STAT201A – Sec. 102
Homework #8.

Steven Pollack
24112977

#1.

Proof. Given a generic prediction, $W = g(X)$, and setting $h(X) = E(Y \mid X) - W$, the MSE of W is $E[(Y-W)^2]$ and can be rewritten as

$$\begin{aligned}
E\left[(Y-W)^2\right] &= E\left(E\left[(Y-W)^2 \mid X\right]\right) \\
&= E\left(E\left[\{(Y - E(Y \mid X)) + h(X)\}^2 \mid X\right]\right) \\
&= E\left(E\left[(Y - E(Y \mid X))^2 \mid X\right]\right) + 2E\left(E\left[(Y - E(Y \mid X))h(X) \mid X\right]\right) + E\left(E\left[h(X)^2 \mid X\right]\right) \\
&= \operatorname{var}(Y \mid X) + 2E\left(h(X)\underbrace{E\left[Y - E(Y \mid X) \mid X\right]}_{=0}\right) + E\left[h(X)^2\right] \\
&= \operatorname{var}(Y \mid X) + E\left[h(X)^2\right]
\end{aligned}$$

Hence, the MSE of Y and W is minimized when $h(X) = 0 \iff g(X) = E(Y \mid X)$. □

#2.

Proof. Let I_1 indicate the event that a widget of the first kind is drawn, $I_2 = 1 - I_1$ indicate for a widget of the second kind and set Y to be the random variable that identifies the drawn widget. It follows that

$$E(Y \mid I_1) = \mu I_1 + \nu I_2$$

and

$$\operatorname{var}(Y \mid I_1) = \sigma^2 I_1 + \tau^2 I_2$$

Thus, the law of iterated expectations says that

$$E(Y) = E(E(Y \mid I_1)) = \frac{\mu + 2\nu}{3}$$

and our iterated variance rule shows that

$$\begin{aligned}
\operatorname{var}(Y) &= E(\operatorname{var}(Y \mid I_1)) + \operatorname{var}(E(Y \mid I_1)) \\
&= E\left(\sigma^2 I_1 + \tau^2 I_2\right) + \operatorname{var}\left(\mu I_1 + \nu I_2\right) \\
&= \frac{\sigma^2 + 2\tau^2}{3} + \mu^2 \operatorname{var}(I_1) + \nu^2 \operatorname{var}(I_2) + 2\mu\nu \operatorname{cov}(I_1, I_2) \\
&= \frac{3(\sigma^2 + 2\tau^2) + 2(\mu - \nu)^2}{9}
\end{aligned}$$

□

#3. Consider $B \sim \text{HypGeo}(b_0 + w_0, b_0, d)$, and β the number of black balls in a sample of size n. Then $\beta \sim \text{HypGeo}(N, b + B, n)$, where $N = d + b + w$. Then, $E(\beta \mid B) = n\frac{b+B}{N}$ and

$$\begin{aligned} E(\beta) &= E(E(\beta \mid B)) \\ &= \frac{n}{N}E(b+B) \\ &= \frac{n}{N}(b + E(B)) \\ &= \frac{n}{N}\left(b + d\frac{b_0}{b_0+w_0}\right) \end{aligned}$$

and

$$\begin{aligned} \text{var}(\beta) &= \text{var}(E(\beta \mid B)) + E(\text{var}(\beta \mid B)) \\ &= \text{var}\left(n\frac{b+B}{N}\right) + E\left(n\frac{b+B}{N}\frac{d+w-B}{N}\frac{N-n}{N-1}\right) \\ &= \frac{n^2}{N^2}\,\text{var}(b+B) + \frac{n}{N^2}\frac{N-n}{N-1}E((b+B)(d+w-B)) \\ &= \frac{n^2}{N^2}\,\text{var}(B) + \frac{n}{N^2}\frac{N-n}{N-1}E(bd + (d+w-b)B - B^2) \\ &= \frac{n^2}{N^2}\,\text{var}(B) + \frac{n}{N^2}\frac{N-n}{N-1}\left(bd + (d+w-b)E(B) - E(B^2)\right) \\ &= \frac{n^2}{N^2}\,\text{var}(B) + \frac{n}{N^2}\frac{N-n}{N-1}\left(bd + (d+w-b)E(B) - \text{var}(B) - E(B)^2\right) \\ &= \frac{n}{N^2}\left(n - \frac{N-n}{N-1}\right)\text{var}(B) + \frac{n}{N^2}\frac{N-n}{N-1}\left(bd + (d+w-b)E(B) - E(B)^2\right) \\ &= \frac{n(n-1)}{N(N-1)}\,\text{var}(B) + \frac{n}{N^2}\frac{N-n}{N-1}\left(bd + (d+w-b)E(B) - E(B)^2\right) \end{aligned}$$

