Assignment Two

Ryan Honea

10/9/2017

Exercise One

The following are 30 time lapses in minutes between eruptions of Old Faithful geyser in Yellowstone National Park, recorded between the hours of 8 a.m. and 10 p.m. on a certain day, and measured from the beginning of one eruption to the beginning of the next:

68	63	66	63	61	44	60	62	71	62	62	55	62	67	73
72	55	67	68	65	60	61	71	60	68	67	72	69	65	66

A researcher wants to use these data for inference purposes, but is concerned about whether it is reasonable to treat such data as a random sample. Do a hypothesis test for randomness. Also compute the exact and approximate p-value of the test.

Solution

First, we enter the data.

```
ys.geyser <- c(68, 63, 66, 63, 61, 44, 60, 62, 71, 62, 62, 55, 62, 67, 73,
               72, 55, 67, 68, 65, 60, 62, 72, 60, 68, 67, 72, 69, 65, 66)
```

Now, I will convert this into a way that it can be checked with the Wald-Wolfowitz test for randomness. Anything to the left of the median will be considered a zero and anything to the right will be considered a 1. Anything else will be ignored and deleted.

```
ys.geyser.median <- median(ys.geyser)
sequence <- c()
for (i in 1:30) {
  if (ys.geyser[i] < ys.geyser.median) sequence[i] <- 0
  else if (ys.geyser[i] > ys.geyser.median) sequence[i] <- 1
  else sequence[i] <- 2
}
sequence <- sequence[sequence != 2]
```

Below is a series of functions that calculate the Wald-Wolfowitz Test's p-value.

```
ww.num_runs <- function(seq) {
  sum <- 1
  for (i in 1:(length(seq) - 1)) {
    if (seq[i] != seq[i+1]) sum <- sum + 1
  }
  return(sum)
}
ww.p_calc <- function(runs, n, m) {
  if (runs %% 2 == 0) {
     k = runs/2
    numerator <- choose(n-1, k-1)*choose(m-1,k-1)
    denominator <- choose(m + n, n)
    return(2*numerator/denominator)
```

```
  } else {
    k <- runs%/%2
    numerator <- choose(n-1,k)*choose(m-1,k-1) + choose(n-1,k-1)*choose(m-1,k)
    denominator <- choose(m + n, n)
    return(numerator/denominator)
  }
}
ww.test <- function(seq) {
  left <- 0
  right <- 0
  seq <- factor(seq)
  levels(seq) <- c(0,1)
  runs <- ww.num_runs(seq)
  n <- length(seq[seq == 0])
  m <- length(seq[seq == 1])
  for (i in 1:runs) {
    left <- left + ww.p_calc(i, n, m)
  }
  for (i in runs:(n+m)) {
    right <- right + ww.p_calc(i, n, m)
  }
  return(2*min(left, right))
}
```

And so using this function on the Yellowstone data gives the following results.

```
cat("The p-value is", ww.test(sequence))
```

```
## The p-value is 0.5595854
```

Based on this p-value, we fail to reject the null hypothesis that it is not random, and so we can reasonably assume that this is a random sample.

To find the normal approximation's, p-value, we follow a similar procedure.

```
randomness.normapprox <- function(seq) {
  seq <- factor(seq)
  levels(seq) <- c(0,1)
  runs <- ww.num_runs(seq)
  n_1 <- length(seq[seq == 0])
  n_2 <- length(seq[seq == 1])
  n <- n_1 + n_2
  numerator <- runs - 0.5 - 2*n_1*n_2/n
  denominator <- sqrt(2*n_1*n_2*(2*n_1*n_2 - n)/(n^2*(n-1)))
  z = numerator/denominator
  return(pnorm(z))
}
```

Using this function, we can find our approximate p-value.

```
cat("The p-value is", 2*randomness.normapprox(sequence))
```

```
## The p-value is 0.5634352
```

Therefore, again we fail to reject.

