Lab 01 Report

Mudith Chathuranga Silva

4/7/2020

01.

Bernoulli ... again.
Let $y_1, ..., y_n|\theta \sim \text{Bern}(\theta)$, and assume that you have obtained a sample with $s = 5$ successes in $n = 20$ trials. Assume a $\text{Beta}(\alpha_0, \beta_0)$ prior for θ and let $\alpha_0 = \beta_0 = 2$.

(A)

(a) Draw random numbers from the posterior $\theta|y \sim \text{Beta}(\alpha_0 + s, \beta_0 + f)$, $y = (y_1, \dots, y_n)$, and verify graphically that the posterior mean and standard deviation converges to the true values as the number of random draws grows large.

Model :-

$$x_1, ..., x_n|\theta \overset{iid}{\sim} Bern(\theta)$$

Prior :-

$$\theta \sim Beta(\alpha_0, \beta_0)$$
$$\theta \sim Beta(2, 2)$$

Posterior :-

$$P(\theta|x_1, ..., x_n) \propto P(x_1, ..., x_n|\theta) * P(\theta)$$

$$P(\theta|x_1, ..., x_n) \propto \theta^s(1-\theta)^f * \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$P(\theta|x_1, ..., x_n) \propto \theta^{s+\alpha-1}(1-\theta)^{f+\beta-1}$$

$$\theta \propto Beta(\alpha, \beta) \overset{x_1,...,x_n}{\rightarrow} \theta|x_1, ..., x_n \sim Beta(\alpha + s, \beta + f)$$

$$\theta \propto Beta(2, 2) \overset{x_1,...,x_n}{\rightarrow} \theta|x_1, ..., x_n \sim Beta(2 + 5, 2 + 15)$$

$$\theta \propto Beta(2, 2) \overset{x_1,...,x_n}{\rightarrow} \theta|x_1, ..., x_n \sim Beta(7, 17)$$

```r
s = 5
f = 15
a0 = 2
b0 = 2

sample01 = rbeta(100,a0 + s, b0 + f)
sample02 = rbeta(1000,a0 + s, b0 + f)
sample03 = rbeta(10000,a0 + s, b0 + f)
sample04 = rbeta(100000,a0 + s, b0 + f)
```

True Mean and Variance for Beta Distribution is given by :-

$$E(\theta) = \frac{\alpha}{\alpha + \beta}$$

$$E(\theta) = \frac{7}{7 + 17}$$

$$E(\theta) = 0.2916$$

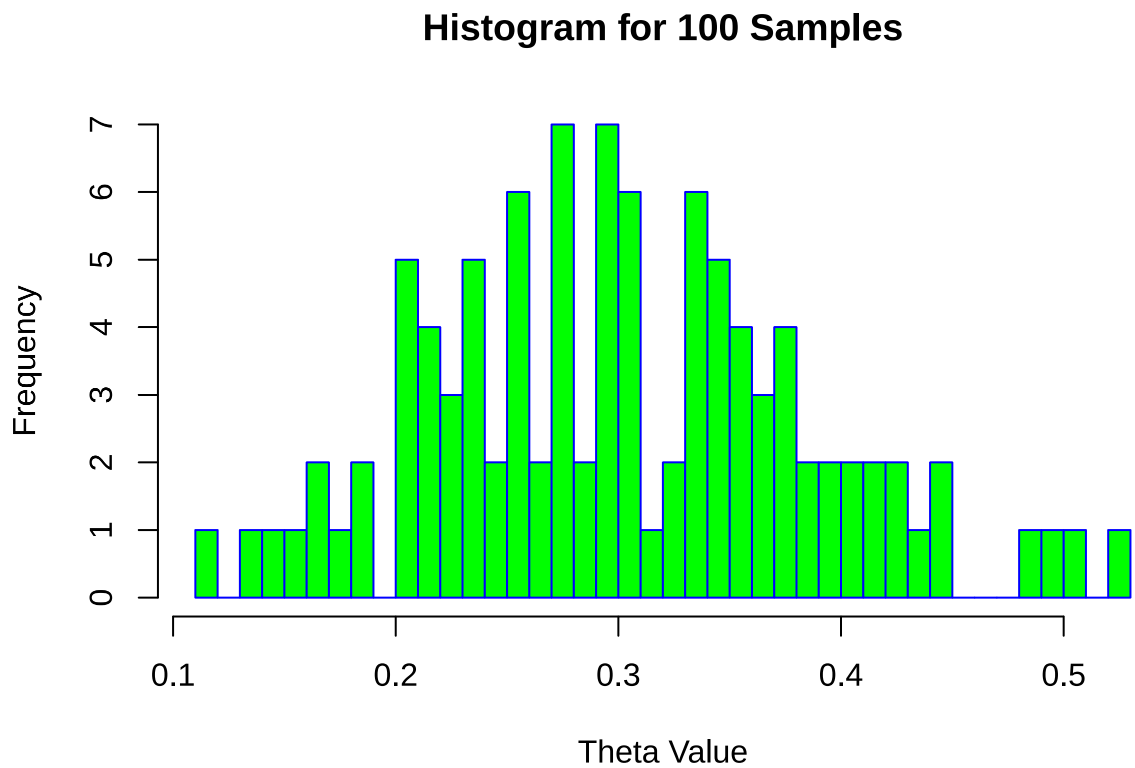
$$var(\theta) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