where

$$\text{var}(B) = d\left(\frac{b_0}{b_0+w_0}\right)\left(\frac{w_0}{b_0+w_0}\right)\left(\frac{b_0+w_0-d}{b_0+w_0-1}\right)$$

#4.

1. Given $X, X_1, X_2, \ldots \overset{iid}{\sim} F_X$, where X_i have MGF ψ_X,

$$\begin{aligned} \psi_S(t) &= E\left[e^{tS}\right] \\ &= E\left[E\left(e^{t\sum_{i=1}^N X_i}\middle| N\right)\right] \\ &= E\left[\psi_X(t)^N\right] \\ &= E\left[\exp\left\{N\log(\psi_X(t))\right\}\right] \\ &= \psi_N(\log(\psi_X(t)) \end{aligned}$$

where the third equality comes from the fact that $E\left(\exp\left\{t\sum_{i=1}^N X_i\right\} \mid N = n\right) = \psi_X(t)^n$.

2. Given $N \sim$ Poisson(λ):

$$
\begin{aligned}
\psi_N(t) &= E(e^{tN}) \\
&= \sum_{n=0}^{\infty} e^{tn} e^{-\lambda} \frac{\lambda^n}{n!} \\
&= e^{-\lambda} \sum_{n=0}^{\infty} \frac{(e^t \lambda)^n}{n!} \\
&= e^{-\lambda} e^{e^t \lambda} \\
&= \exp\left\{\lambda(e^t - 1)\right\}
\end{aligned}
$$

3. For I an indicator random variable with $P(I = 1) = p$, the mgf of I is:

$$\psi_I(t) = 1 + p(e^t - 1)$$

Putting this all together, if we toss a coin $N \sim$ Poisson(λ) times, where each head has probability p, then X, the number of heads in N tosses looks like $\sum_{i=1}^{N} I_i$, where I_i indicates a heads on the i^{th} toss. Thus,

$$
\begin{aligned}
\psi_X(t) &= \psi_N(\log(\psi_{I_1}(t))) \\
&= \exp\{\lambda(\exp\{\log \psi_{I_1}(t)\} - 1)\} \\
&= \exp\{\lambda(\psi_{I_1}(t) - 1)\} \\
&= \exp\left\{\lambda p(e^t - 1)\right\}
\end{aligned}
$$

which shows that $X \sim$ Poisson(λp).

#5.

1. Let $\Theta \sim \beta$eta(r, s) be the probability of getting a heads (endowed with a $\beta(r, s)$ prior density), and note that our experiment follows $X \sim$ Geo(Θ). Hence, our likelihood is $P(X = k \,|\, \Theta \in d\theta) = (1 - d\theta)^{k-1} d\theta$ and $P(\Theta \in d\theta) \propto d\theta^{r-1}(1 - d\theta)^{s-1}$. Using the formula, posterior $\propto$ likelihood $\times$ prior, we have

$$
\begin{aligned}
P(\Theta \in d\theta \,|\, X = k) &\propto P(X = k \,|\, \Theta \in d\theta) \cdot P(\Theta \in d\theta) \\
&\propto d\theta^r (1 - d\theta)^{s+k-2}
\end{aligned}
$$

Which means our posterior density is βeta($r+1, s+k-1$), making the beta densities are a family of conjugate priors, here.

