Problem Set 4 - Group F

Azka Javaid, Tasheena Narraidoo, Daniel Law

02/28/2016

10.47

In this question we are given the PDF:

$$f(x \mid x_0, \theta) = \theta x_0^{\theta} x^{-\theta-1}, \; x \geq x_0, \theta > 1$$

Assume that $x_0 > 0$ is given and that $X_1, ..., Xn$ is an i.i.d sample.

(a) Find method of moments estimate of θ

$$E[X] = \int_{x_0}^{\infty} \theta x_0^{\theta} x^{-\theta-1}(x)dx = \int_{x_0}^{\infty} \theta x_0^{\theta} x^{-\theta} dx$$

$$= \theta x_0^{\theta} \int_{x_0}^{\infty} x^{-\theta} dx = \theta x_0^{\theta} \left[\frac{x^{-\theta+1}}{-\theta+1} \right]_{x_0}^{\infty}$$

$$= \theta x_0^{\theta} (\frac{-x_0^{-\theta+1}}{-\theta+1}) = \frac{-\theta x_0^{\theta} x_0^{-\theta+1}}{-\theta+1} = \frac{-\theta x_0}{-\theta+1}$$

Now we have

$$E[X] = \frac{-\theta x_0}{-\theta+1} = \mu_1$$

rearranging for θ

$$\theta = \frac{\mu_1}{(\mu_1 - x_0)}$$

$$\hat{\theta} = \frac{\bar{X}}{(\bar{X} - x_0)}$$

(b) Find mle of θ

$$f(x \mid x_0, \theta) = \theta x_0^{\theta} x^{-\theta-1}$$

$$l(\theta) = \sum_{i=1}^{n} (log(\theta) + log \; x_0 + (-\theta-1) logx) = nlog(\theta) + \theta \sum_{i=1}^{n} logx_0 + (-\theta-1) \sum_{i=1}^{n} logx_i$$

$$\frac{\partial l}{\partial \theta} = \frac{n}{\theta} + \sum_{i=1}^{n} (log(x_0) - \sum_{i=1}^{n} (log(X_i) = \frac{n}{\theta} + nlogx_0 - \sum_{i=1}^{n} (log(X_i)$$

$$\frac{n}{\theta} + nlogx_0 - \sum_{i=1}^{n}(log(X_i) = 0$$

$$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} logX_i - nlogx_0}$$

(c) Find asymptotic variance of mle

$$Var(\theta) = \frac{1}{nI(\theta)}, f(x \mid x_0, \theta) = \theta x_0^{\theta} x^{-\theta-1}$$

$$I(\theta) = -E[\frac{\partial^2}{\partial\theta^2} logf(x \mid \theta_0)]$$

$$logf(x \mid \theta_0) = log\theta + \theta logx_0 + (-\theta - 1)logx!$$

$$\frac{\partial^2}{\partial\theta^2} logf(x \mid \theta_0) = \frac{-1}{\theta^2}$$

$$I(\theta) = \frac{1}{\theta^2} \rightarrow Var(\tilde{\theta}) = \frac{\theta^2}{n}$$

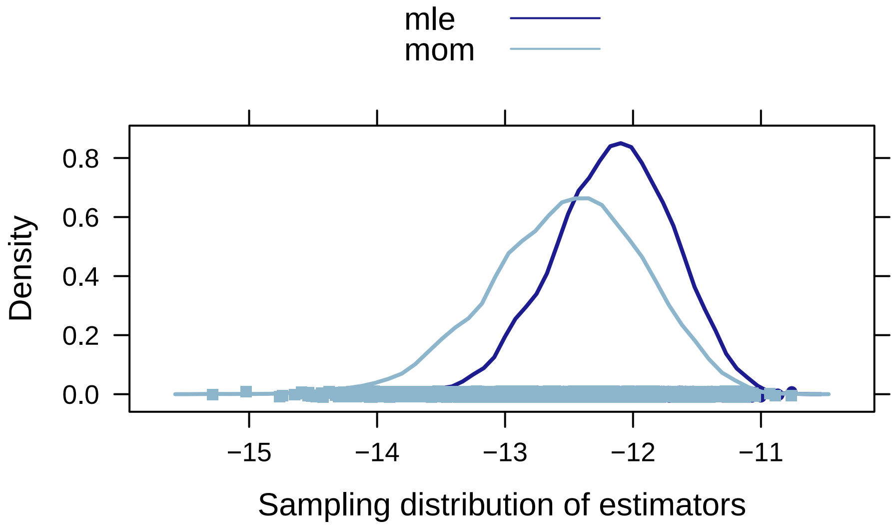
(d) Find the sufficient statistic for θ

Corollary A on pg 309 states that if T is sufficient for θ, the maximum likelihood estimate is a funtion of T. So, since MLE for $\theta = \frac{n}{\sum_{i=1}^{n} logX_i - nlogx_0}$,
the sufficient statistic $= \sum_{i=1}^{n} logX_i$

