Group Assignment 4

Sarah Bailey, Richard Hsia, Lillian Yan Lin, Yue Ma, Will Ruth, Michelle Thiessen

November 9, 2016

Question 7.7

We can estimate $F(x)$ with the empirical distribution function $\hat{F}_n$, which is defined as

$$\hat{F}_n(x) = \frac{\sum_{i=1}^{n} I(X_i \leq x)}{n},$$

where

$$I(X_i \leq x) = \begin{cases} 1 & \text{if } X_i \leq x \\ 0 & \text{if } X_i > x \end{cases}.$$

Adapting the code `chapter7.R` (see appendix) provided in class, the empirical distribution (Fig. 1) and a 95 percent confidence envelope for F (Fig. 2) can be found.

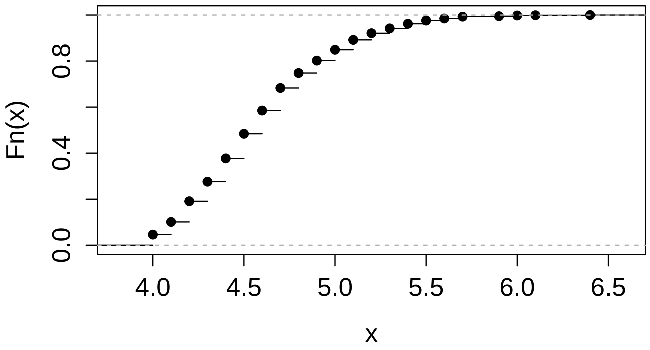


Figure 1: Empirical distribution $\hat{F}_n$ to estimate $F(x)$.

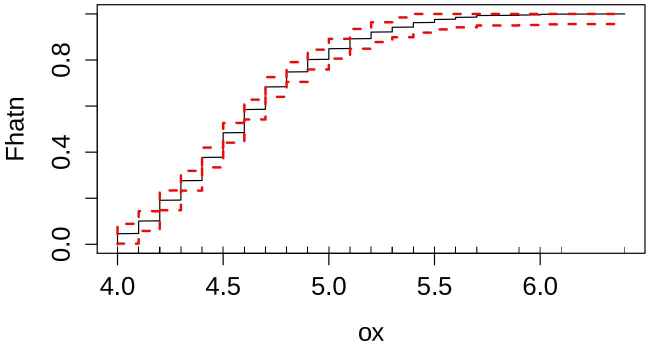


Figure 2: Empirical distribution $\hat{F}_n$ to estimate $F(x)$.

An approximate 95 percent confidence interval for the linear functional $T(F) = F(4.9) - F(4.3)$ can be found by first obtaining the estimated standard error of $\hat{\theta}_n = T(\hat{F}_n) = \hat{F}_n(4.9) -$

$\hat{F}_n(4.3) = 0.526$, which is the plug-in estimator of $\theta = T(F)$. The estimated standard error is found to be

$$\begin{aligned}\hat{\text{se}} &= \sqrt{\frac{\left(\hat{F}_n(4.9) - \hat{F}_n(4.3)\right)\left(1 - \hat{F}_n(4.9) + \hat{F}_n(4.3)\right)}{n}} \\ &= \sqrt{\frac{0.526(1-0.526)}{1000}} \\ &= 0.01579.\end{aligned}$$

The confidence interval can then be constructed as $T(\hat{F}_n) \pm z_{\alpha/2}\,\hat{\text{se}}$, and $z_{.05/2} = 1.96$. So the 95 percent confidence interval for $F(4.9) - F(4.3)$ is

$$0.526 \pm 1.96 \times 0.01579 = 0.526 \pm 0.031.$$

Appendix

```
# Plot ECDF for Fiji quakes data
x <- read.table("http://stat.cmu.edu/~larry/all-of-statistics/
                =data/fijiquakes.dat",skip=1)$V5
ee <- ecdf(x)
plot(ee)

ee(4.9)
abline(v=4.9)
abline(h=ee(4.9))
abline(v=4.3)
abline(h=ee(4.3))

ee(4.9) - ee(4.3)

#-------------------------------------------------------
# Draw our own and add nonparametric confidence band.
# Adapted from R code on author's website:
# www.stat.cmu.edu/~larry/all-of-statistics/=Rprograms/edf.r
ecdf2 = function(x,CI=TRUE) {
  ox <- sort(x) # ordered x
  n <- length(x)
  Fhatn <- (1:n)/n
  plot(ox,Fhatn,type="l")
  rug(x,ticksize=0.025)
  if(CI) {
    alpha <- 0.05
    eps <- sqrt(log(2/alpha)/(2*n))
    upper <- pmin(Fhatn + eps,1)
    lower <- pmax(Fhatn - eps,0)
    lines(ox,upper,type="s",lwd=2,col=2,lty=2)
    lines(ox,lower,type="s",lwd=2,col=2,lty=2)
  }
}
ecdf2(x)
```