2. For a fixed r, s we see that increasing k puts more mass to the left of 1/2, which makes sense: a heavier "head" implies a lower probability of flipping heads, and hence a longer waiting time.

#6.

1. Set $f_\rho(u,v) = \rho u + \sqrt{1-\rho^2}v$, and note that for $-1 \leq \rho \leq 1$, and $U, V \sim$ N(0,1) with $U \perp\!\!\!\perp V$, $(U, f(U,V))$ is distributed according to the standard bivariate normal distribution with correlation ρ. Thus, if we take $X \sim$ N(μ_X, σ_X^2), and $V \sim$ N(μ_Y, σ_Y^2), $X \perp\!\!\!\perp V$, then

$$\left(\frac{X-\mu_X}{\sigma_X}, f_\rho\left(\frac{X-\mu_X}{\sigma_X}, \frac{V-\mu_Y}{\sigma_Y}\right)\right) \sim N\left(\mu = (0,0), \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

Hence, if we perform the standard shift and scale, we'll get

$$\left(X, \sigma_Y f_\rho\left(\frac{X-\mu_X}{\sigma_X}, \frac{V-\mu_Y}{\sigma_Y}\right) + \mu_Y\right) \sim N\left(\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_X & 0 \\ 0 & \sigma_Y \end{pmatrix}\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\begin{pmatrix} \sigma_X & 0 \\ 0 & \sigma_Y \end{pmatrix}\right)$$

That is, if $Y = \sigma_Y f_\rho\left(\frac{X-\mu_X}{\sigma_X}, \frac{V-\mu_Y}{\sigma_Y}\right) + \mu_Y$, then

$$(X,Y) \sim N\left(\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}\right)$$

So, putting this machinery to good use, let $X = MSAT$, with $\mu_X = 500$, $\sigma_X = 90$, and $Y = VSAT$ with $\mu_Y = 480$ and $\sigma_Y = 100$ and suppose the correlation between X and Y is $\rho = 0.5$. Then, the above construction shows that

$$\begin{aligned}
P(MSAT > VSAT) &= P(X > \sigma_Y f_\rho(X^*, V^*) + \mu_Y) \\
&= P(X > \sigma_Y \rho X^* + \sigma_Y\sqrt{1-\rho^2}V^* + \mu_Y) \\
&= P\left(\frac{X - \sigma_Y\rho X^* - \mu_Y}{\sigma_Y\sqrt{1-\rho^2}} > V^*\right) \\
&= \int_{x=-\infty}^{\infty}\int_{v=-\infty}^{g(x,\rho,\sigma_X,\mu_Y,\sigma_Y)} f_{X,V}(x,y)\,dv\,dx \\
&= \int_{x=-\infty}^{\infty} f_X(x)\int_{v=-\infty}^{g(x,\rho,\sigma_X,\mu_Y,\sigma_Y)} f_{V^*}(v)\,dv\,dx \\
&= \int_{x=-\infty}^{\infty} f_X(x)\Phi\left(\frac{x - \sigma_Y\rho\sigma_X^{-1}(x-\mu_X) - \mu_Y}{\sigma_Y\sqrt{1-\rho^2}}\right)\,dx
\end{aligned}$$

Where the fifth equation is justified since $X \perp\!\!\!\perp V$, by assumption and thus $X \perp\!\!\!\perp V^*$. Plugging in the appropriate parameters in the equation and having R run the integral, we find

$$P(MSAT > VSAT) \approx 0.5830323$$

2. Using the setup above, we have that for $MSAT = 550$, our random vector, $(MSAT, VSAT)$, looks like

$$(550, \sigma_Y f_\rho(550^*, V^*) + \mu_Y)$$

which has normal distribution with mean = 507.7778 and sd = 86.6025. Thus,

$$\begin{aligned} P(MSAT > VSAT \mid MSAT = 550) &= P(550 > VSAT \mid MSAT = 550) \\ &= \Phi\left(\frac{550 - 507.7778}{86.6025}\right) \\ &\approx 0.687062 \end{aligned}$$

#7.