Exercise Two

Show that $D_n^- = \max\{\max_{1\leq i\leq n}[F_0(X_{(i)}) - \frac{i-1}{n}], 0\}$

Solution

We know by defition that $D_n^- = \sup_x[F_n(x) - F_0(x)]$ and $F_n(x) = \frac{i}{n}$.

$$\begin{aligned} D_n^- &= \sup_x \left[F_0(x) - \frac{i}{n}\right] \\ &= \max_{0\leq i\leq n} \left[\sup_{X_{(i)}\leq x<X_{(i+1)}} \left[F_0(x) - \frac{i}{n}\right]\right] \end{aligned}$$

The supremememum of the difference is whatever value maximized $F_0(x)$, that is $\sup_{X_{(i)}\leq x<X_{(i+1)}} \left[F_0(x) - \frac{i}{n}\right] = \sup_{X_{(i)}\leq x<X_{(i+1)}} [F_0(x)] - \frac{i}{n}$

$$\begin{aligned} D_n^- &= \max_{0\leq i\leq n} \left[\sup_{X_{(i)}\leq x<X_{(i+1)}} [F_0(x)] - \frac{i}{n}\right] \\ &= \max_{0\leq i\leq n} \left[F_0(X_{(i+1)}) - \frac{i}{n}\right] \\ &= \max_{0\leq i\leq n} \left[F_0(X_{(i)}) - \frac{i-1}{n}\right] \\ &= \max\left\{\max_{1\leq i\leq n} \left[F_0(X_{(i)}) - \frac{i-1}{n}\right], 0\right\} \end{aligned}$$

Note that $\left[F_0(X_{(i+1)}) - \frac{i}{n}\right]$ is rewritten in terms of i and $i-1$ so as to make it similar to D_n^+.

Thus we have our desired result.

Exercise Three

Use the asymptotic distribution of D_n^+ to compute the critical value, $D_{n,\alpha}^+$, of the hypothesis test problem: $H_0 : F(x) = F_0(x)$, $H_a : F(x) > F_0(x)$, based on the sample size $n = 50$ and $\alpha = 0.01$.

Solution

We have from Theorem 4.3.5 ad Corollary 4.3.5.1 in Gibbons and Chakraborti that we can find $D_{n,\alpha}$ from $V = 4nD_n^{+2}$ where V is the χ^2 distribution with 2 degrees of freedom. So, we have that

$$\begin{aligned}
\chi^2_{2,\alpha} &= 4nD_n^{+2} \\
D^+_{n,\alpha} &= \sqrt{\frac{\chi^2_{2,\alpha}}{4n}} \\
D^+_{2,.99} &= \sqrt{\frac{\chi^2_{2,\alpha}}{4 * 50}}
\end{aligned}$$

```
sqrt(qchisq(.99, 2)/200)
```

```
## [1] 0.2145966
```

And so our result is that $D_{2,.99} = 0.2145$.

This can also be done in the following way.

$$\begin{aligned}
P(D_n^+ \leq \frac{d}{\sqrt{n}}) &= 1 - e^{-2d^2} \\
.01 &= e^{-2d^2} \\
\ln .01 &= -2d^2 \\
d &= \sqrt{-\frac{1}{2}\ln .01} \\
D_n^+ &= \frac{d}{\sqrt{n}} \\
&= \sqrt{-frac12n \ln .01} \\
&= .2145
\end{aligned}$$

Exercise Four

Explain why $\omega_n^2 = \int_{-\infty}^{\infty}[F_n(x) - F_0(x)]^2 f_X(x)dx$ can be used to test $H_0 : F(x) = F_0(x), H_a : F(x) \neq F_0(x)$. Also show that the test is distribution free.

Solution

This test essentially calculates the difference in some hypothetical distribution versus an observed (or empirical) distriution. That difference is squared to result in positive values, and thus the higher ω_n^2, the more likely we are to reject the null hypothesis.

