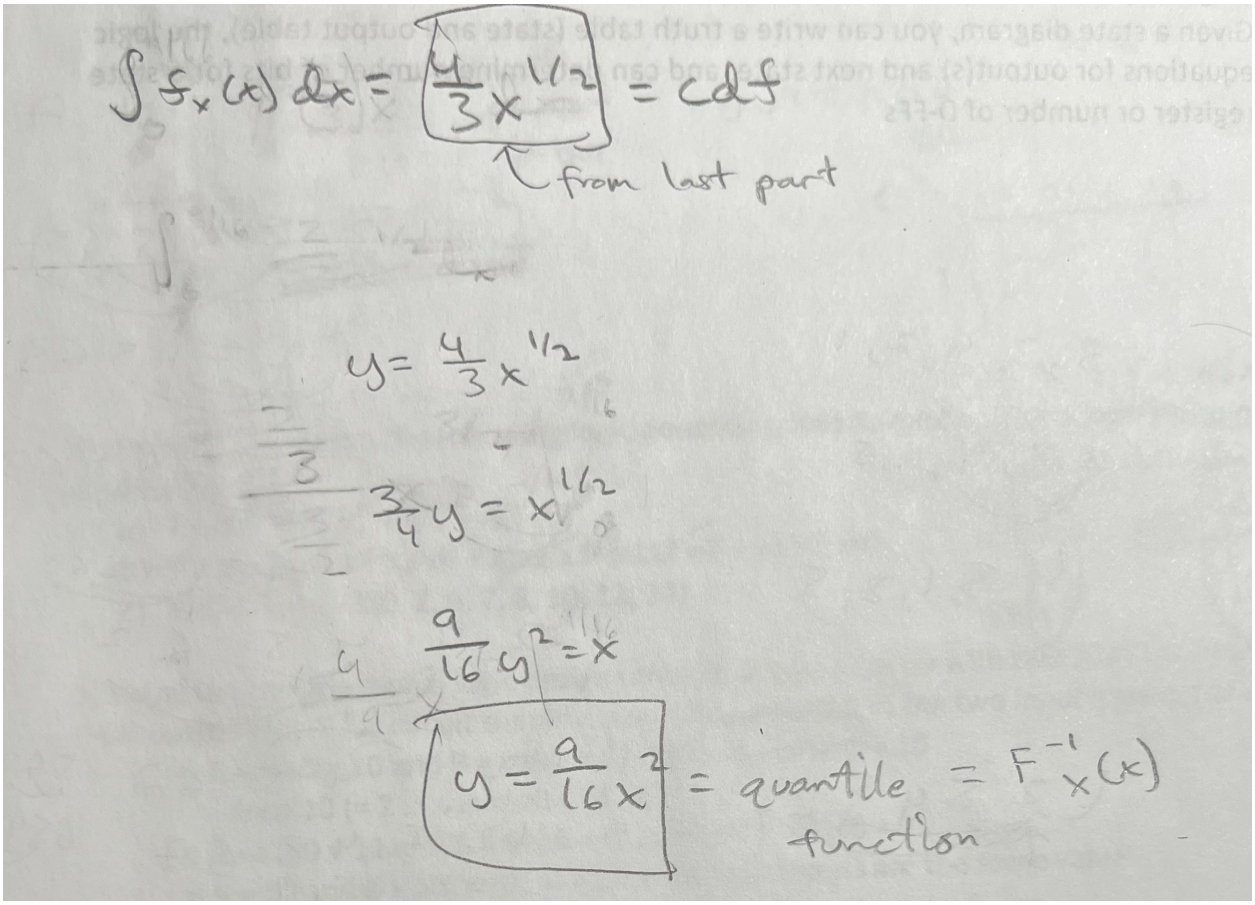
$$f_X(x) = \begin{cases} \frac{2}{3\sqrt{x}}, & 0 < x \leq 9/16, \\ 0 & \text{otherwise} \end{cases}$$

Hint: Recall that a valid density function is positive and has area under the curve equal to one over the interval of interest.