$$var(\theta) = \frac{7 * 17}{(7 + 17)^2(7 + 17 + 1)}$$

$$var(\theta) = \frac{7 * 17}{(7 + 17)^2(7 + 17 + 1)}$$

$$var(\theta) = 0.008263$$

```r
hist(sample01,
     main = "Histogram for 100 Samples",
     xlab = "Theta Value",
     border = "blue",
     col = "green",
     breaks = 30)
```



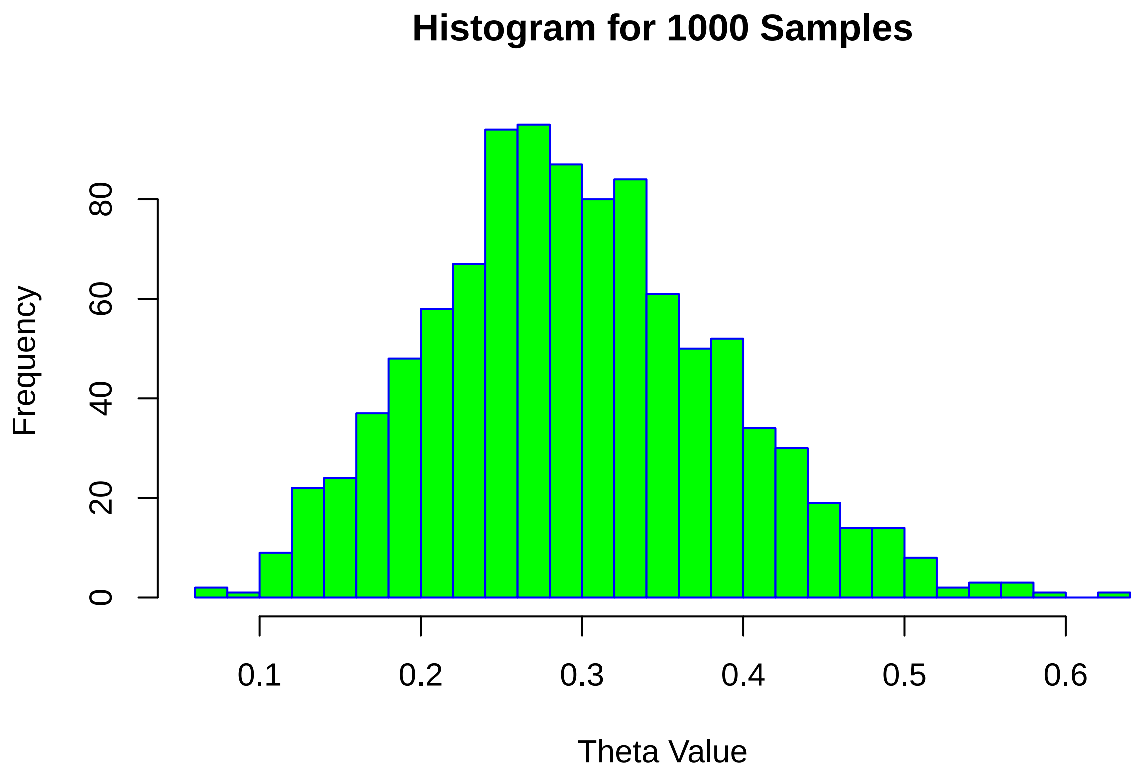

```r
cat('Mean for 100 samples:- ', mean(sample01), '\n')
```

```
## Mean for 100 samples:-  0.3019287
```

```r
cat('Variance for 100 samples:- ', var(sample01), '\n')
```

```
## Variance for 100 samples:-  0.007292786
```

```r
hist(sample02,
     main = "Histogram for 1000 Samples",
     xlab = "Theta Value",
     border = "blue",
     col = "green",
     breaks = 30)
```