We can show that the test is distribution free by showing that the test statistic can be written in terms of just the ordered observations. For the sake of this argument, consider $x_{(0)} = -\infty$ and $x_{(n+1)} = \infty$. A similar argument is made in the proof that the Kolgomorov-Smirnoff test is distribution free.

$$
\begin{aligned}
\omega_n^2 &= \int_{-\infty}^{\infty} \left[F_n(x) - F_0(x)\right]^2 f_0(x)dx \\
&= \int_{-\infty}^{x_{(1)}} \left[F_n(x) - F_0(x)\right]^2 f_0(x)dx + \int_{x_{(1)}}^{x_{(2)}} \left[F_n(x) - F_0(x)\right]^2 f_0(x)dx + ... + \int_{x_{(n)}}^{\infty} \left[F_n(x) - F_0(x)\right]^2 f_0(x)dx \\
&= \sum_{i=0}^{n} \int_{x_{(i)}}^{x_{(i+1)}} \left[F_n(x) - F_0(x)\right]^2 f_0(x)dx \\
&= \sum_{i=0}^{n} \int_{x_{(i)}}^{x_{(i+1)}} \left[\frac{2i+1}{2n} - F_0(x)\right]^2 f_0(x)dx \\
u &= \frac{2i+1}{2n} - F_0(x) \\
du &= -f_0(x)dx \\
\omega_n^2 &= -\sum_{i=0}^{n} \int_{x_{(i)}}^{x_{(i+1)}} u^2 du \\
&= -\sum_{i=0}^{n} \frac{u^3}{3}\Big|_{x_{(i)}}^{x_{(i+1)}} \\
&= -\sum_{i=0}^{n} \left[\frac{\left(\frac{2i+1}{2n} - F_0(x_{(i+1)})\right)^3}{3} - \frac{\left(\frac{2i+1}{2n} - F_0(x_{(i)})\right)^3}{3}\right]
\end{aligned}
$$

So we have shown that ω_n^2 can be written in terms of the order statistics and so it is distribution free. Note that $F_n(x)$ here is $\frac{2i+1}{2n}$ here because we are integrating between two $x_(i)$ and so it is appropriate to consider the average of $F_n(x_{(i)})$ and $F_n(x_{(i-1)})$.

Exercise Five

A random sample of size 13 is drawn from an unknown continuous population $F_X(x)$, with the following results after array

