Problem 46

Overview

Chromatin folding is process in biology. We assume that chromatin polymers follow a random walk model. In this model, the two-dimensional distance between two chromatin fibers, R, follows a Raleigh distribution, with density

$$f(r|\theta) = \frac{r}{\theta^2} e^{\frac{-r^2}{2\theta^2}}$$

where $r \geq 0$ and $\theta > 0$.

Estimating Theta

Maximum Likelihood

Derivation of the Estimator

The likelihood function is

$$lik(\theta) = \frac{1}{\theta^{2n}} e^{-\frac{1}{2\theta^2}\sum_{i=1}^{n} r_i^2} \prod_{i=1}^{n} r_i$$

The log likelihood is thus

$$l(\theta) = -2n\ln(\theta) + \sum_{i=1}^{n} \ln(r_i) - \frac{1}{2\theta^2}\sum_{i=1}^{n} r_i^2$$

The first-order condition for maximizing the likelihood satisfies

$$0 = \frac{-2n}{\hat{\theta}} + \frac{1}{\hat{\theta}^3}(\sum_{i=1}^{n} r_i^2)$$

Which, after rearranging, gives the maximum likelihood estimator for θ, $\hat{\theta}_{MLE}$

$$\hat{\theta}_{MLE} = \frac{1}{\sqrt{2}}\sqrt{\frac{1}{n}\sum_{i=1}^{n} r_i^2}$$

Asymptotic Variance

As usual, the asymptotic variance of a maximum likelihood estimator is roughly

$$Var(\hat{\theta}_{MLE}) \approx \frac{1}{nI(\theta)}$$

where under sufficient smoothness conditions that at the level of this textbook we will assume hold,

$$I(\theta) = E\left[\frac{\partial}{\partial\theta}\log(f(x|\theta))\right]^2 = -E\left[\frac{\partial^2}{\partial\theta^2}\log(f(x|\theta))\right]$$

$f(r|\theta) = \frac{r}{\theta^2}e^{\frac{-r^2}{2\theta^2}}$, so

$$\begin{aligned}\frac{\partial}{\partial\theta}\log f(r|\theta) &= -\frac{2}{\theta} + \frac{r^2}{\theta^3},\\ \frac{\partial^2}{\partial\theta^2}\log f(r|\theta) &= \frac{2}{\theta^2} - 3\frac{r^2}{\theta^4},\\ I(\theta) &= -\frac{2}{\theta^2} + 3\frac{E(R^2)}{\theta^4}\end{aligned}$$

The expectation of R^2 is

$$E(R^2) = \frac{1}{\theta^2}\int_0^\infty r^3 e^{\frac{-r^2}{2\theta^2}}dr = 4\theta^2\int_0^\infty x^3 e^{-x^2}dx = 2\theta^2\int_0^\infty ue^{-u}du$$

by making the substitutions $x = \frac{r}{\sqrt{2}\theta}$ and $u = x^2$. Integrating by parts,

$$E(R^2) = 2\theta^2\left[-ue^{-u}|_0^\infty + \int_0^\infty e^{-u}du\right] = -2\theta^2(e^{-u})|_0^\infty = 2\theta^2$$

Thus

$$\begin{aligned}I(\theta) &= \frac{4}{\theta^2},\\ Var(\hat{\theta}_{MLE}) &\to \frac{\theta^2}{4n}\end{aligned}$$

Method of Moments

Derivation of the Estimator

The expectation of R is

$$E(R) = \frac{1}{\theta^2}\int_0^\infty r^2 e^{\frac{-r^2}{2\theta^2}}dr = (2\sqrt{2})\theta\int_0^\infty x^2 e^{-x^2}dx$$

by making the substitution $x = \frac{r}{\sqrt{2}\theta}$. Integrating by parts,

$$E(R) = (2\sqrt{2})\theta\left[\frac{1}{2}xe^{-x^2}|_0^\infty + \frac{1}{2}\int_0^\infty e^{-x^2}dx\right] = \sqrt{2}\theta\int_0^\infty e^{-x^2}dx = \theta\sqrt{\frac{\pi}{2}}$$

since $\int_0^\infty e^{-x^2} dx = \sqrt{\pi}$, and the integrand is an even function. Thus

$$\hat{\theta}_{MoM} = \bar{X}\sqrt{\frac{2}{\pi}}$$

Asymptotic Variance

By the Central Limit Theorem, $\bar{X}$ converges to a normal random variable with mean $E(R)$ and variance $\frac{Var(R)}{n}$ as n approaches infinity. From previous results.

