STAT 621 Homework 4

Ben Buzzee

10/4/2019

1.

(a) Find $P(max(U_i) > .75)$

Since $max(U_i)$ is an order statistic, we can show it's cdf is $F(x)^n$. Since we are dealing with a continuous uniform distribution, $P(max(U_i) > .75) = 1 - P(max(U_i) < .75)^3 = 1 - .75^3$ for n = 3

This tells us $P(max(U_i) > .75) = .578$

(b)

Next we will do Monte Carlo simulation to compare.

```
M=50000
out=rep(NA, M)
for (i in 1:M){
  sample <- runif(n = 3)
  out[i] <- any(sample > .75)
}

sum(out/M)
```

```
## [1] 0.58076
```

And we find $P(max(U_i) > .75)$ to be very close to the value we derived above.

(c)

I think the probability of observing a value above .75 will increase with n. Each additional sample will "contribute" some probability to the cumulative probability.

(d)

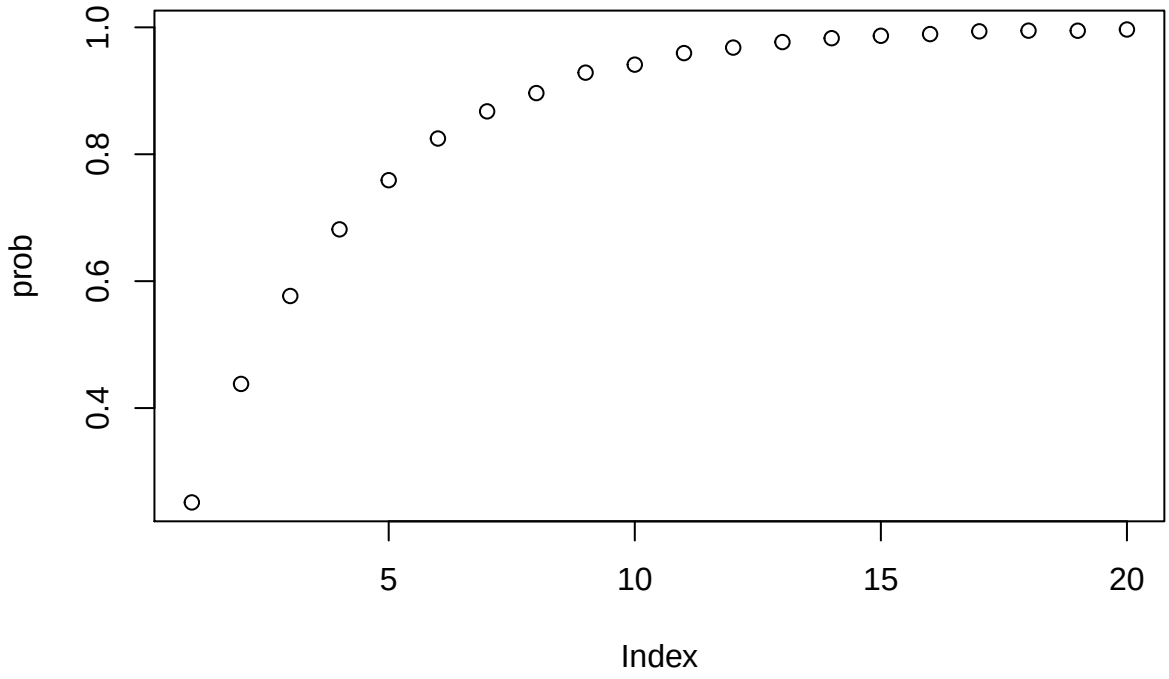
To check this assumption we will do a Monte Carlo Simulation:

```
nreps <- 20
prob <- rep(NA, max(nreps))
M = 10000
out=rep(NA, M)


for (j in 1:nreps){
  for (i in 1:M){
    sample <- runif(n = j)
    out[i] <- any(sample > .75)
  }
```

```
  prob[j] <- sum(out/M)
}

plot(prob)
```



We find that our hunch was correct and $P(max(U_i) > .75)$ increases as the sample size increases until it approaches an asymptote.