3.5 4.1 4.8 5.0 6.3 7.1 7.2 7.8 8.1 8.4 8.6 9.0

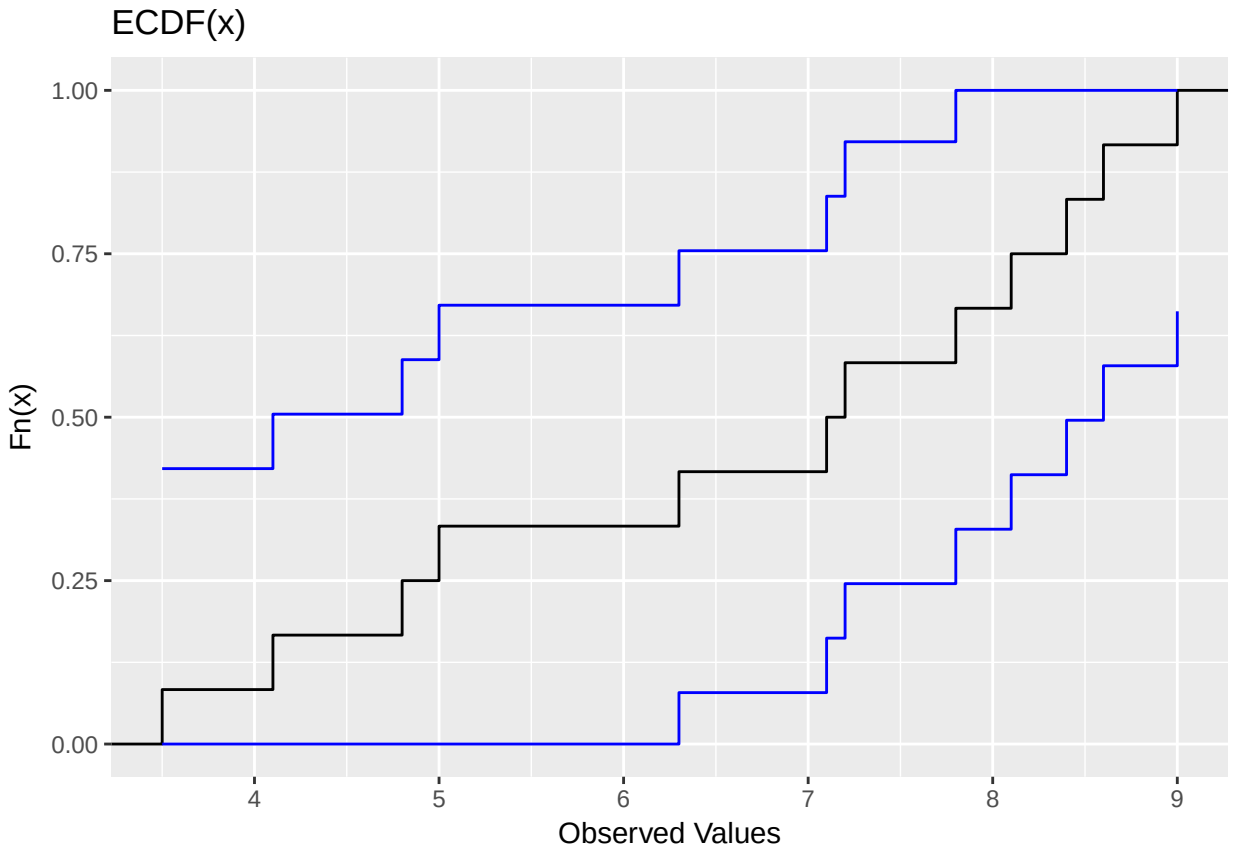
A 90% confidence band is desired for $F_X(x)$. Plot a graph of the empirical distribution function $F_n(x)$ and resulting confidence bands.

Solution

For a sample of size 12 and $\alpha = .10$, we can use a table to find the $D_{n,\alpha}$ value that can be used to create a confidence band. With this table's results, we find the value to be 0.338. So, below, we create the empirical distribution with R's commands and then find the upper and lower confidence bands.

```
x <- c(3.5, 4.1, 4.8, 5.0, 6.3, 7.1, 7.2, 7.8, 8.1, 8.4, 8.6, 9.0)
dist.x <- ecdf(x)
lower <- c()
upper <- c()
for (i in 1:12) {
  lower[i] <- max(dist.x(x[i]) - .338, 0)
  upper[i] <- min(dist.x(x[i]) + .338, 1)
}
pl = data.frame(x = x, y = dist.x(x))
ggplot(pl,aes(x)) +
```

```
  stat_ecdf(geom = "step") +
  geom_step(aes(y=lower), colour="blue") +
  geom_step(aes(y=upper), colour="blue") +
  labs(x = "Observed Values", y = ("Fn(x)")) +
  labs(title = "ECDF(x)")
```



The black lines represents the ECDF where the blue lines represent the confidence bands.

Exercise Six

Plot the distribution function of $N(\mu = 0, \sigma = 1)$. Simulate $n = 10$ number from the distribution and plot the Empirical CDF. Do the same for $n = 50$, $n = 100$, and $n = 500$. Comment on the behavior of the Empirical CDF.

