Assignment One

Ryan Honea

9/8/2017

Exercise One

On the basis of the following random sample of pairs:

X	126	131	153	125	119	102	116	163
Y	120	126	152	129	102	105	100	175

Test at a significance level not exceeding 0.10 the null hypothesis $H_0 : M = 2$ against the alternative $H_1 : M \neq 2$, where M is the median of the continuous and symmetric population of difference $D = X - Y$. Compute the exact probability of type 1 error.

Solution

```
X <- c(126, 131, 153, 125, 119, 102, 116, 163)
Y <- c(120, 126, 152, 129, 102, 105, 100, 175)
D <- X - Y
```

Because the population of D is continuous and symmetric, the Wilcox Signed Rank test will perform stronger than the simple signed test.

```
results <- wilcox.test(D, mu = 2)
cat("P-value for Wilcox test with mu/median = 2 is", results$p.value, "\n")
```

```
## P-value for Wilcox test with mu/median = 2 is 0.888502
```

The results of this test, that is a p-value of .3828 leads us to fail to reject the null hypothesis, that is $M = 2$.

I will, however, also manually compute the p-value of the simple sign test and it's Type I error.

```
successes <- length(D[D > 2])
p <- 2*min(pbinom(4,8,.5),pbinom(8,8,.5) - pbinom(3,8,.5))
cat("The p-value for the sign test is",p)
```

```
## The p-value for the sign test is 1.273437
```

Obviously that p-value is impossible, but it is occurring due to overlap. So, the true p-value for the sign-test is 1 which leads us to fail to reject. We can find the probability of Type I error by finding the probability that we observe a value of successes that would lead us to reject the null hypothesis. Below gives the probabilies of observing various results for different numbers of successes.

```
for (i in 0:8) {
  cat("The probability of", i, "successes is", dbinom(i,8,.5),"\n")
}
```

```
## The probability of 0 successes is 0.00390625
## The probability of 1 successes is 0.03125
## The probability of 2 successes is 0.109375
## The probability of 3 successes is 0.21875
## The probability of 4 successes is 0.2734375
## The probability of 5 successes is 0.21875
## The probability of 6 successes is 0.109375
## The probability of 7 successes is 0.03125
## The probability of 8 successes is 0.00390625
```

The values that would lead to rejecting the null hypothesis at a level of .10 are 0, 1, 7, and 8.

```
typeI <- 1 - (pbinom(6,8,.5) - pbinom(1,8,.5))
cat("The probability of Type I Error is", typeI, "\n")
```

```
## The probability of Type I Error is 0.0703125
```

Exercise Two

Recent studies of the private practices of physicians who saw no Medicaid patients suggested that the median length of each patient visit was 22 minutes. It is believed that the median visit length in practices with a large Medicaid load is shorter than 22 minutes. A random sample of 20 visits in practices with a large Medicaid load yielded, in order, the following visit lengths:

9.4	13.4	15.6	16.2	16.4	16.8	18.1	18.7	18.9	19.1
19.3	20.1	20.4	21.6	21.9	23.4	23.5	24.8	24.9	26.8

Based on these data, is there sufficient evidence to conclude that the median visit length in practices with a large Medicaid load is shorter than 22 minutes?

Solution

We start by loading the data.

```
medicaid <- c(09.4, 13.4, 15.6, 16.2, 16.4, 16.8, 18.1, 18.7, 18.9, 19.1,
              19.3, 20.1, 20.4, 21.6, 21.9, 23.4, 23.5, 24.8, 24.9, 26.8)
```

After loading the data, one only needs to run the sign test specifying that the alternative hypothesis is less than instead of the default two-tailed test.

```
test <- SIGN.test(medicaid, md = 22, alternative = "less")
cat("P-Value for Sign Test is", test$p.value, "\n")
```

```
## P-Value for Sign Test is 0.02069473
```

This p-value suggests that we reject the null hypothesis and conlcude that there is indeed evidence to suggest that the median wait time is less than 22 minutes.