2. Simulation Exercise

For this simulation exercise we will investigate the performance of the Kolmogorov-Smirnov test. The Alaska Department of Fish and Game often performs Mark-Recapture experiments to determine the size of local populations of fish. One of the key assumptions in these experiments is that there is no size-selective sampling during either the mark or the recapture events. To check for violations of this assumption, we perform two-sample KS tests on the distribution of lengths from each of the mark and recapture events.

To keep the simulation exercise simple, we will loosely base our simulation off of real world data. In 2018 ADFG did a mark-recapture study in the three-mile complex of lakes in Beluga, AK. The observed range of lengths was 200mm - 800mm with a slight right skew, and the mean difference of lengths between the mark event and the recapture event was 10mm. So we will sample from a normal(450, 50^2) distribution.

```
# set parameters
mean = 450
sd = 50
diff = c(5, 10, 20)
```

```r
samp_size = c(10, 50, 100, 250)
M = 1000

# power matrix
results.1 <- matrix(NA, nrow=length(samp_size), ncol=3)

# final result matrix
results.2 <- NA

for (h in 1:length(diff)){

  for (i in 1:length(samp_size)){

    p.val <- rep(NA, times = M)

    for (j in 1:M){

    # simulate data
    sample1 <- rnorm(samp_size[i], mean = mean, sd = sd)
    sample2 <- rnorm(samp_size[i], mean = mean + diff[h], sd = sd)

    # concerned with general size selectivity in either event
    # so we will do a two-sided test
    p.val[j] <- ks.test(sample1, sample2, alternative = "two.sided")$p.value
    }

    power <- sum(p.val < .05)/M
    results.1[i,] <- cbind(samp_size[i], diff[h], power)

  }
  results.2 <- rbind(results.2, results.1)

}


results.2 <- as.data.frame(results.2[-1,], row.names = FALSE)

names(results.2) <- c("Sample Size", "True Diff.", "Power")

knitr::kable(results.2)
```

Sample Size	True Diff.	Power
10	5	0.011
50	5	0.066
100	5	0.067
250	5	0.134
10	10	0.023
50	10	0.112
100	10	0.194
250	10	0.450
10	20	0.044
50	20	0.375
100	20	0.629

Sample Size	True Diff.	Power
250	20	0.969

Given the assumption of normally distributed data and our parameter values, we see that the power of our KS tests are quite low. Even if we have 250 measurements in both the mark and recapture datasets, we have less than a 50% chance of correctly rejecting the null given the true difference in mean lengths is 10mm.

The next step would be to determine the amount of bias a 10mm difference in mean lengths has on our final estimate, and whether the amount of bias is negligible or should be corrected for. We could also vary our distributional and parameter assumptions.

3

(a) Spearmans Correlation

```
states <- read.csv("states.txt", sep = " ")

spending <- states$spending
rate <- states$rate


cor.test(spending, rate, method = "spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  spending and rate
## S = 60, p-value = 0.1777
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
## rho
## 0.5
```

With a test statistic of S = 60, we fail to reject the null and have no evidence rho is non-zero.

(b) Kendalls Tao

```
d <- cor.test(spending, rate, method = "kendall")

d$statistic
```

```
##  T
## 25
```