Proof. 1. Consider the experiment where we have two particles of type x and y and their frequency of appearance follows $X \sim$ Exp(λ) and $Y \sim$ Exp(μ) for types x and y, respectively. If I indicates on the event that a particle of type y appearance first, then $E(I) = P(X > Y)$. But, using the tower rule

$$E(I) = E[E(I \mid Y)] = E[e^{-\lambda Y}] = \psi_Y(-\lambda) = \left(1 + \frac{\lambda}{\mu}\right)^{-1} = \frac{\mu}{\mu + \lambda}$$

since

$$E(I \mid Y \in (y - dy, y + dy)) = P(X > y) = \int_y^\infty \lambda e^{-\lambda t}\, dt = e^{-\lambda y}$$

and $\psi_Y(t) = (1 - t/\mu)^{-1}$. This answer still makes sense for $\mu = \lambda$, since in this instance, both particles have the same frequency of occurrence, so it's 50/50 that you'll get one over the other.

2. From the method of MGF's we have that $\psi_{cY}(t) = \psi_Y(ct) = (1 - t/(\mu/c))^{-1}$, hence $cY \sim$ Exp(μ/c), and performing the entire argument above with $Y' = cY$, we have

$$P(X > cY) = \frac{\mu/c}{\mu/c + \lambda} = \frac{\mu}{\mu + c\lambda}$$

3. From b)

$$P(X > cY) = P(X/Y > c) = \frac{\mu}{\mu + c\lambda} \iff F_{X/Y}(c) = 1 - \frac{\mu}{\mu + c\lambda} = \frac{c\lambda}{\mu + c\lambda}$$

4. To find the medium, we want m such that $F_{X/Y}(m) = 50\%$. So,

$$F_{X/Y}(m) = \frac{1}{2} \iff \frac{m\lambda}{\mu + m\lambda} = \frac{1}{2} \iff 2m\lambda = \mu + m\lambda \iff m = \frac{\mu}{\lambda}$$

Since the median of Y is $m_Y = \ln(2)/\mu$, we have that $m = m_X/m_Y$. Also, $E(Y) = \mu^{-1}$ shows that $m = E(X)/E(Y)$.

5. It's a trap! If you differentiate $F_{X/Y}$ with respect to c, you recover the density of X/Y:

$$f_{X/Y}(c) = \frac{\lambda\mu}{(\mu + \lambda c)^2} = \Theta\left(\frac{1}{c^2}\right)$$

Hence, $E(X/Y) = \int_0^\infty c f_{X/Y}(c)\, dc = +\infty$, since $c f_{X/Y}(c) = \Theta(c^{-1})$. Thus, X/Y has a median but no density!

□

#8.

Proof. 1. For $X \sim \Gamma\text{amma}(r, \lambda)$, $\mu_X = r/\lambda$ and $\sigma_X^2 = r/\lambda^2$, hence

$$r = \mu_X \lambda = \sigma_X^2 \lambda^2 \iff r = \frac{\mu_X^2}{\sigma_X^2}$$

So, we may as well try estimating r via $\hat{r} = \hat{\mu}_1^2/(\hat{\mu}_2 - \hat{\mu}_1^2)$. My rationale for this is that $\hat{\mu}_1 \approx \mu_X$ and $\hat{\mu}_2 - \hat{\mu}_1^2 \approx E(X^2) - E(X)^2 = \sigma_X^2$ for n large, so hopefully the limit of their ratio will tend to $\mu_X/\sigma_X^2 = r$. By this same rationale, we might try using $\hat{\lambda} = \hat{\mu}_1/(\hat{\mu}_2 - \hat{\mu}_1^2)$.