Empirical component for Problem 47:

For the simulation component of problem 47, we simulated the mom and mle estimates from a pareto distribution. A pareto distribution necessitates specification of parameters scale and shape. We chose arbitrary values for the scale and shape parameters (scale = 5, shape = 10). The x_0 value was also arbitrarily chosen to be 6 for x_0. The variance of the mle and mom estimates as well as the efficiency of the two parameters was calculated.

```r
set.seed(13)
numsim <- 5000
simfum <- function(n=1000, scale=5, shape=10){
  x<-VGAM::rpareto(n, scale=scale, shape=shape) #sampling from pareto distribution
  mom <- mean(x)/(mean(x)-6) #calculating mom estimate with arbitrary value for x0
  mle <- n/(sum(log(x))-n*log(6)) #calculating mle estimate with arbitrary value for x0
  return(data.frame(mom=mom, mle=mle))
}
res<-do(numsim)*simfum()
densityplot(~mle+mom, auto.key=TRUE, xlab="Sampling distribution of estimators", lwd=2, data=res)
```




```
var(~mom, data=res) #variance of the mom estimate
```

```
## [1] 0.3563144
```

```
var(~mle, data=res) #variance of the mle estimate
```

```
## [1] 0.2166633
```

```
var(~mom, data=res)/var(~mle, data=res) #calculating efficiency
```

```
## [1] 1.644554
```

An approximation of the variance of the mle estimate is 0.217. In comparison, variance of the mom estimate is 0.356. The efficiency of the mom estimate relative to the mle estimate is about 1.64. This indicates that the variance of the mom estimate is about 0.64 higher that the variance of the mle estimate.

8.58

If the gene frequencies are in equilibrium, the genotypes AA, Aa, and aa occur with probabilities $(1-\theta)^2$, $2\theta(1-\theta)$ and θ^2, respectively. Plato et al. (1964) published the following data on haptoglobin type in a sample of 90 people:

```
##   Haptoglobin Type
##    Hp1-1 Hp1-2 Hp2-2
##       10    68   112
```

(a) Find the mle of θ.

$$l(p_1, ..., p_m) = logn! - \sum_{i=1}^{m} logx_i! + \sum_{i=1}^{m} logp_i$$

$$l(\theta) = logn! - \sum_{i=1}^{3} logX_i! + X_1 log(1-\theta^2) + X_2 log 2\theta(1-\theta) + X_3 log\theta^2$$

$$l(\theta) = logn! - \sum_{i=1}^{3} logX_i! + (2X_1 + X_2) log(1-\theta) + (2X_3 + X_2) log\theta + X_2 log 2$$

Set the derivitive with respect to θ equal to 0:

$$l'(\theta) = -\frac{2X_1 + X_2}{1-\theta} + \frac{2X_3 + X_2}{\theta} = 0$$

Solving the above equation in terms of θ to obtain the MLE:

$$\hat{\theta} = \frac{2X_3 + X_2}{2X_1 + 2X_2 + 2X_3}$$

$$\hat{\theta} = \frac{2X_3 + X_2}{2n}$$

$$\hat{\theta} = \frac{2(112) + 68}{2(190)} = .768$$

(b) Find the asymptotic variance of the mle.