$$Var(R) = E(R^2) - E(R)^2 = 2\theta^2 - \frac{\pi}{2}\theta^2 = \frac{4-\pi}{2}\theta^2$$

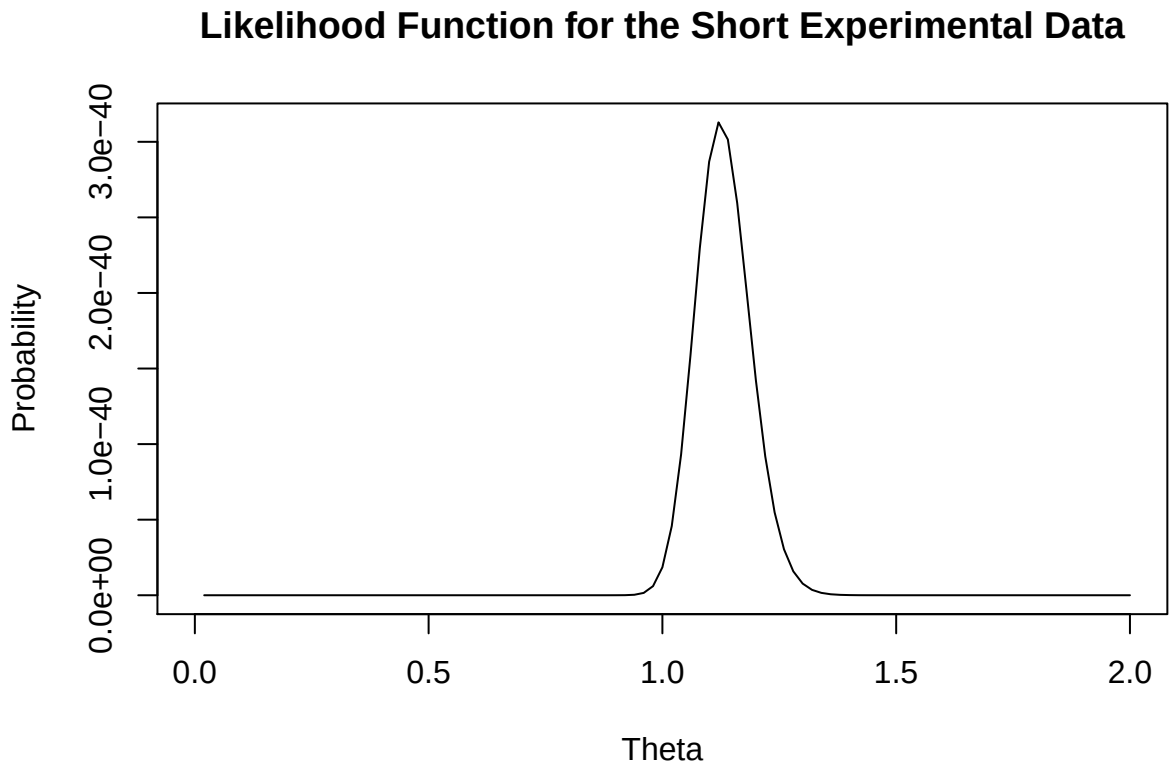
and the asymptotic variance of the method of moments estimator for θ is