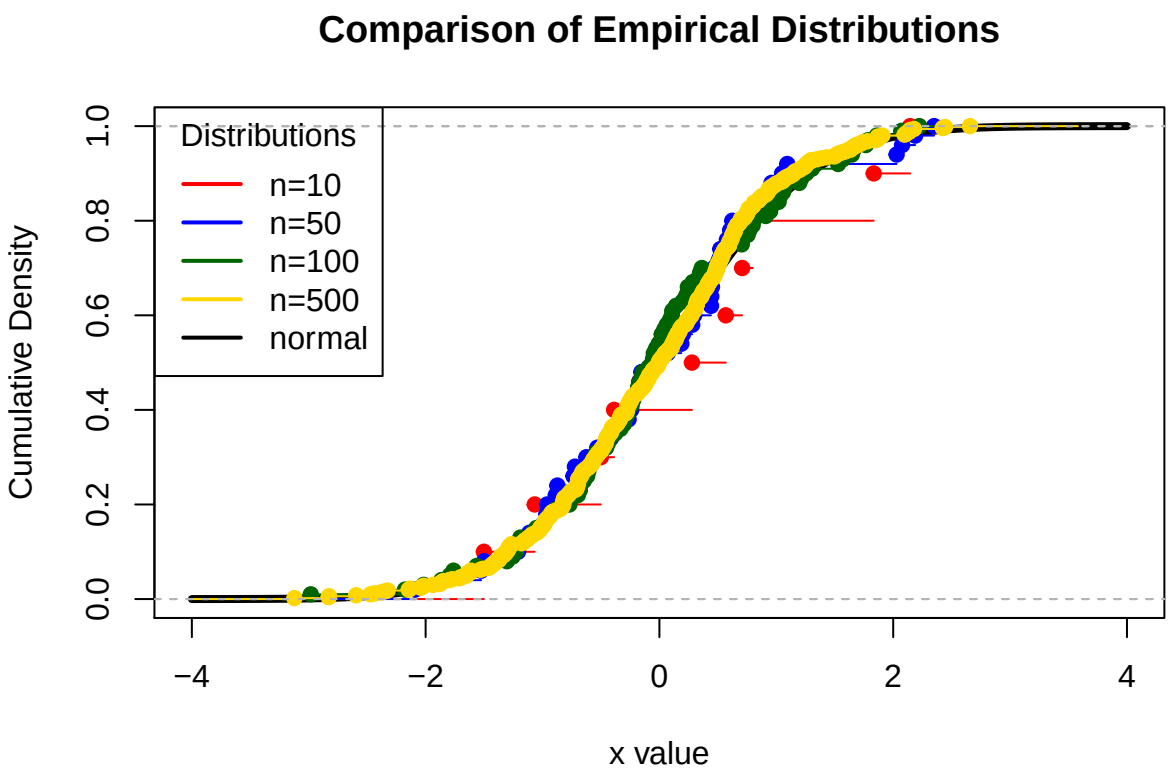
```
x <- seq(-4, 4, length=100)
hx <- pnorm(x)
sim.10 <- ecdf(rnorm(10))
sim.50 <- ecdf(rnorm(50))
sim.100 <- ecdf(rnorm(100))
sim.500 <- ecdf(rnorm(500))

colors <- c("red", "blue", "darkgreen", "gold", "black")
labels <- c("n=10", "n=50", "n=100", "n=500", "normal")

plot(x, hx, type="l", lwd = 4, xlab="x value",
  ylab="Cumulative Density", main="Comparison of Empirical Distributions")
```

```
lines(sim.10, lwd=1, col=colors[1])
lines(sim.50, lwd=1, col=colors[2])
lines(sim.100, lwd=1, col=colors[3])
lines(sim.500, col=colors[4])

legend("topleft", title="Distributions",
  labels, lwd=2, lty=c(1, 1, 1, 1, 1), col=colors)
```



It's clear here that with increasing n, the empirical distribution's CDF approaches that of the normal distribution's CDF.

Exercise Seven

Suppose that the monthly collections for home delivery of the New York Times in a large suburb of New York City are approximately normally distributed with mean \$150 and standard deviation \$20. A random sample of 10 delivery persons in a nearby suburb is taken; the arrayed data for monthly collections in dollars are

90 106 109 117 130 145 156 170 174 190

Test the null hypothesis that the same normal distribution model applies to this suburb, using the most appropriate tests (Use 2 tests at least).