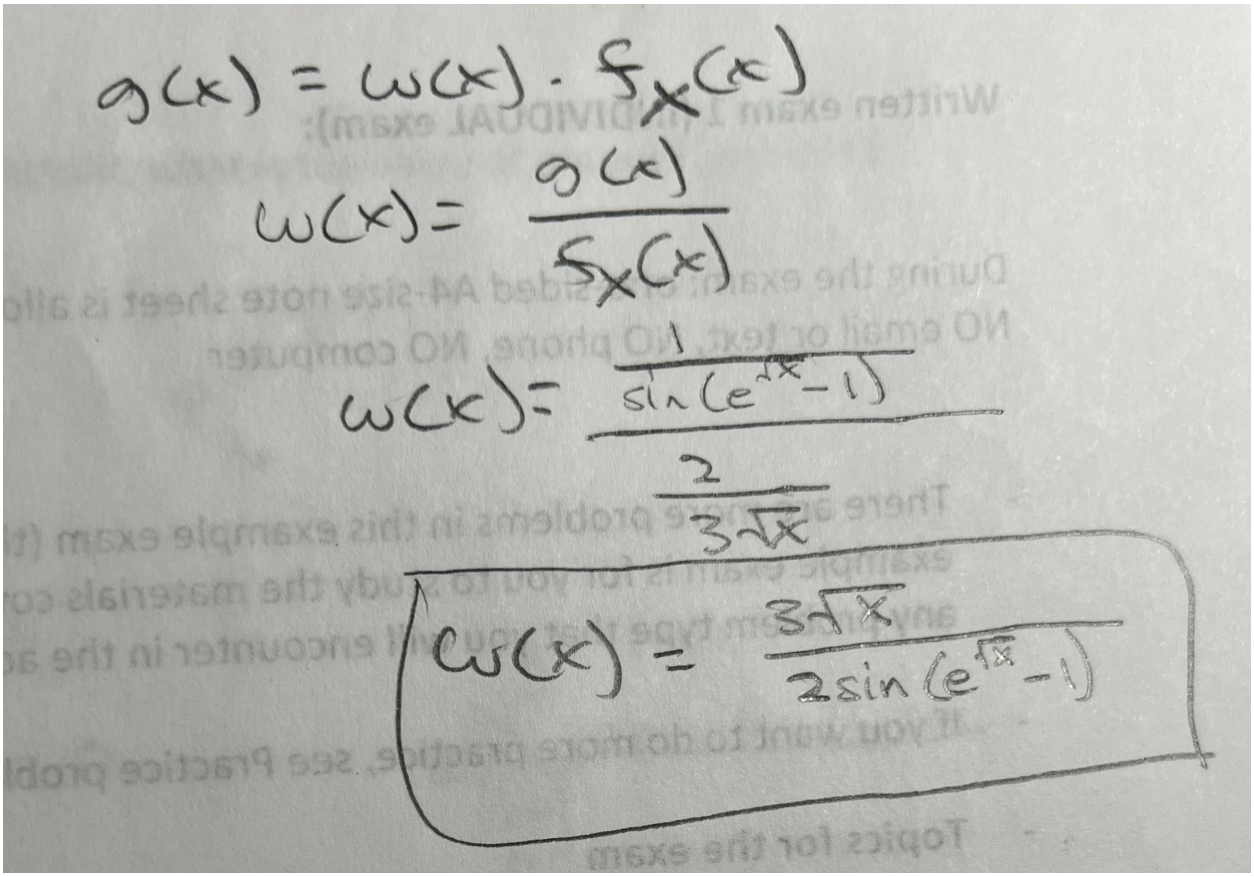
Problem 1 Part d) Determine the cumulative distribution function of the density function from part (c). Additionally, determine the corresponding probability transform.

Hint: Recall that we can integrate the density function, $f_X(x)$, in order to determine the cumulative distribution function, $F_X(x)$. Then, finding the probability transform, $F_X^{-1}(x)$, is the same as finding the inverse of the cumulative distribution function.



Problem 1 Part e) If we use the density function $f_X(x)$ from part (c) to perform Monte Carlo integration with importance sampling, what would the weight function $w(x)$ be?

Hint: Recall that we want $g(x) = w(x) \cdot f_X(x)$ where $g(x) = \frac{1}{\sin(\exp(\sqrt{x})-1)}$.



