Math 152 - Statistical Theory - Exam 1

Ethan Ashby

Due: Monday, Oct 5, 2020, 5pm PDT

3. χ^2 distribution part (c)

c. Consider the following dataset of 100 observations from $N(0, 47)$. Bootstrap the data to find the SE of both of your estimators: $\hat{\theta}_1$ and $\hat{\theta}_2$. Recall that in part b. you calculated the variability (as part of the MSE). Are the analytic values from part b. consistent with the bootstrap values in part c.? Explain.

```
set.seed(123)
testdata <- rnorm(100,0,sqrt(47))  # do not change the dataset

#bootstrap
reps <- 1000
esttheta1<-c()
esttheta2<-c()
for (i in 1:reps){
  bootstrap_data<-sample(testdata, replace=TRUE)
  esttheta1 <- c(esttheta1, sum(bootstrap_data^2)/100)
  esttheta2 <- c(esttheta2, var(bootstrap_data))
}

paste("SE(estimator1):", round(sd(esttheta1), 3))
```

```
## [1] "SE(estimator1): 5.196"
```

```
paste("SE(estimator2):", round(sd(esttheta2), 3))
```

```
## [1] "SE(estimator2): 5.095"
```

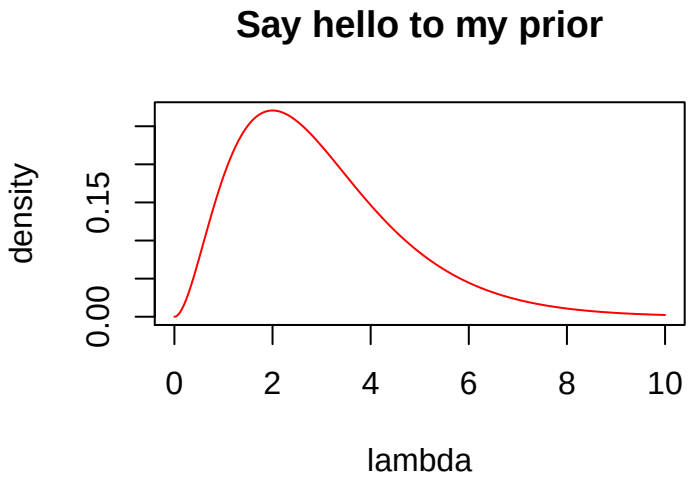
The bootstrap standard error of these estimators are both approximately 5. The variance of the estimators (as part of the MSE) calculated in part (b) was $\frac{2\theta^2}{n} = 44.18$ and $\frac{2\theta^2}{n} = 44.63$ (SD approx 6.6). Thus, the analytic values are in the same ballpark as the bootstrap values, although the boostrap variability is understimating the analytic variability slightly.

4. Starbucks

a. See description on sheet. In short, I believe on a typical day and time, there will be 3 ppl in front of me. I also wanted my prior to be appropriately variable to account for variability in line length as a function of day (weekday vs weekend) and time of day (rush hour or late morning). I also like the right skewed shape, since there will be rare occasions when lines are really long. Also the gamma distribution plays nicely with the Poisson distribution (conjugacy), which makes my life easier :)

b. Draw the prior distribution in R

```
plot(seq(0,10,0.01), dgamma(seq(0,10,0.01), shape=3, rate=1), "l", col="red", ylab="density", xlab="lam
```



c. Derivations on page. Data calculations are conducted below

```
data1 <- c(7,2,4,5,3,4,2,3,3,6,1,4,1,3,3,5,2,0,1,3,1,3,3,1,4,2,0,0,3,4)
data2 <- c(4,2,3,1,5,3,2,7,3,4,1,2,3,6,2,2,4,6,2,5,1,4,1,0,0,3,0,0,3,3,4,2,3,2,1, 3,4,1,4,1,3,1,2,4,3,2
           5,4,6,2,7,2,1,2,0,5,5,4,1,0,2,2,2,2,2,2,3,5,3,3,6,1,5,3,3,3)

#case 1
alpha1=sum(data1)+3
beta1=length(data1)+1
alpha1
```

```
## [1] 86
```

```
beta1
```

```
## [1] 31
```

```
#case 2
alpha2=sum(data2)+3
beta2=length(data2)+1
alpha2
```

```
## [1] 292
```