Solution

First, we enter the data.

```
observed <- c(90, 106, 109, 117, 130, 145, 156, 170, 174, 190)
simulated <- rnorm(10,150,20)
```

The two tests used with be the Wald-Wolfowitz test and the Mann-Whitney test which are both below. Note that the code for the Wald-Wolfowitz test is using functions written previously.

```
ww.dist.test <- function(s_1, s_2) {
  n <- length(s_1)
  m <- length(s_2)
  seq <- rbind(cbind(s_1,rep(0,n)),cbind(s_2,rep(1,m)))
  seq <- seq[order(seq[,1]),]
  runs <- ww.num_runs(seq[,2])
  p = 0
  for (i in 2:runs) {
    p <- p + ww.p_calc(i, n, m)
  }
  return(p)
}
```

```
mw.calc_U <- function(seq) {
  sum_right <- 0
  r <- seq
  first_1 <- match(1,r)
  while (length(r) > 1) {
    r <- r[first_1:length(r)]
    sum_right <- sum_right + length(r[r == 0])
    r <- r[2:length(r)]
    first_1 <- match(1, r)
    if (is.na(first_1)) break
  }
  return(sum_right)
}
arrangements <- function(m,n,U) {
  if (U < 0) return(0)
  else if ((U == 0 && m == 0) || (U == 0 && n == 0)) return(1)
  else if (n == 0 || m == 0) return(0)
  else return(arrangements(m,n-1,U) + arrangements(m-1,n,U-n))
}
mw.test <- function(s_1, s_2) {
  n <- length(s_1)
  m <- length(s_2)
```

```
  seq <- rbind(cbind(s_1,rep(0,n)),cbind(s_2,rep(1,m)))
  seq <- seq[order(seq[,1]),]
  U <- mw.calc_U(seq[,2])
  p.value <- 0
  for (i in 0:U) {
    p.value <- p.value + arrangements(m,n,i)
  }
  p.value <- p.value/choose(m+n,m)
  return(p.value)
}
```

Now, we can perform the tests on our values. Note that the Mann-Whitney test will have its value multiplied by two as it is a two-tailed test in this case.

```
cat("The P-Value for the Wald-Wolfowitz Test is", ww.dist.test(observed,simulated),"\n")
```

```
## The P-Value for the Wald-Wolfowitz Test is 0.5859296
```

```
cat("The P-Value for the Mann-Whitney Test is", 2*mw.test(observed,simulated),"\n")
```

```
## The P-Value for the Mann-Whitney Test is 0.3149992
```

Both tests fail to reject the null hypothesis, so we can assume that they follow the same distribution.

Exercise Eight

Each student in a class of 18 is asked to list three people he likes and three he dislikes and label the people 0, 1, 2, 3, 4, 5 according to how much he likes them, with 0 denoting least liked and 5 denoting most liked. From this list, each student selects the number assigned to the person he thinks is the wealthiest of the six. The results in the form of an array are as follows:

$$0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 2 \quad 2 \quad 2 \quad 2 \quad 3 \quad 3 \quad 4 \quad 4 \quad 4 \quad 4 \quad 4 \quad 4 \quad 5$$

Test the null hypothesis that the students are equally likely to select any of the numbers 0, 1, 2, 3, 4, or 5 using the most appropriate test and the 0.05 level of significance.

Solution

I will test this using the Kolgomorov-Smirnoff Test. I will run the simulation 100 times and see how often the null hypothesis is rejected to form a proper analysis.

```
rejections <- 0
student_selections <- c(0,0,0,0,1,2,2,2,2,3,3,4,4,4,4,4,4,5)
for (i in 1:100) {
  sim <- runif(18,0,5)
  test <- ks.test(student_selections, sim)
  if (test$p.value < .05) rejections <- rejections + 1
}
cat("The KS.Test Rejected the Null Hypothesis", rejections, "times.")
```

```
## The KS.Test Rejected the Null Hypothesis 0 times.
```

Based on this result, we can see that this distribution is similar to the uniform and thus each student is equally likely to be assigned any of the numbers.