2. Using the following code

```
set.seed(112)

r <- round(runif(n=1,min=2,max=5),digits=2) # r = 3.13
lambda <- round(runif(n=1,min=1,max=2),digits=2) # lambda = 1.92

n <- 100
X <- rgamma(n=n,shape=r,rate=lambda)
mu_hat_1 <- (1/n)*sum(X^1)
mu_hat_2 <- (1/n)*sum(X^2)

r_hat <-  (mu_hat_1^2)/(mu_hat_2 - mu_hat_1^2) # r_hat = 4.097228
print(abs(r-r_hat))

[1] 0.9672

lambda_hat <- mu_hat_1/(mu_hat_2 - mu_hat_1^2) # lambda_hat = 2.603451
print(abs(lambda-lambda_hat))

[1] 0.6835
```

and

```
p1 <- ggplot(data.frame(x=c(0,5))) +
  geom_histogram(aes(x = X, y = ..density..), alpha = 0.25,
                 binwidth=0.3333) +
  stat_function(fun=dgamma, arg=list(shape=r,rate=lambda),
                colour="darkgreen")
print(p1)
```

We can see that the histogram in figure 1 resembles the density of our sample distribution (though, it could stand to resemble it more closely); however, I'm remiss to say my estimators are not very close to their targets...

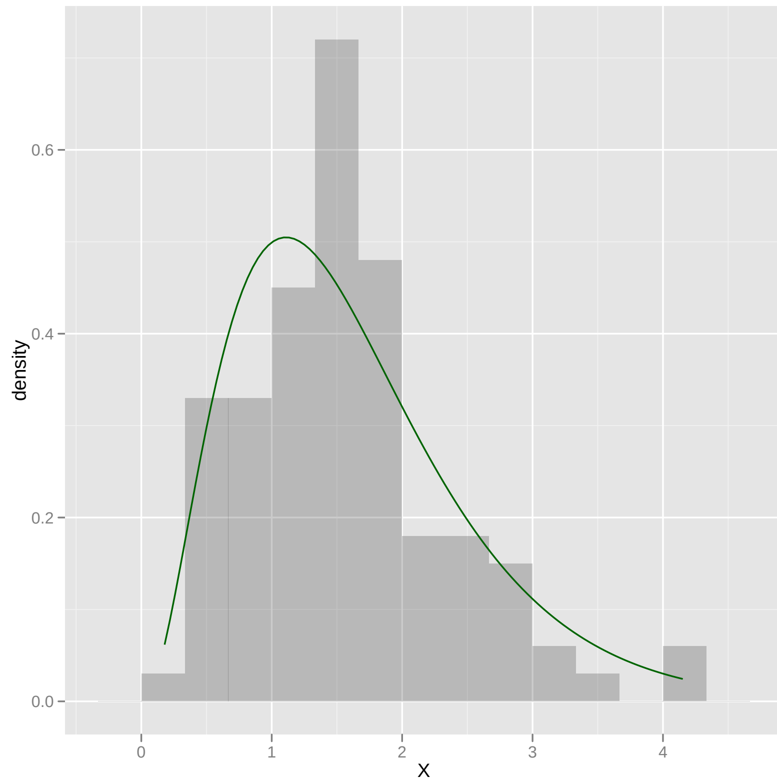

Figure 1: Histogram of $n = 100$ iid samples from Γ(r=3.13, λ=1.92) with density overlayed.

3. The mean our 1000 estimations of r and λ are 3.2619 and 2.0066, respectively. The SD of our the 1000 estimations is 0.5382 and 0.349. See figure 2 for histograms of the estimators.

□

#9. Using the following code

```
n <- 1e3
times <- 1e3

new_estimators <- t(mapply(FUN=generateEstimators,
                           n=rep(n,times=times),
                           r=rep(r,times=times),
                           lambda=rep(lambda,times=times)))

r_hat <- new_estimators[,1]
lambda_hat <- new_estimators[,2]
```

The mean our *new* 1000 estimations of r and λ are 3.1447 and 1.9287, respectively. The SD of our the 1000 estimations is 0.1654 and 0.1062. It seems that our estimators are,