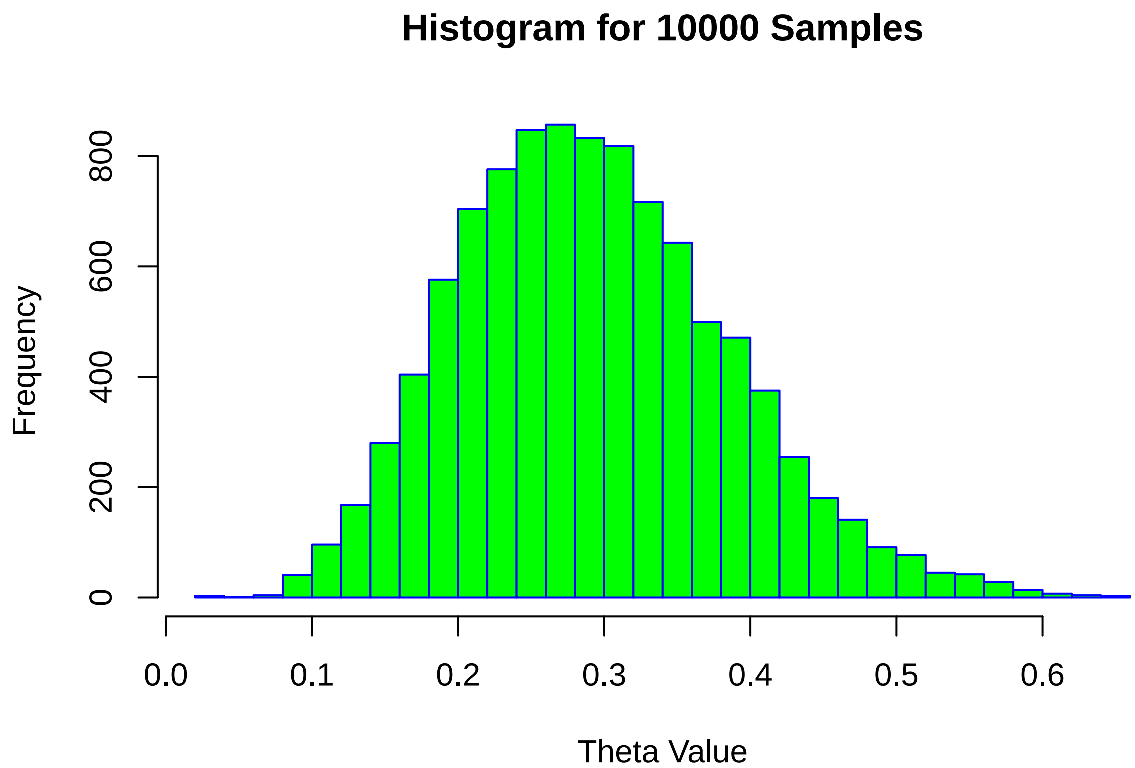

```r
cat('Mean for 1000 samples:- ', mean(sample02), '\n')
```

```
## Mean for 1000 samples:-  0.2959214
```

```r
cat('Variance for 1000 samples:- ', var(sample02), '\n')
```

```
## Variance for 1000 samples:-  0.008201021
```

```r
hist(sample03,
     main = "Histogram for 10000 Samples",
     xlab = "Theta Value",
     border = "blue",
     col = "green",
     breaks = 30)
```