For parameters estimated from random multinomial counts:

$$Var(\hat{\theta}) \approx -\frac{1}{E[l''(\theta_0)]}$$

$$l'\theta = -\frac{2X_1 + X_2}{1-\theta} + \frac{2X_3 + X_2}{\theta}$$

$$l''\theta = -\frac{2X_1 + X_2}{(1-\theta)^2} + \frac{2X_3 + X_2}{(\theta)^2}$$

Since X_1 are binomially distributed:

$$E(X_1) = n(1-\theta)^2$$

$$E(X_2) = 2n\theta(1-\theta)$$

$$E(X_3) = n\theta^2$$

$$E[l''(\theta)] = -\frac{2n}{\theta(1-\theta)}$$

$$Var(\hat{\theta}) \approx \frac{\hat{\theta}(1-\hat{\theta})}{2n}$$

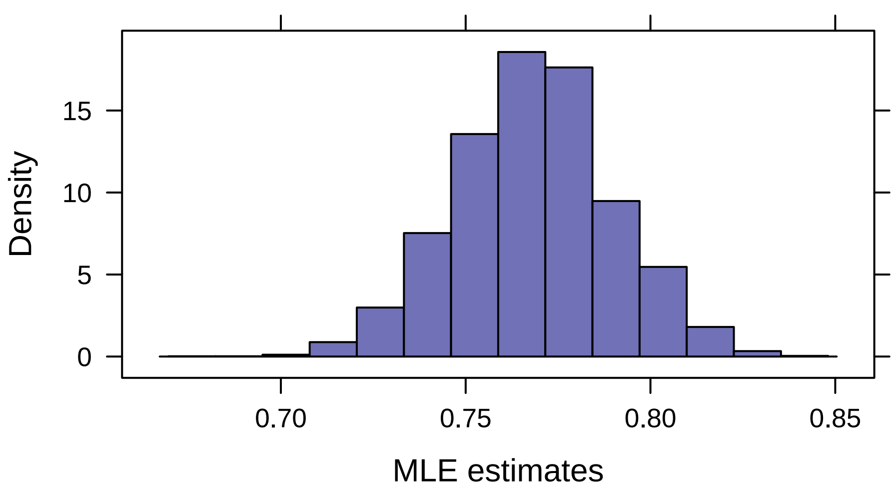
$$Var(\hat{\theta}) \approx 0.0004689$$

(c) Find an approximate 99% confidence interval for θ.

An approximate 99% confidence interval for θ is $\hat{\theta} \pm 2.57 s_{\hat{\theta}}$ where $s_{\hat{\theta}} = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{2n}} = 0.0216$. So the confidence interval is (0.712, 0.824)

(d) Use the bootstrap to find the approximate standard deviation of the mle and compare to the result of part (b).

```
theta_est = 292/380
n_total = 190
numsim = 10000

original_mle_est <- (2*112+68)/(2*n_total)

#Asymptotic variance from part (b).
asym_var <- theta_est*(1-theta_est)/(2*n_total)

#This function generates random variables 'AA', 'Aa' and 'aa' based on the probability distribution

RV <- function(num,theta){
  x <- runif(num)
  rv <- ifelse( x <= (1-theta)^2, 'AA', x)
  rv <- ifelse( rv > (1-theta)^2 & rv<= (1-theta)^2 + 2*theta*(1-theta), 'Aa', rv)
  rv <- ifelse( rv != 'AA' & rv != 'Aa', 'aa', rv)
  return(rv)
}

#This function generates a sample of the RVs above and calculates the MLE of the sample

simFun <- function(theta = theta_est, n=n_total){
  sim <- tally(RV(n,theta))
  mle_est <- (2*sim['aa']+sim['Aa'])/(2*n)
  return(mle_est)
}

mle <- do(numsim)*simFun()

histogram(mle$aa, xlab = "MLE estimates")
```



```r
mean(mle$aa)
```

```
## [1] 0.7687303
```

```r
#looking at the bootstrap variance vs the analytic asymptotic variance
var(mle$aa)
```

```
## [1] 0.0004584966
```

```r
asym_var
```

```
## [1] 0.0004682898
```

We see above that the asymptotic variance is in agreement with the variance of our bootstrapped MLE estimates. Suggesting that with 10000 simulations we are close to the asymptotic variance.

(e) Use the bootstrap to find an approximate 99% confidence interval and compare to part (c).

```r
#working out the 99% CI interval from the bootstrap
quants <- as.numeric(quantile(mle$aa,c(0.005,0.995)))
up_low <- quants - original_mle_est

#lower 0.5% bound
original_mle_est - up_low[2]
```

```
## [1] 0.7157895
```

```r
#upper 99.5% bound
original_mle_est - up_low[1]
```

```
## [1] 0.8236842
```

From part (c) we have a 99% CI of (0.712, 0.824), which is very close to our bootstrapped 99% CI above. Suggesting that in this case, both methods work well in calculating confidence intervals.