Exercise Three

A psychologist claims that the number of repeat offenders will decrease if first time offenders complete a particular rehabilitation course. You randomly select 10 prisons and record the number of repeat offenders during a two-year period. Then, after first-time offenders complete the course, you record the number of repeat offenders at each prison for another two-year period. The results are shown in the following table. At 0.05 significance level, can you support the psychologist's claim?

Prison	1	2	3	4	5	6	7	8	9	10
Before	21	34	9	45	30	54	37	36	33	40
After	19	22	16	31	21	30	22	18	17	21

Solutions

We start by loading the data.

```r
before <- c(21, 34, 09, 45, 30, 54, 37, 36, 33, 40)
after  <- c(19, 22, 16, 31, 21, 30, 22, 18, 17, 21)
diff <- after - before
```

In this case, we can support the psychologist's claim if the median of the difference is less than 0. This would suggest that there has been an increase.

```r
test <- SIGN.test(diff, md = 0, alternative = "less")
cat("P-Value for Sign Test is", test$p.value, "\n")
```

```
## P-Value for Sign Test is 0.01074219
```

Based on the p-value for the sign test, there is evidence to support the psychologist's claim.

Exercise Four

Develop a hypothesis test procedure for testing a 3rd Quantile (Q3). Also find a method to build confidence interval for Q3.

Solution

First, we develop a test statistic utilizing the following pieces of information

$$\phi_i = \begin{cases} 1, & \text{if } x_i > \theta \\ 0, & \text{if } x_i < \theta \end{cases}, \qquad S = \sum_{i=1}^{n} \phi_i \sim \text{Bin}(n, .25)$$

And our rejection regions is

$$R = \begin{cases} s > c_1, & \text{if } H_a : \theta > \theta_0 \\ s < c_2, & \text{if } H_a : \theta < \theta_0 \\ s > c_1 \cup s < c_2, & \text{if } H_a : \theta \neq \theta_9 \end{cases}$$

For purpose of example, the below dataset will be used.

$$1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9$$

Typically with a two-tailed test, one would just find the p-value for one side of your results and then multiply that result by two for a p-value. That does not work here due to the binomial distribution for $p = .25$ not being symmetric. In the case of the dataset above, let's choose $H_0 : p_.75 = 6.5$ and $H_a : p_.75 \neq 6.5$. Let's look at the probability of each number of positives occuring given $S \sim \text{Bin}(n, .25)$.

```r
for (i in 0:9) {
  cat("P of", i, "successes is", sprintf("%.3f", dbinom(i,9,.25)),"\n")
}
```

```
## P of 0 successes is 0.075
## P of 1 successes is 0.225
## P of 2 successes is 0.300
## P of 3 successes is 0.234
## P of 4 successes is 0.117
## P of 5 successes is 0.039
## P of 6 successes is 0.009
## P of 7 successes is 0.001
## P of 8 successes is 0.000
## P of 9 successes is 0.000
```

It's clear that this is not a symmetric tail, and that there will be cases where the two-tailed test has significance equal to the significance of a one-tail test.

Creating a confidence interval is by finding for what values of r and s does the following hold,

$$P(x_{(r)} \leq \theta \leq x_{(s)}) \leq 1 - \alpha$$

There are obviously going to be several cases where this is true, and so the objective is to maximize $P(x_{(r)} \leq \theta \leq x_{(s)})$ subject to the condition that is is $\leq 1 - \alpha$.

Exercise Five

To test $H_0 : M = M_0$ vs. $H_a : M > M_0$, plot the approximate (Normal Approximation) power function (as a function of the $p = P(X > M_0|H_a)$ for sign test. From the power function, also compute a sample size determination formula. Plot that function agaisnt sample size n when $p = 0.2$.