Problem 1 Part f) Use Monte Carlo integration with importance sampling to provide 1000 estimates of $\int_0^{9/16} g(x)\,dx$ based on a sample of 240 random variables with proposal density $f_X(x)$. Determine the mean and standard deviation of the 1000 estimates of $\int_0^{9/16} g(x)\,dx$.

```
set.seed(2022)
int_gx_importance240 = rep(0, 1000)  # vector to store the 1000 estimates of int_a^b g(x)

# NOTE: We will perform the MC integration with importance sampling 1000 times
# and store each integral estimate in the k-th index of vector
# int_gx_importance240

for (k in 1:1000) {

    # NOTE: First, we want a vector of 240 uniformly distributed RVs that are
    # probabilities
    u = runif(240)
    x = 9/16 * (u^2)
    # NOTE: Next, we want a vector of RVs distributed according to your density
    # function

    # NOTE: x = a vector of 240 RVs distributed according to your density
    # function NOTE: which you get from applying the probability transform to u

    # NOTE: Finally, we calculate the mean of w(x) for 240 xvals NOTE: This is
```

```
    # the value of the integral estimated by MC with importance sampling

    int_gx_importance240[k] = mean(3 * sqrt(x)/(2 * sin(exp(sqrt(x)) - 1)))

}

# NOTE: Compute the mean of the vector int_gx_importance240 below FILL IN

mean(int_gx_importance240)
```

```
## [1] 1.310694
```

```
# NOTE: Compute the standard deviation of the vector int_gx_importance240 FILL
# IN

sd(int_gx_importance240)
```

```
## [1] 0.00529576
```

Problem 1 Part g) Determine the value of the ratio $SD_{\text{simple MC}}/SD_{\text{importance MC}}$, where $SD_{\text{simple MC}}$ was calculated in part (b) and $SD_{\text{importance MC}}$ was calculated in part (f). What does this ratio convey? Which method (simple MC or MC with importance sampling) is more likely to produce an estimate closer to the true value of the definite integral? Briefly explain why this method is more accurate, keeping in mind your response from part (a).

```
# ratio simple MC sd/importance MC sd
0.1785548/0.00529576
```

```
## [1] 33.71656
```

This ratio conveys that importance MC estimate had a lot less variance and a lot more accuracy, with a standard deviation over 30 times smaller than the simple MC. The importance sampling MC is more likely to produce an estimate closer to the true value of the definite integral because as explained in part (a), importance sampling gives us a better distribution of the relevant x values to use in extimating area under g(x)

Problem 2. Let L be the length of a pendulum and g be the acceleration due to gravity.
For small angles, the period, T, of a pendulum is given by

$$T = 2\pi\sqrt{\frac{L}{g}}. \tag{1}$$

In 2005, the Huygens space probe landed on the surface of Titan. Pictures of Titan were taken with a camera designed at the University of Arizona. We will use this relationship and observations from a swinging pendulum to estimate the acceleration due to gravity on the moon.

Problem 2 Part a) Assume that the length of the pendulum is fixed such that $L = 1$ meter. Now suppose the probe makes repeated independent measurements, $T_1, T_2, \cdots T_{25}$, of the period. If these measurements have population mean $\mu_T = 5.40$ seconds and population standard deviation $\sigma_T = 0.12$ seconds, determine the mean and standard deviation of $\bar{T}$. That is, determine $\mu_{\bar{T}}$ and $\sigma_{\bar{T}}$.

Hint: Recall that we can use the population mean and population standard deviation, together with the law of large numbers, to compute the mean and standard deviation of $\bar{T}$ (the mean of the 25 repeated independent measurements). Note that, in relation to the LoLN, each of the 25 repeated measurements/samplings can be viewed as the sample mean of the corresponding sampling taken from the probe's sensors.

```
# mean of T bar
5.4
```

```
## [1] 5.4
```

```
# std of T bar = sd pop / sqrt (n)

0.12/sqrt(25)
```

```
## [1] 0.024
```

Problem 2 Part b) Use equation (1) to create an estimator, $\hat{g}$, for the acceleration due to gravity, g.

Hint: Recall that an estimator $\hat{g}$ is a function that will calculate an estimate for the acceleration due to gravity based on the input of experimental data for length and period. Mathematically, this equivalent to solving equation (1) for g, then labeling this expression as the estimator $\hat{g}$. Notice that $\hat{g}$ will be a function of T (since L is fixed), so we can write $\hat{g}(T)$ to be more precise in our notation.