$$Var(\hat{\theta}_{MoM}) \to \left(\frac{4}{\pi} - 1\right)\frac{\theta^2}{n}$$

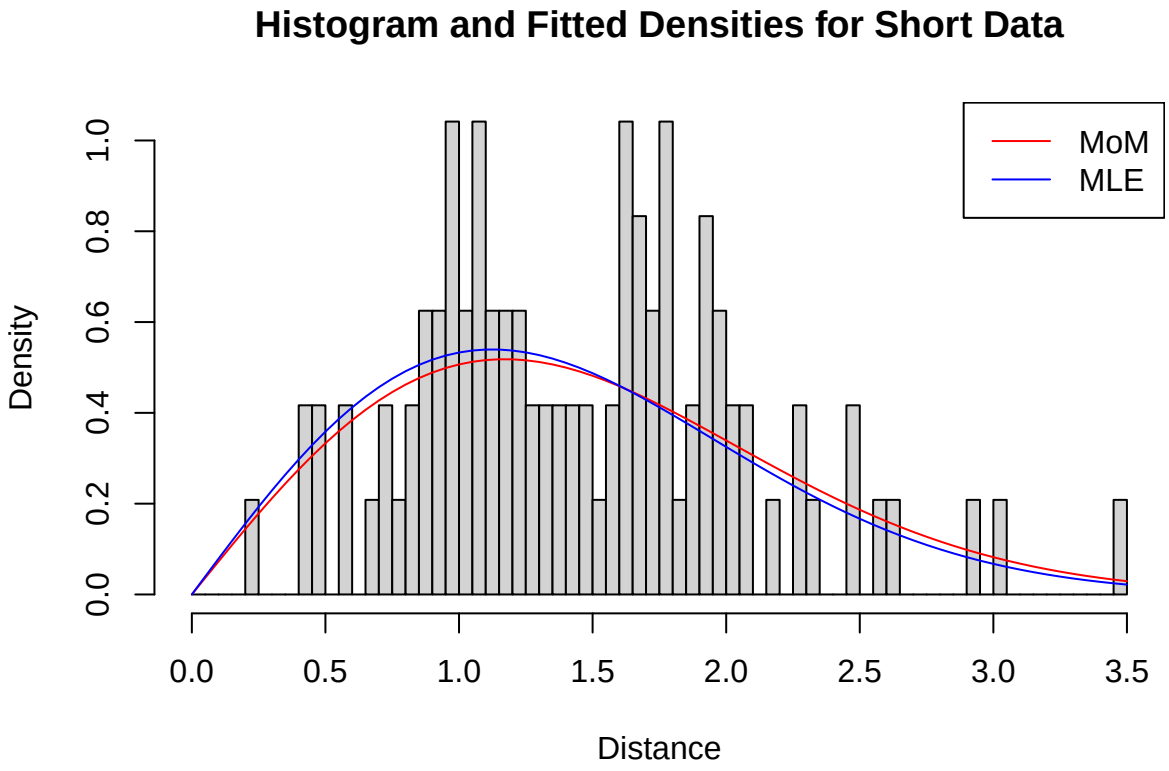
in which $\frac{4}{\pi} - 1 \approx .273295...$