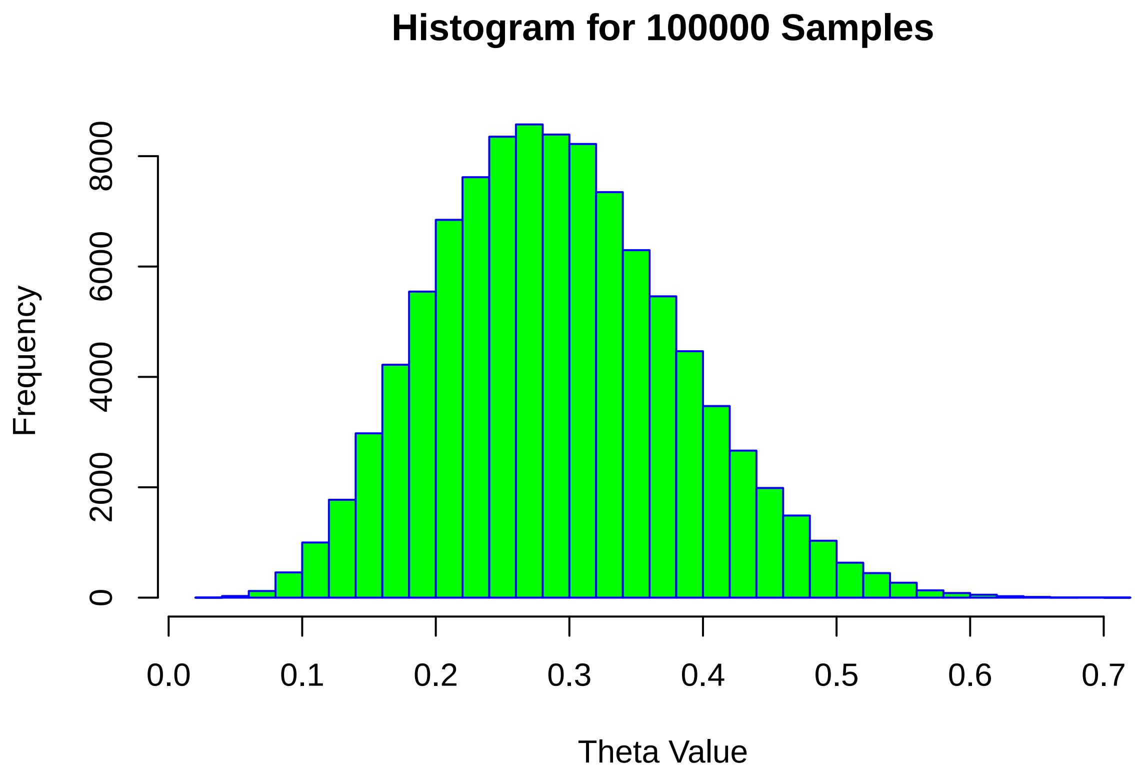

```r
cat('Mean for 10000 samples:- ', mean(sample03), '\n')
```

```
## Mean for 10000 samples:-  0.2925951
```

```r
cat('Variance for 10000 samples:- ', var(sample03), '\n')
```

```
## Variance for 10000 samples:-  0.008420138
```

```r
hist(sample04,
     main = "Histogram for 100000 Samples",
     xlab = "Theta Value",
     border = "blue",
     col = "green",
     breaks = 30)
```




```r
cat('Mean for 100000 samples:- ', mean(sample04), '\n')
```

```
## Mean for 100000 samples:-  0.2913501
```

```r
cat('Variance for 100000 samples:- ', var(sample04), '\n')
```

```
## Variance for 100000 samples:-  0.008254777
```

It's clear that Posterior Mean and Variance Converges to the true Mean and Variance when the number of random draw grows larger.