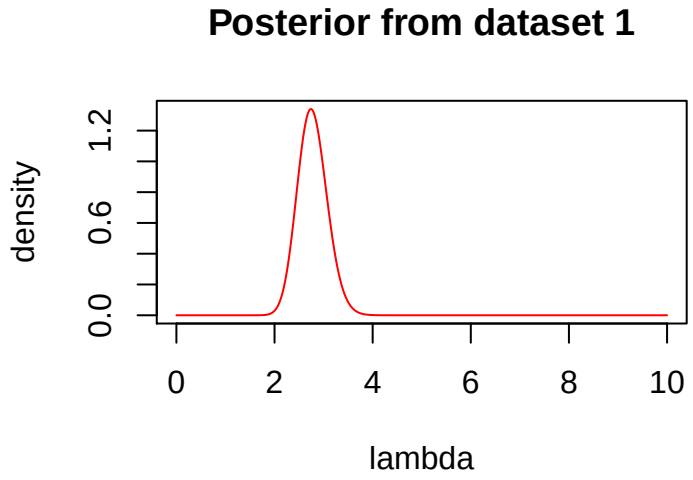
```
beta2
```

```
## [1] 101
```

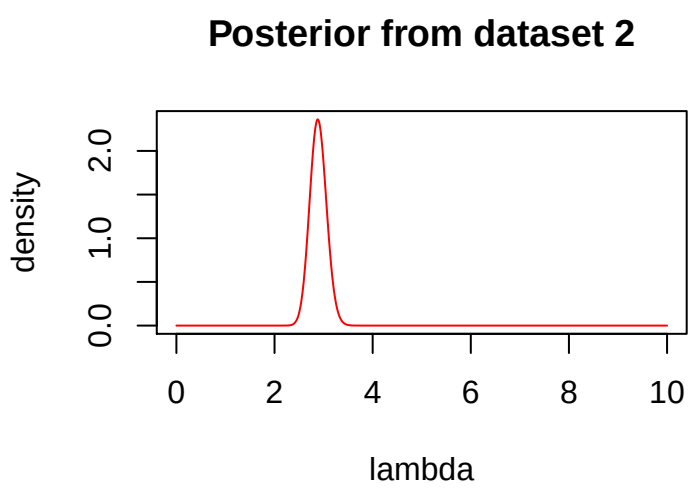
The alpha and beta values for the gamma distributions are calculated above. Dataset 1 produces the posterior Gamma(86, 31) and Dataset 2 produces the posterior Gamma(292, 101).

d. Draw posteriors

```
plot(seq(0,10,0.01), dgamma(seq(0,10,0.01), shape=86, rate=31), "l", col="red", ylab="density", xlab="la
```




```
plot(seq(0,10,0.01), dgamma(seq(0,10,0.01), shape=292, rate=101), "l", col="red", ylab="density", xlab=
```



e. On sheet. The priors and posteriors have similar centers but the posteriors have much narrower spread. With more data, the posterior narrows, which is why Posterior 2 is narrower/less variable than Posterior 1.

5. Normal distribution The estimators I will work with are as follows: the mean, variance, and two weighted averages of the mean and variance.