Data Analysis

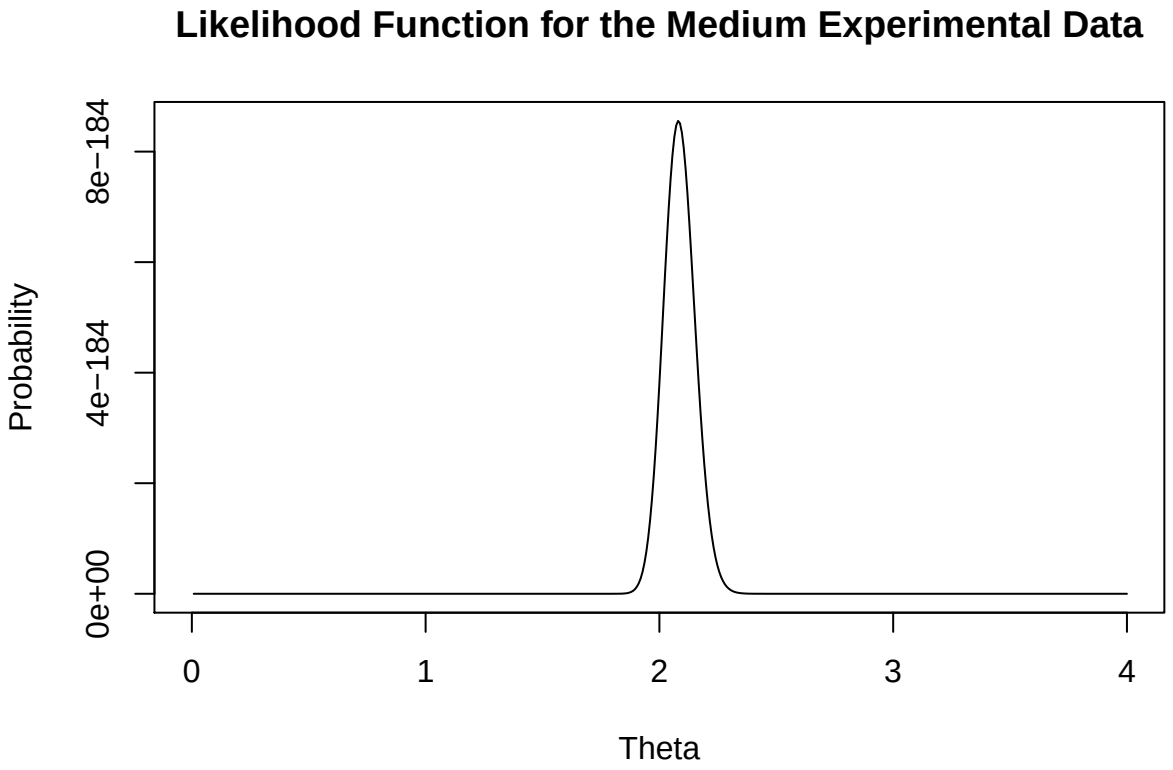
Short Experiment



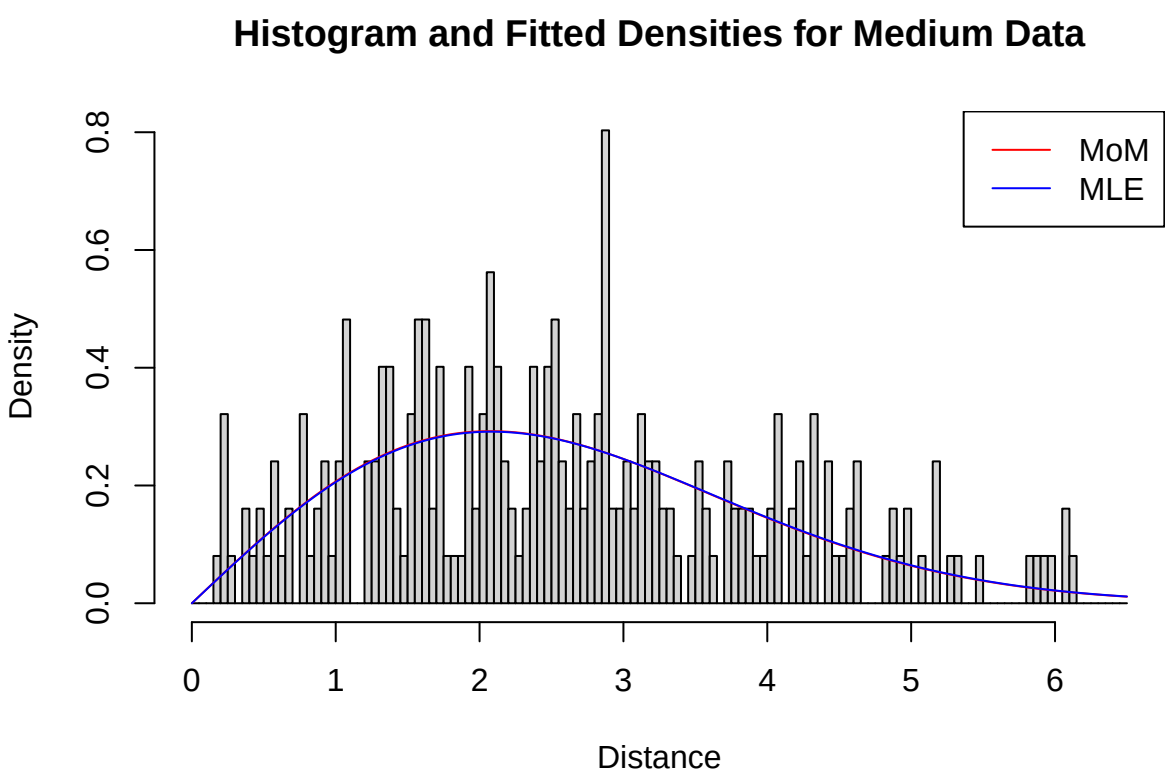
The MLE estimate is $\hat{\theta}_{MLE} = 1.1237776$, with an estimated variance of 0.0032887. The MoM estimate is $\hat{\theta}_{MoM} = 1.1705631$, with an estimated variance of 0.0039.