Question 8.7

(a) If $X_1, X_2, ..., X_n \sim Uniform(0, \theta)$ a special property of the order statistics $X_{(1)}, X_{(2)}, ..., X_{(n)}$ is that

$$\frac{X_{(i)}}{\theta} \sim Beta(i, n - i + 1).$$

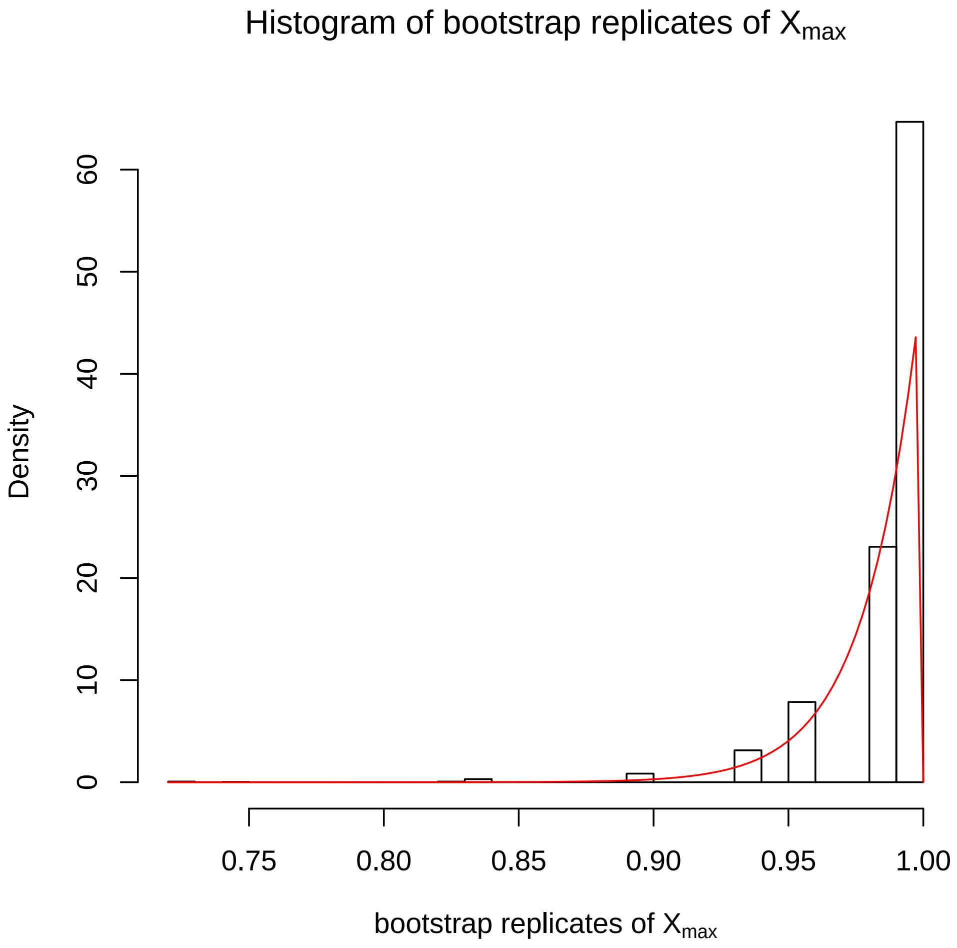
So the distribution of $\hat{\theta} = X_{max} = X_{(n)}$ is

$$g(\hat{\theta}; \theta) = \frac{n\hat{\theta}^{n-1}}{\theta^n}; 0 < \hat{\theta} < \theta.$$

You can also find it through the CDF method. Let $Y = X_{(n)}$.

$$F_Y(y) = P(Y \leq y) = P(max(X_1, ..., X_n) \leq y) = P(X_1 \leq y, ..., X_n \leq y) = \prod_{i=1}^{n} P(X_i \leq y) = [F_X(y)]^n$$

$$f_Y(y) = \frac{d}{dy} [F_X(y)]^n = n f_X(y) [F_X(y)]^{n-1} = \frac{n\hat{\theta}^{n-1}}{\theta^n}; 0 < \hat{\theta} < \theta.$$



(b) $P(\hat{\theta}^* = \hat{\theta}) = P(max(X_1^*, ..., X_n^*) = max(X_1, ..., X_n)) = 1 - P(max(X_1^*, ..., X_n^*) \neq max(X_1, ..., X_n))$.

Recall that $P(X^* = X_i) = \frac{1}{n}$ for $i = 1, ..., n$, so

$$P(max(X_1^*, ..., X_n^*) \neq max(X_1, ..., X_n)) = P(X^* \neq X_{(n)})^n = \left(1 - \frac{1}{n}\right)^n$$

giving us

$$P(max(X_1^*, ..., X_n^*) = max(X_1, ..., X_n)) = 1 - \left(1 - \frac{1}{n}\right)^n.$$

Then we take the limit as $n \to \infty$,

$$\lim_{n\to\infty} 1 - \left(1 - \frac{1}{n}\right)^n = 1 - e^{-1} \approx 0.632$$

Meanwhile $P(\hat{\theta} = \theta) = 0$ since $\hat{\theta}$ is continuous.

Appendix

```
set.seed(25)
theta <- 1;n <- 50;R.x <- runif(n,0,theta) #Our 'original' sample
bootstrap.max <- function(original.sample,n){
  selected.elements <- sample(1:n, size = n, replace=TRUE)
  bootstrap.sample <- original.sample[selected.elements]
  return(max(bootstrap.sample))
}
replicates.boot <- replicate(5000,bootstrap.max(R.x,n))
hist(replicates.boot,breaks=30,freq=FALSE,
     main=expression(paste("Histogram of bootstrap replicates of ",X[max])),
     xlab=expression(paste("bootstrap replicates of ",X[max])))
curve(dbeta(x,n,1),col="red",add=T)
```