(B)

(b) Use simulation (`nDraws = 10000`) to compute the posterior probability $\Pr(\theta > 0.3|y)$ and compare with the exact value [Hint: `pbeta()`].

```r
sampleProb = mean(sample03 > 0.3)
trueProb = 1 - pbeta(0.3,7,17) # pbeta gives p(theta < 0.3)

cat('Sample Probability :- ', sampleProb, '\n')
```

```
## Sample Probability :-  0.441
```

```r
cat('True Probability :- ', trueProb, '\n')
```

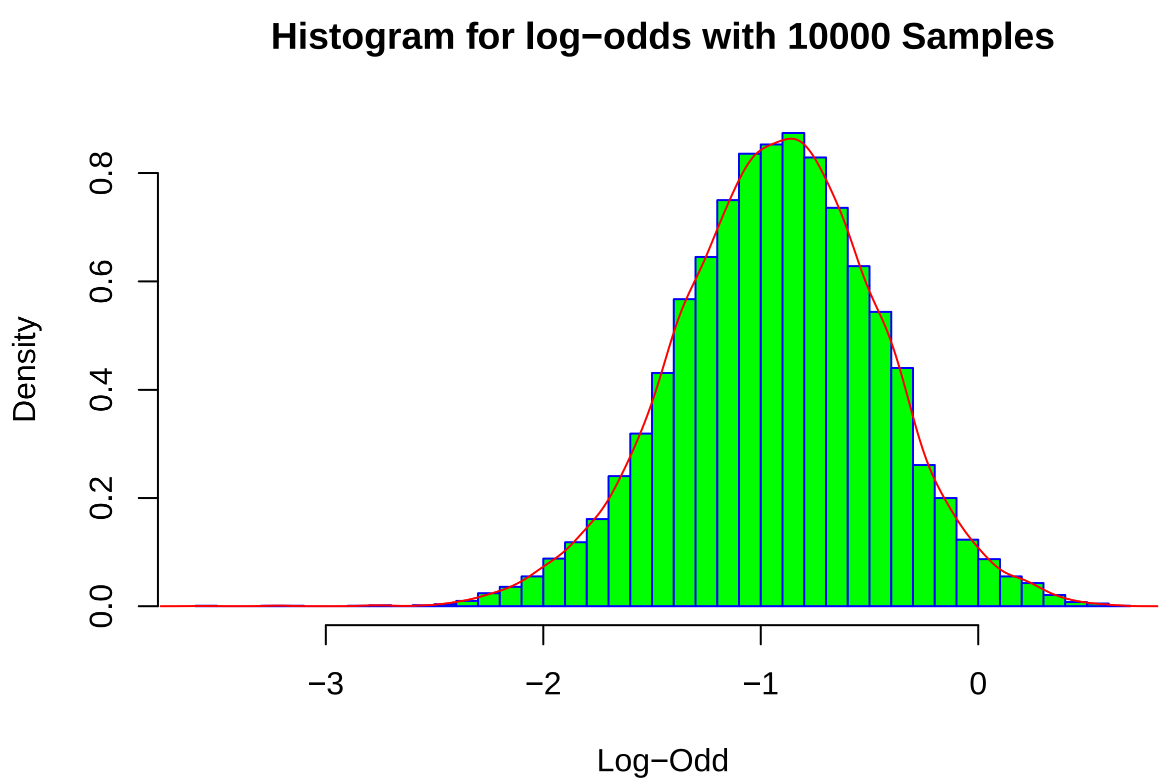
```
## True Probability :-  0.4399472
```

True $Pr(\theta > 0.3|y) = 0.4399472$ is so close to the Sample $Pr(\theta > 0.3|y) = 0.441$ for 10000 samples

(C)

(c) Compute the posterior distribution of the log-odds $\phi = \log \frac{\theta}{1-\theta}$ by simulation (`nDraws = 10000`). [Hint: `hist()` and `density()` might come in handy]

```r
log_odds = log(sample03 / (1 - sample03))
hist(log_odds,
     main = "Histogram for log-odds with 10000 Samples",
     xlab = "Log-Odd",
     border = "blue",
     col = "green",
     breaks = 30,
     probability = T)
lines(density(log_odds),
      lwd = 1,
      col = "red")
```



02.