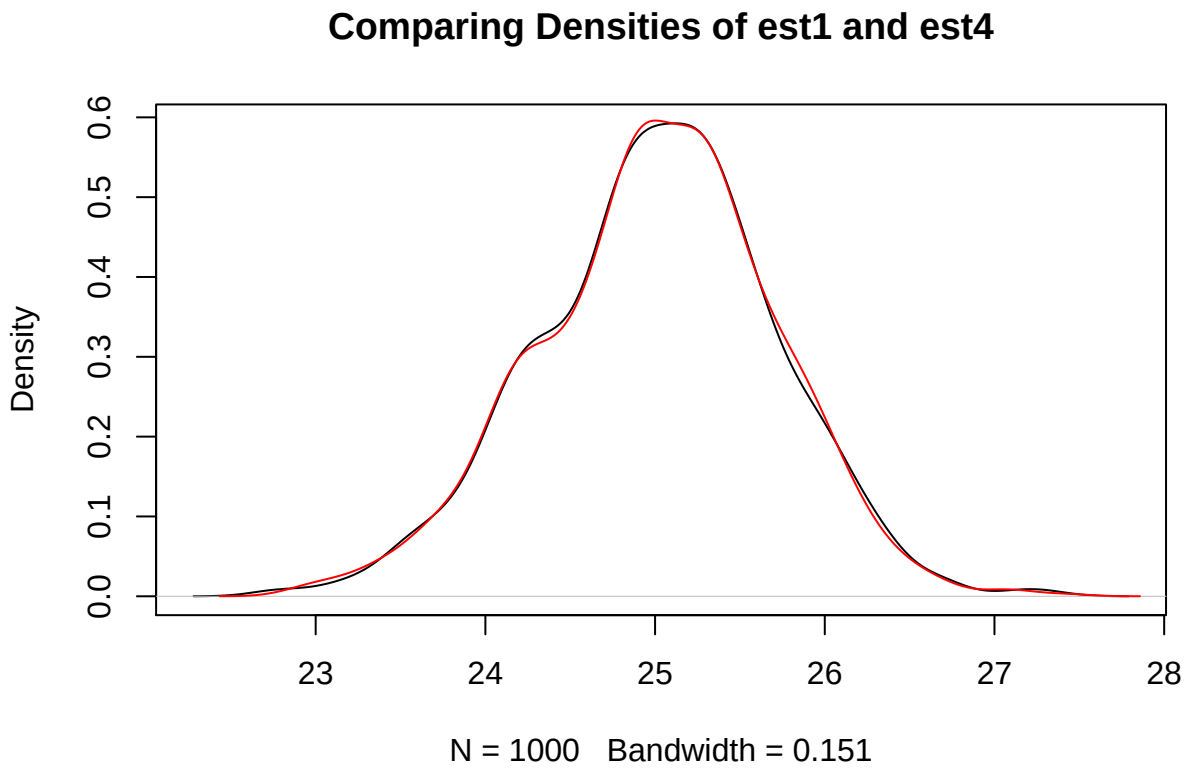
1. $\bar{X}$
2. $\text{Var}(X)$
3. $\frac{\bar{X}+\text{Var}(X)}{2}$
4. $0.98\bar{X} + 0.02\text{Var}(X)$

The first two choices of estimators were easy targets, since we know the normal distribution from which the data are drawn have mean=θ and variance=θ. But these two estimators should be somewhat limited because they only rely on half the information, and don't incorporate the fact that **BOTH** the mean and variance are equal to θ. So I constructed two more estimators which are weighted averages of these two quantities.