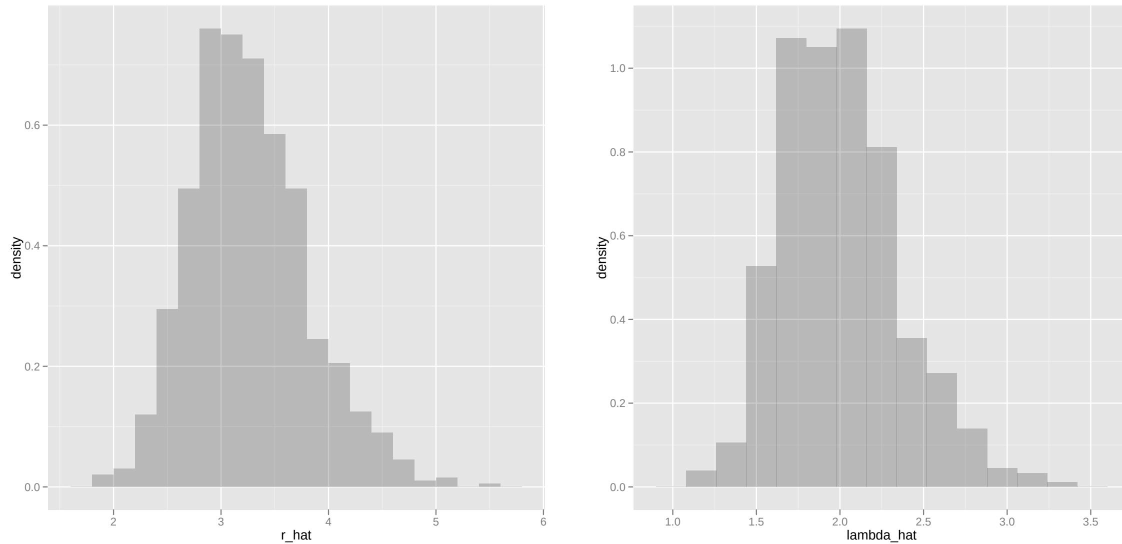

Figure 2: Histograms for $\hat{r}$ and $\hat{\lambda}$ when $n = 100$

indeed, converging to their targets. I'm willing to say this is most likely a consequence of the fact that $\hat{\mu}_k \to E(X^k)$ with whatever speed the weak-law of large numbers affords us. Figure 3 shows the histograms for these new estimates. Notice that they more closely resemble bell curves.