To plot the power, we need to put the function in a more expressable form. Specifically,

$$P_{H_a}(S > c) = P_{H_a}(\frac{s - np}{\sqrt{np(1-p)}} > \frac{c - np}{\sqrt{np(1-p)}})$$
$$P_{H_a}(S > c) = P_{H_a}(z > \frac{c - np}{\sqrt{np(1-p)}})$$

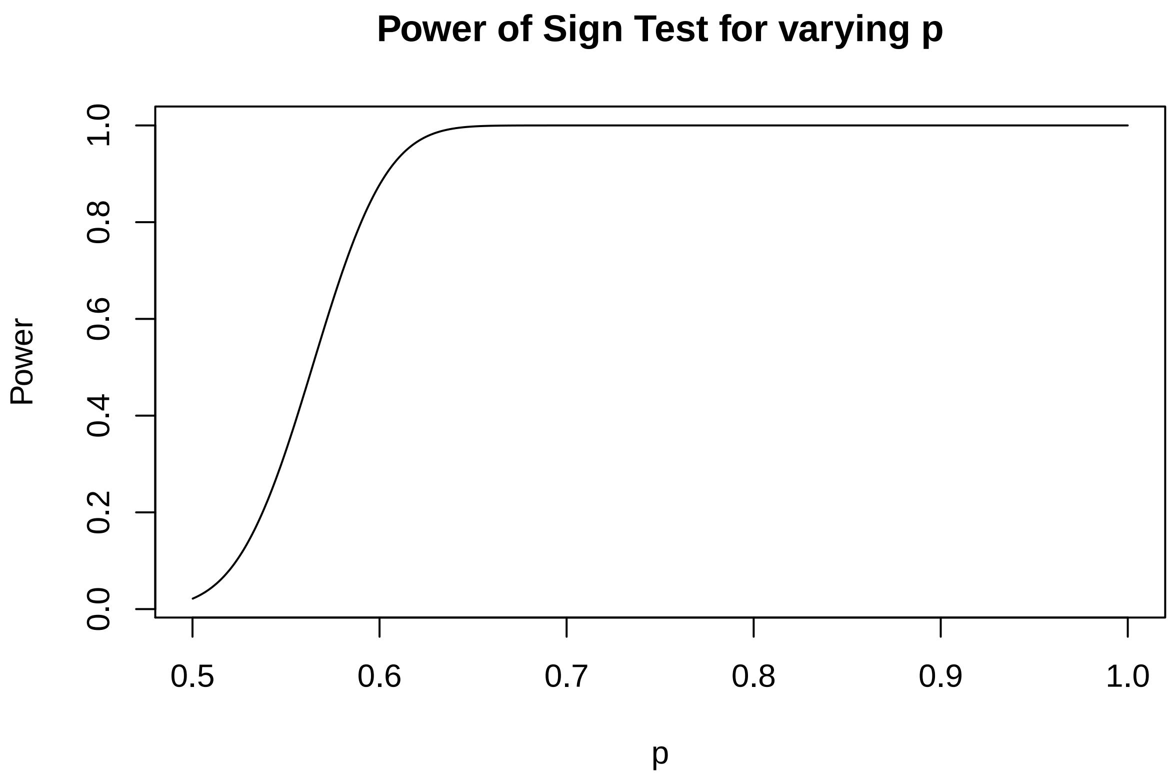
We can find our value for c if we set $\alpha = 0.05$ and $n = 250$ and use R to find our value for c under H_0.

```r
cat("c = ", qnorm(.975, 250*.5, sqrt(250/4)))
```

```
## c =  140.4949
```