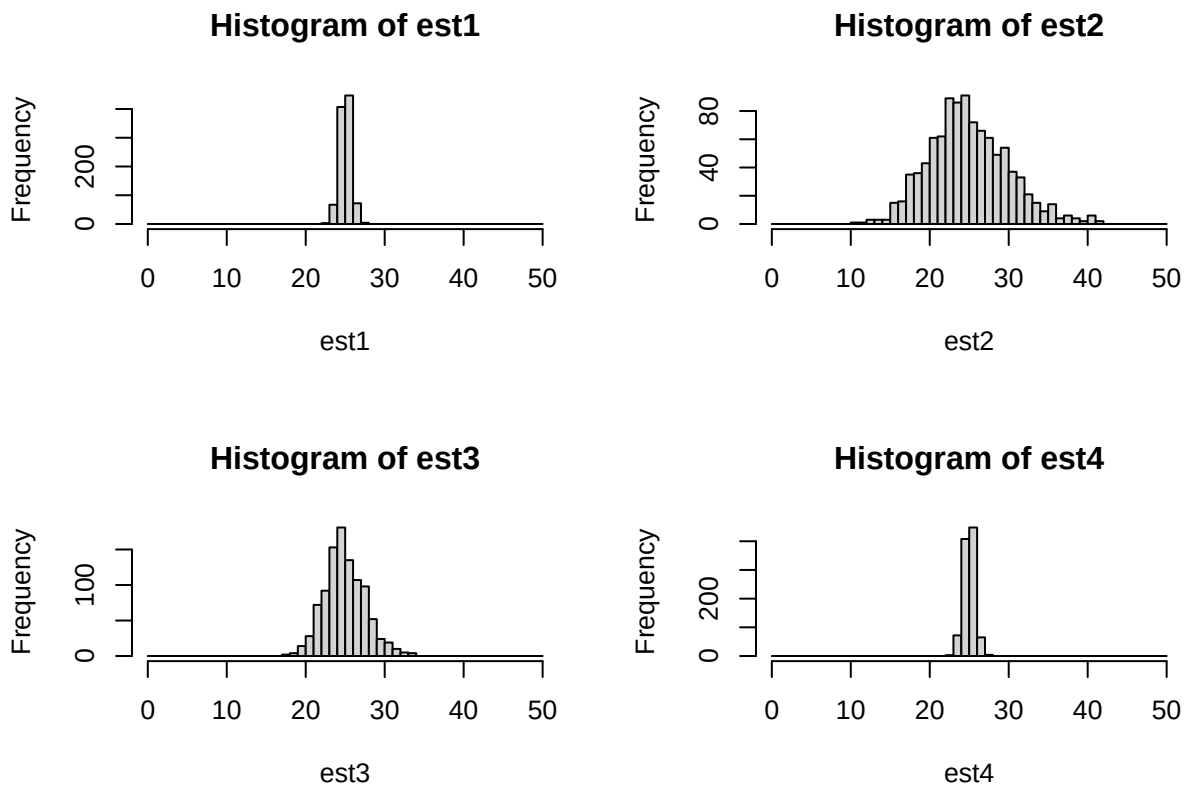
```r
set.seed(11)

est1<-c()
est2<-c()
est3<-c()
est4<-c()

for (i in 1:1000){
data<-rnorm(50, mean=25, sd=sqrt(25))

est1<-c(est1, mean(data))
est2<-c(est2, var(data))
est3<-c(est3, (mean(data)+var(data))/2)
est4<-c(est4, 0.98*mean(data)+0.02*var(data))
}

results=data.frame("Estimator"=c(1,2,3,4), "Mean"=c(mean(est1), mean(est2), mean(est3), mean(est4)), "Va
knitr::kable(results)
```

Estimator	Mean	Variance	Bias	MSE
1	25.01129	0.4839547	0.0001275	0.4840822
2	24.93920	25.6770231	0.0036969	25.6807200
3	24.97524	6.5926416	0.0006128	6.5932544
4	25.00985	0.4791688	0.0000970	0.4792659

```r
plot(density(est1), main="Comparing Densities of est1 and est4")
lines(density(est4), col="red")
```




```
par(mfrow=c(2,2))
hist(est1, breaks=0:50)
hist(est2, breaks=0:50)
hist(est3, breaks=0:50)
hist(est4, breaks=0:50)
```



The results in the table above display the mean, variance, bias, and MSE of these estimators over 1000 random samples of size 50 drawn from N(25,25). Estimator 2 was the worst estimator, as it led to the largest bias, variance, and MSE by factors of up to 50. This makes sense, since the variance is a more unstable quantity than a measure of central tendency, like the mean. Estimator 3 was the next worst, since it generated a large variance, relatively large bias, and large MSE. The results show that **Estimator 4** had the lowest variance of ~0.004 smaller than the next best estimator (Estimator 1). Estimator 4 also generated a slight reduction in bias compared to Estimator 1, which resulted in a MSE of ~0.005 smaller than Estimator 1.

The empirical sampling distributions, as shown in the histograms above, reflect the results shown in the table. The histograms of Estimators 1 and 4 are nearly identical and are the narrowest by far. Estimators 2 and 3 show really wide empirical sampling distributions, indicating that they have much larger variance. On these zoomed-out plots, its difficult to see any differences in bias, which while present, are small relative to the major differences in variance. Estimators 1, 3, 4 are built off of means, and since the sampling distributions of means are normal, these empirical sampling distributions are symmetric about their means. Estimator 2 (which is based off the variance) is roughly symmetric as well, and looks to be centered at its mean.

Superimposing the kernel densities of the empirical sampling distributions of estimators 1 and 4 show that these estimators have nearly identical shape. If you look *really* closely, you can spot that estimator 4's density (in red) is slighly higher at 25, intimating a reduction in bias. The red curve also looks slightly skinnier, reflecting the reduction in variance.

In short, **Estimator 4**: $0.98\bar{X} + 0.02\text{Var}(X)$ is my estimator of choice.