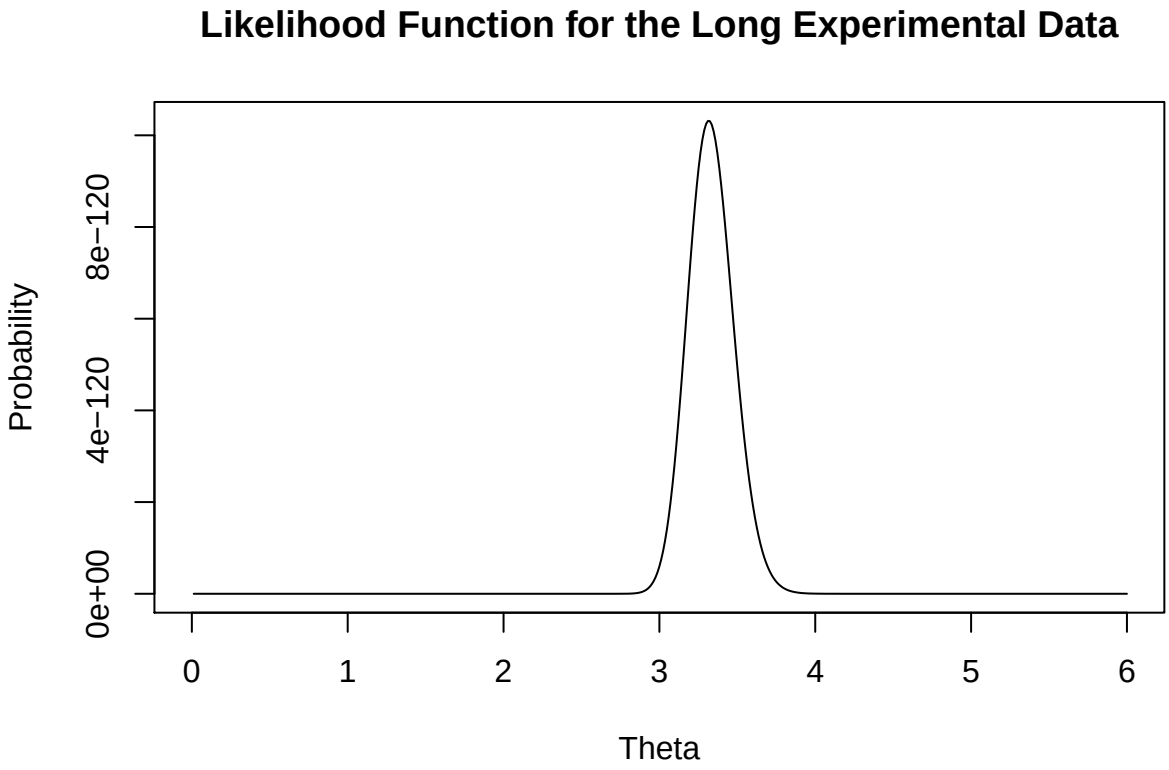
Medium Experiment



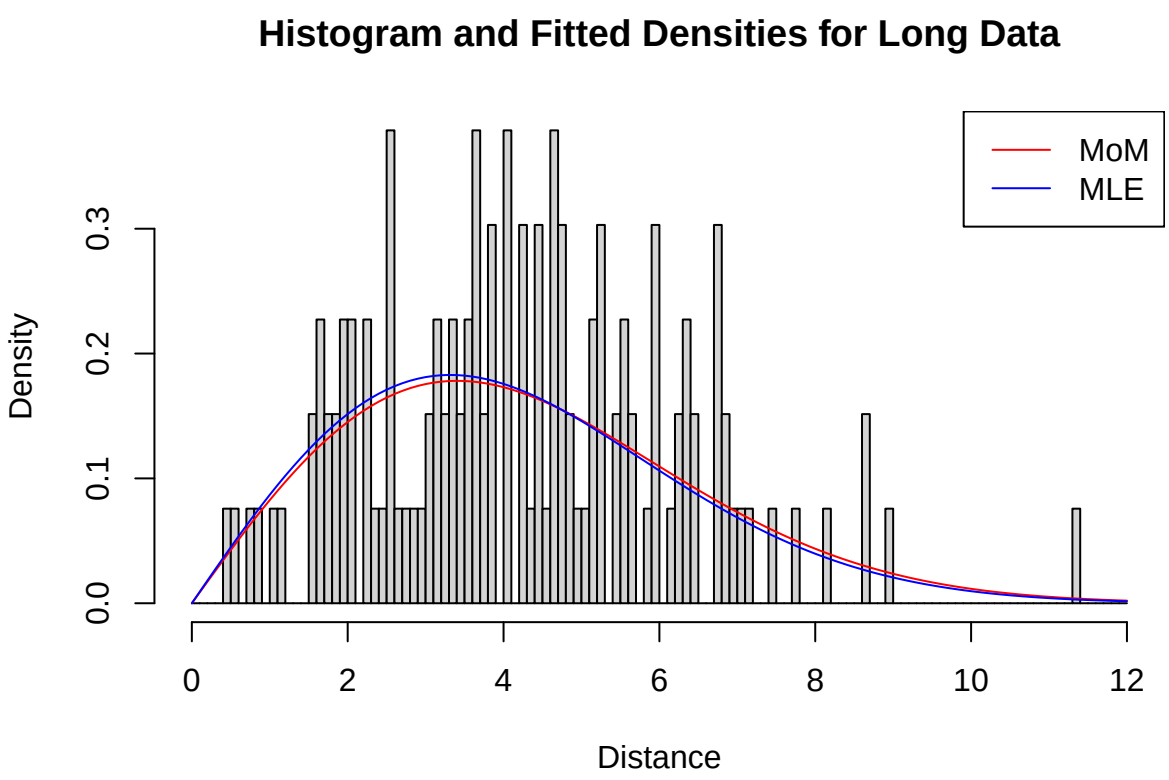
The MLE estimate is $\hat{\theta}_{MLE} = 2.0812133$, with an estimated variance of 0.0043488. The MoM estimate is $\hat{\theta}_{MoM} = 2.0748203$, with an estimated variance of 0.0047239.


Histogram and Fitted Densities for Medium Data
MoM
MLE
Density
0.0
0.2
0.4
0.6
0.8
0
1
2
3
4
5
6
Distance

Long Experiment



The MLE estimate is $\hat{\theta}_{MLE} = 3.3168463$, with an estimated variance of 0.0208361. The MoM estimate is $\hat{\theta}_{MoM} = 3.4044888$, with an estimated variance of 0.0239924.



Tentative Conclusions

We see there is little practical difference between the method of moments estimates and the maximum likelihood fits estimates for theta. The method of moments estimates are slightly larger, though. The larger the genomic separation, the larger the estimate for theta, indicating that the densities become flatter.

Estimating Theta via Bootstrap

For the medium experiment, we estimate the distribution of θ by bootstrap, and compare it to our results derived via asymptotic theory. As a reminder, the estimated value of $\hat{\theta}$ is 2.0812133, and the estimated variance using the asymptotic variance, $Var(\hat{\theta}_{MLE})$ is 0.0043488.

We will now show that if $X \sim Rayleigh(1)$, then $Y = \theta X$ is distributed $Y \sim Rayleigh(\theta)$. The density function of X is