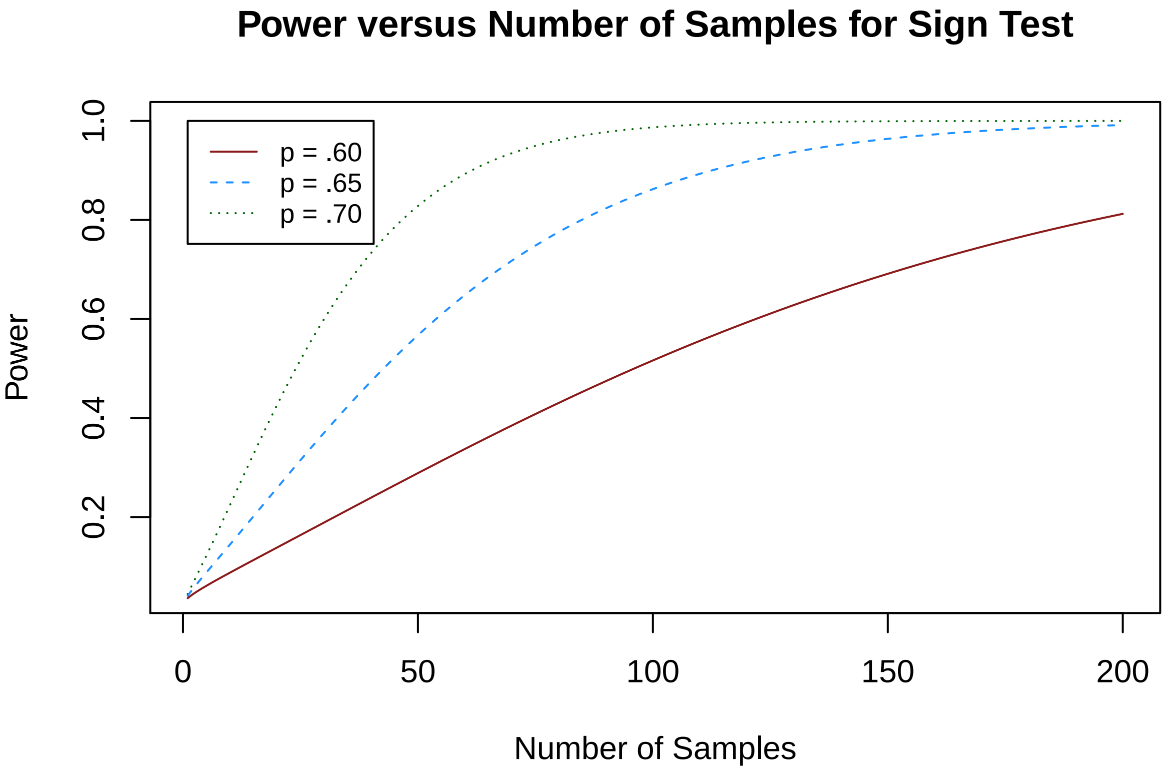
So now, we can create a collection of z-scores based on a range of probabilities and obtain our powers for those probabilities.

```r
z = function(n,c,p) {
  val = (c - n*p)/(sqrt(n*p*(1-p)))
  return(val)
}
power <- matrix(NA, nrow = 5000, ncol = 1)
prob <- matrix(NA, nrow = 5000, ncol = 1)
n = 250
c = 141
for (i in 1:5000) {
  prob[i,] <- .5 + i/10000
  z_val <- z(n, c, prob[i,])
  power[i,] <- (1 - pnorm(z_val, 0, 1))
}
plot(prob, power, main = "Power of Sign Test for varying p", xlab = "p",
     ylab = "Power", type = "l")
```



We can do the same for a plot with varying n where we set p to different levels and keep $\alpha = .05$.

```
probabilities = c(.600, .650, .700)
n = seq(1,200, by = 1)
power = matrix(NA, nrow = 200, ncol = 1)
z_val = matrix(NA, nrow = 200, ncol = 1)
prob = matrix(NA, nrow = 200, ncol = 1)
c = matrix(NA, nrow = 200, ncol = 1)
dfs <- list()
for (j in 1:3) {
  for (i in 1:200) {
    c[i,] = qnorm(.975, n[i]*.5, sqrt(n[i]*.5*(1-.5)))
    z_val[i,] = z(n[i], c[i,], probabilities[j])
    power[i,] = (1 - pnorm(z_val[i,], 0, 1))
    prob[i,] = probabilities[j]
  }
  dfs[[j]] <- data.frame(n, power, prob)
}
plot(dfs[[1]]$n, dfs[[3]]$power,
     main = "Power versus Number of Samples for Sign Test",
     xlab = "Number of Samples", ylab = "Power",
     col = "darkgreen", type = "l", lty=3)
lines(dfs[[1]]$n, dfs[[2]]$power, col = "dodgerblue", lty = 2)
lines(dfs[[1]]$n, dfs[[1]]$power, col = "firebrick4", lty = 1)
legend(1, legend=c("p = .60", "p = .65", "p = .70"),
       col=c("firebrick4","dodgerblue","darkgreen"), lty=1:3, cex=0.8)
```


Power versus Number of Samples for Sign Test
Power
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
Number of Samples
p = .60
p = .65
p = .70