$$\frac{T}{2\pi} = \sqrt{\frac{L}{g}}$$

$$\frac{T^2}{4\pi^2} = \frac{L}{g}$$

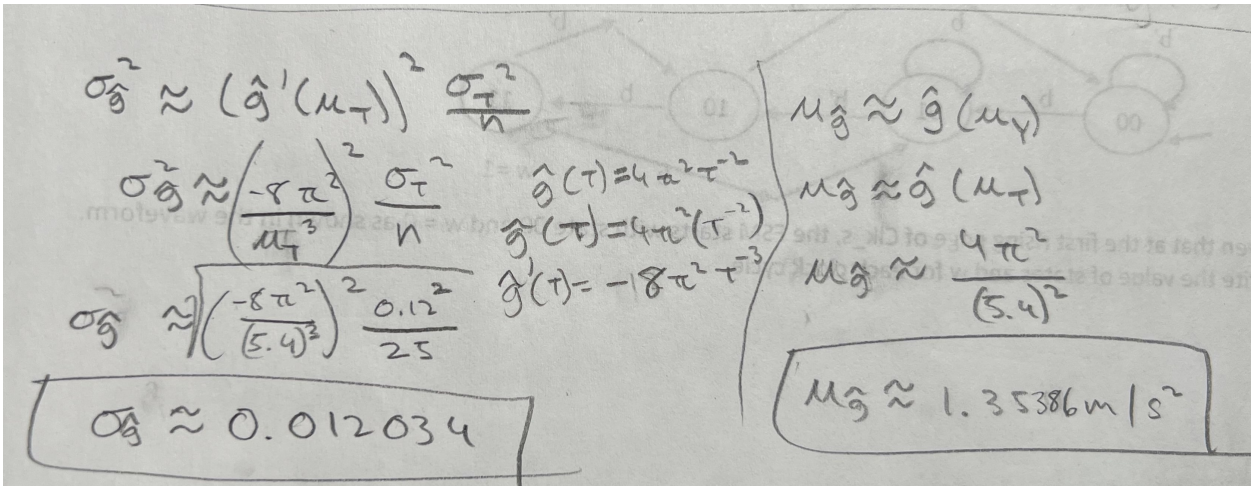
$$g = \frac{L 4\pi^2}{T^2}$$

$$\hat{g}(T) = \frac{L 4\pi^2}{T^2}$$

$$\hat{g}(T) = \frac{4\pi^2}{T^2}$$

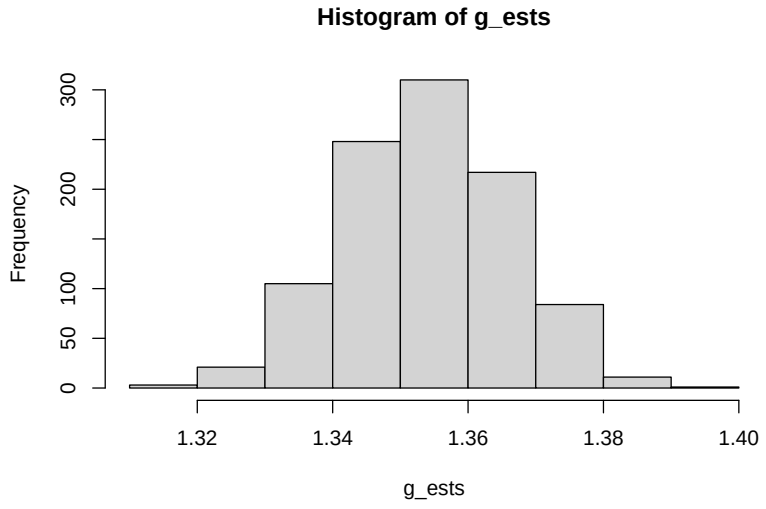
Problem 2 Part c) Using the delta method, estimate the mean and standard deviation of the estimator $\hat{g}$. That is, determine $\mu_{\hat{g}}$ and $\sigma_{\hat{g}}$.

Hint: Recall that, per the delta method as defined in Session 13 notes, for an estimator $\hat{g}(Y)$ we have that the mean $\mu_{\hat{g}} \approx \hat{g}(\mu_Y)$ and variance $\sigma^2_{\hat{g}} \approx [\hat{g}'(\mu_Y)]^2 \ \sigma^2_Y/n$.



Problem 2 Part d) Below is a simulation of estimates. Describe the histogram of simulated estimates shown below (including center, shape, skewness, etc.). Additionally, compare the mean and standard deviation of the estimates in the simulation below to the values given by the delta method in part (c).

```
set.seed(2021)
Tbar <- rnorm(1000, mean = 5.4, sd = 0.024)
g_ests <- (4 * pi^2)/(Tbar^2)
hist(g_ests)
```




```
mean(g_ests)
```

```
## [1] 1.353785
```

```
sd(g_ests)
```

```
## [1] 0.012259
```

the histogram is in the shape of a normal distribution with center around the 1.35 to 1.36 mark and a very slight right skew. The estimates from the distribution compared to the delta method have very close mean values with the delta method having a slightly better (lower) standard deviation.