Log-normal distribution and the Gini coefficient.
Assume that you have asked 10 randomly selected persons about their monthly income (in thousands Swedish Krona) and obtained the following ten observations: 44, 25, 45, 52, 30, 63, 19, 50, 34 and 67. A common model for non-negative continuous variables is the log-normal distribution. The log-normal distribution $\log \mathcal{N}(\mu, \sigma^2)$ has density function

$$p(y|\mu, \sigma^2) = \frac{1}{y \cdot \sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(\log y - \mu)^2\right],$$

for $y > 0$, $\mu > 0$ and $\sigma^2 > 0$. The log-normal distribution is related to the normal distribution as follows: if $y \sim \log \mathcal{N}(\mu, \sigma^2)$ then $\log y \sim \mathcal{N}(\mu, \sigma^2)$. Let $y_1, ..., y_n|\mu, \sigma^2 \overset{iid}{\sim} \log \mathcal{N}(\mu, \sigma^2)$, where $\mu = 3.7$ is assumed to be known but σ^2 is unknown with non-informative prior $p(\sigma^2) \propto 1/\sigma^2$. The posterior for σ^2 is the $Inv - \chi^2(n, \tau^2)$ distribution, where

$$\tau^2 = \frac{\sum_{i=1}^{n}(\log y_i - \mu)^2}{n}.$$

First let's check how the we do the calculation for the $P(\sigma^2|y)$

Model :-

$$y_1, ..., y_n|\mu, \sigma^2 \overset{iid}{\sim} logN(\mu, \sigma^2), where \mu = 3.7$$

Prior :-

$$P(\sigma^2) \propto \frac{1}{\sigma^2}$$

Posterior :-

$$P(\sigma^2|y) \propto P(y|\mu = 3.5, \sigma^2) * P(\sigma^2)$$

$$P(\sigma^2|y) \propto \prod_{i=1}^{n} \frac{1}{y_i\sqrt{2\pi\sigma^2}} \exp[-\frac{1}{2\sigma^2}(logy_i - \mu)^2] * \frac{1}{\sigma^2}$$

$$P(\sigma^2|y) \propto \frac{1}{\sigma^{n+2}} \exp[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(logy_i - \mu)^2]$$

Now Let's consider the Scaled $Inv - \chi^2$ Distribution.

$$P(\beta|v, s^2) = \frac{(v/2)^{v/2}}{\Gamma(v/2)} s^v \beta^{-(v/2+1)} \exp[\frac{-vs^2}{2\beta}]$$

$$P(\beta|v, s^2) \propto \beta^{-(v/2+1)} \exp[\frac{-vs^2}{2\beta}]$$

Now Let's restructure our Posterior;

$$P(\sigma^2|y) \propto \sigma^{2-(\frac{n}{2}+1)} \exp[-\frac{n}{2\sigma^2}\frac{\sum_{i=1}^{n}(logy_i - \mu)^2}{n}]$$

since $\frac{\sum_{i=1}^{n}(logy_i-\mu)^2}{n}$ is a constant. Let use 's^2' for that term. Then,

$$P(\sigma^2|y) \propto \sigma^{2-(\frac{n}{2}+1)} \exp[-\frac{ns^2}{2\sigma^2}]$$

Now We could express the posterior for σ^2 as the Scaled $Inv - \chi^2$ Distribution.

$$P(\sigma^2|y) \propto ScaledInv - \chi^2(n, s^2)$$

Where $s^2 = \frac{\sum_{i=1}^{n}(logy_i-\mu)^2}{n}$

(A)

(a) Simulate 10,000 draws from the posterior of σ^2 (assuming $\mu = 3.7$) and compare with the theoretical $Inv - \chi^2(n, \tau^2)$ posterior distribution.

