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 approximately .573, which is close to the exact value we found.

(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)

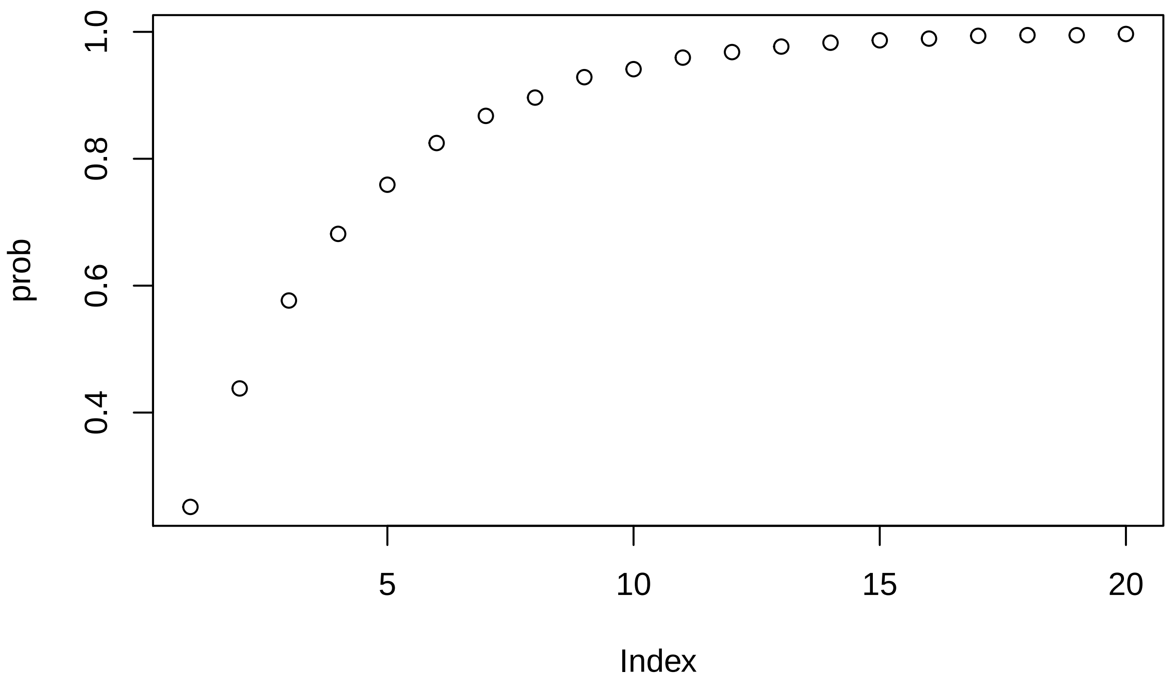
To check this assumption we will do a Monte Carlo Simulation:

```
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)
```

```r
    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 asymptotes.

(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 during the mark and the 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.

We will simulate data from a normal distribution with a mean and variance similar to what was observed.

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

Sample Size	True Diff.	Power
250	10	0.450
10	20	0.044
50	20	0.375
100	20	0.629
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 steps we could take would be to vary the distributional assumptions, as well as the magnitude of the true difference in mean lengths. Then we could compute estimates and determine the amount of bias that could exist in these low-power situations.