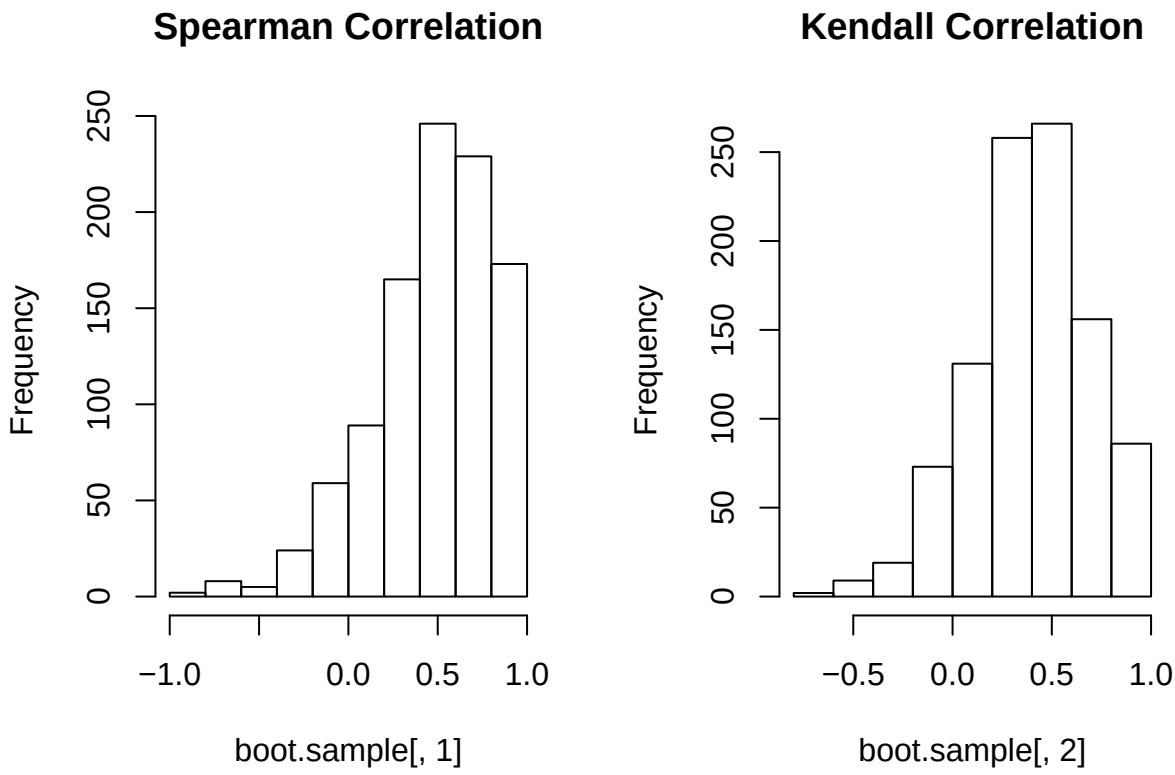
With a test statistic of 25 and p-value of .18, we have no evidence tao is nonzero.

(c) Bootstrapping

```r
nreps=1000
boot.sample <- matrix(NA,nreps, ncol = 2)
for(i in 1:nreps)
{
newSamp <- dplyr::sample_n(states, size = 9, replace = TRUE)

spearman <- cor(newSamp$spending, newSamp$rate, method = "spearman")
kendall <- cor(newSamp$spending, newSamp$rate, method = "kendall")
boot.sample[i,1] <- spearman
boot.sample[i,2] <- kendall
}

par(mfrow = c(1,2) )
hist(boot.sample[,1], main = "Spearman Correlation")
hist(boot.sample[,2], main = "Kendall Correlation")
```



And our 90% confidence intervals are:

```r
print("Spearman 90% Boostrap Interval")
```

```
## [1] "Spearman 90% Boostrap Interval"
```

```r
quantile(boot.sample[,1], c(.05, .95))
```

```
##         5%        95%
## -0.1478261  0.9473684
```

```r
print("Kendall 90% Boostrap Interval")
```

```
## [1] "Kendall 90% Boostrap Interval"
```

```r
quantile(boot.sample[,2], c(.05, .95))
```

```
##          5%         95%
## -0.09781784  0.87096774
```

Since both intervals include 0, these intervals match the results of our hypothesis test.