Steps to Simulate 10,000 Draws from Posterior of σ^2 :-

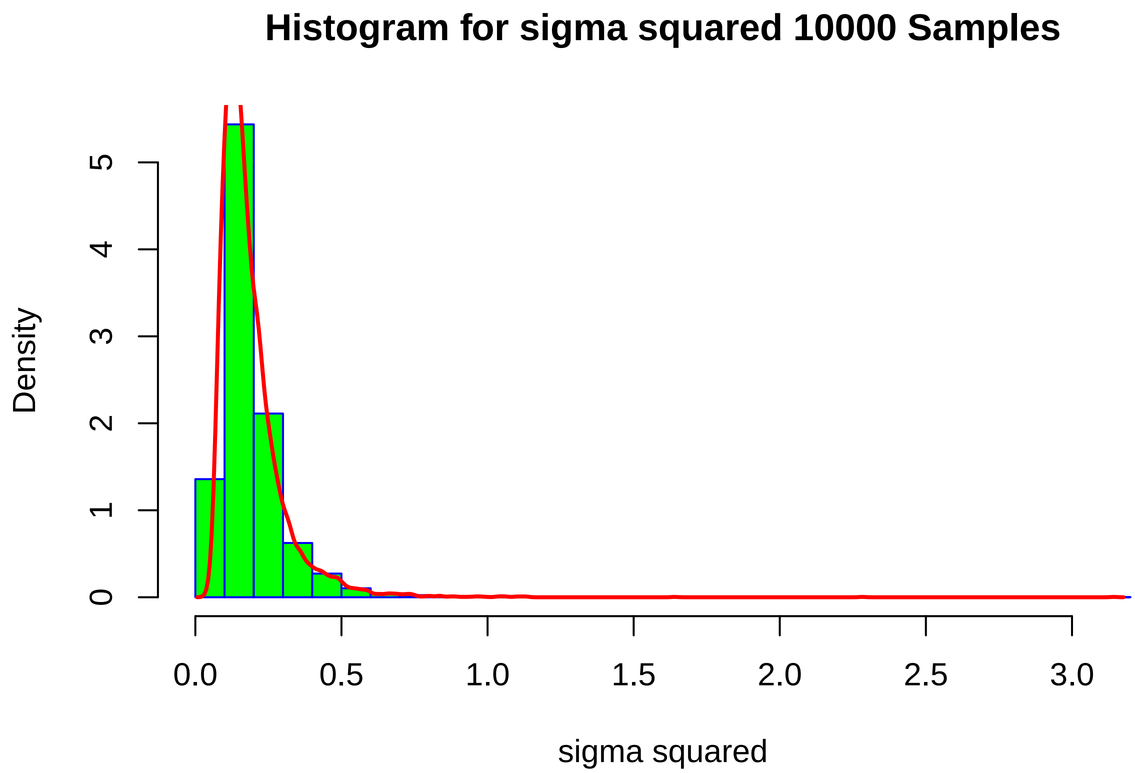
1. Draw $X \sim \chi^2(n)$ [rchisq()]
2. Compute $\sigma^2 = \frac{ns^2}{X}$. This is the draw from $Inv - \chi^2(n, s^2)$

n = 10 (Given Observations)

```
y_data = c(44,25,45,52,30,63,19,50,34,67)
mu = 3.7

s_squared = sum((log(y_data) - mu)**2) / length(y_data)
x_points = rchisq(10000,length(y_data))
sigma_squared = (length(y_data) * s_squared) / x_points

hist(sigma_squared,
     main = "Histogram for sigma squared 10000 Samples",
     xlab = "sigma squared",
     border = "blue",
     col = "green",
     probability = T,
     breaks = 30)
lines(density(sigma_squared),
      lwd = 2,
      col = "red")
```


Histogram for sigma squared 10000 Samples
Density
0
1
2
3
4
5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
sigma squared