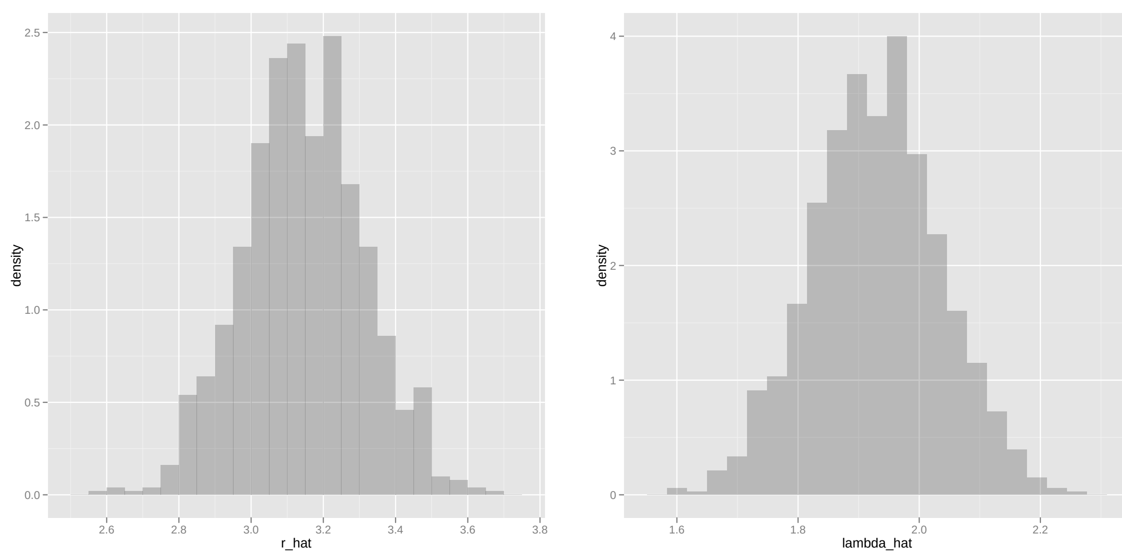

Figure 3: Histograms for $\hat{r}$ and $\hat{\lambda}$ when $n = 1000$

#10. The following code performs the bootstrap algorithm, based on the preceding chunk's result:

```
r_hat1 <- r_hat[1]
lambda_hat1 <- lambda_hat[1]

bootstrap_estimators <- t(mapply(FUN=generateEstimators,
                       n=rep(n,times=times),
                       r=rep(r_hat1,times=times),
                       lambda=rep(lambda_hat1,times=times)))

r_hat_bootstrap <- bootstrap_estimators[,1]
lambda_hat_bootstrap <- bootstrap_estimators[,2]
```

And we find the new bootstrap estimators $\hat{\hat{r}}$ and $\hat{\hat{\lambda}}$ to have means and SD's 2.8622, 0.1498 and 1.7577, 0.0988. Note that our original $\hat{r} = 2.8851$ and $\hat{\lambda} = 1.7675$, so our bootstrap estimators do a very good job mimicing the original estimators.

The central 95% of $\hat{\hat{r}}$'s distribution is in (2.5797,3.1495), and the central 95% of $\hat{\hat{\lambda}}$ is in (1.5668,1.9458). A simple inspection of figure 3 shows that these intervals pretty much capture the same information as the distribution of $\hat{r}$ and $\hat{\lambda}$, so we can be confident that our estimands (r, λ) can be estimated via $\hat{\hat{r}}$ and $\hat{\hat{\lambda}}$.

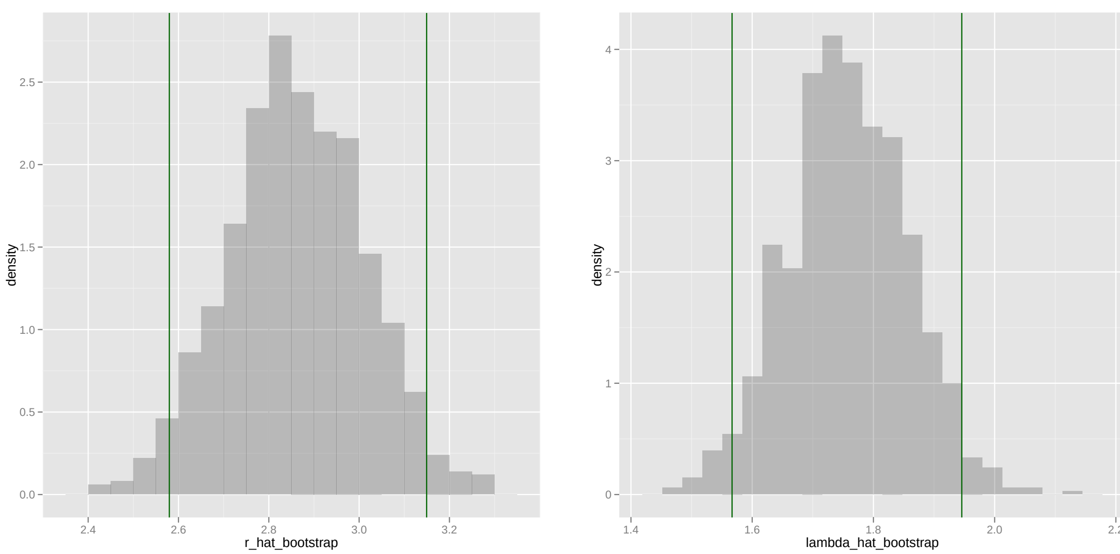

Figure 4: Histograms for $\hat{\hat{r}}$ and $\hat{\hat{\lambda}}$ generated from bootstrap with vertical lines denoting the central 95% of each distribution