$$f_X(x) = xe^{-\frac{x^2}{2}}, x \geq 0$$

Letting $Y = \theta X$,

$$P(Y \leq y) = P(\theta X \leq y) = P(X \leq \frac{y}{\theta})$$

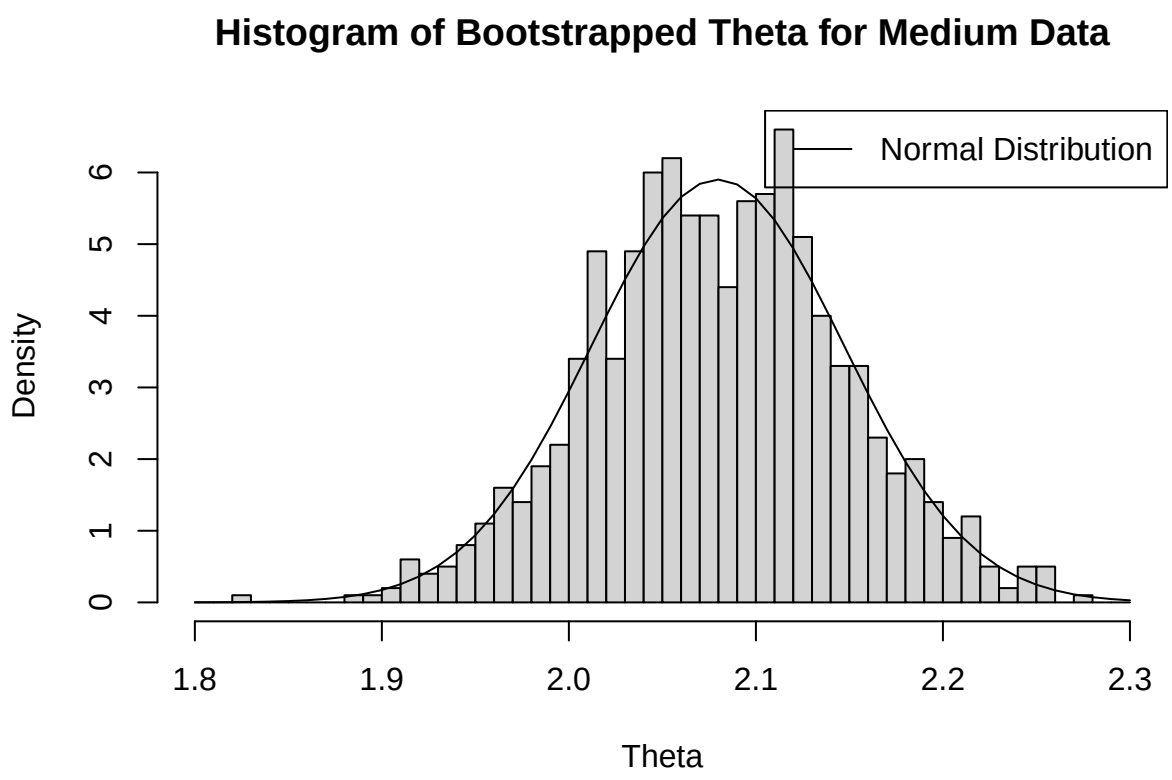
since $\theta > 0$. Deriving, $f_Y(y) = \frac{1}{\theta} f_X(\frac{y}{\theta})$ implies that

$$f_Y(y) = \frac{y}{\theta^2} e^{-\frac{y^2}{2\theta^2}}, y \geq 0$$

which is a Rayleigh variable with parameter θ. Thus, to generate $Rayleigh(\theta)$ random variables, it suffices to generate $Rayleigh(1)$ random variables.

We now find the inverse of the cumulative density function of a $Rayleigh(1)$ random variable. Integrating, $F_X(x) = 1 - e^{-\frac{x^2}{2}}$ when $x \geq 0$ and 0 otherwise. Taking the inverse, $F_X^{-1}(y) = \sqrt{-2\ln(1-y)}$ when $y \geq 0$ and 0 otherwise. Thus, to generate $Rayleigh(1)$ random variables, we can generate Uniform(0, 1) random variables and apply $F_X^{-1}(y)$ to them.

We first generate 1000 samples with size equal to the size of the medium experiment. For each sample, we estimate θ using method of maximum likelihood. By asymptotic theory, the bootstrapped θ's should be normally distributed.



The distribution of bootstrapped θ's looks roughly normal, though with a slight bimodal tendency. The standard deviation of the bootstrapped θ's is 0.0675994, while the standard error from asymptotic theory is 0.0659458. As expected, the two are very close.

We now generate a 95% confidence interval for $\hat{\theta}$. By bootstrap, we take the 25th and 975th percentiles of the estimated θ's, giving us a 95% confidence interval of (1.9471469, 2.2118549). The corresponding 95% confidence interval via asymptotic theory is (1.9519596, 2.210467). Again, the two are very close.