Applications of Econometrics and Data Science Methods Pset1

Jack Ogle and Izzy Allum

Problem 1 (a)

Propose a sequence of random variables that converges in probability to p. This sequence must be a function that maps sample sizes (i.e. natural numbers) to random variables. Hint: Recall the weak law of large numbers.

From the Law of Large numbers and the Central Limit Theorem we know that a sequence of random variables θ_N converges in probability to θ if for any $\epsilon > 0$ and $\delta > 0$ there exists $N^* = N^*(\epsilon, \delta)$ such that for all $N > N^*$

Where θ_N is a function of $(y_i, X_i, i = 1, ...N)$

$$Pr(|\theta_N - \theta| < \epsilon) > 1 - \delta$$

If the limit exists we write that θ_N converges in probability to θ. And because p is the percentage of COVID-19 infected in Chicago we also know that it has a Bernoulli distribution. Our end goal is to estimate the number of total test kits needed for the entire population of Chicago. Therefore we want to estimate the postitivity rate of Chicago. We can't observe the entire population, but we can observe a sample. We use the sample to estimate p and we know that through the law of large numbers and the central limit theorem if you substitute θ for p. Our sample Pn will converge in probability to P.

(b)

The classical Central Limit Theorem assumptions are that:

Let

$$Z_N = \frac{\bar{X}_N - E[\bar{X}_N]}{\sqrt{Var(\bar{X}_N)}}$$

where $\bar{X}_N = N^{-1} \sum_{i=1}^{N} X_i$, the sample mean.

(i) Let X_i be iid with $E[X_i] = \mu$

Since all X_i are Bernoulli and mutually independent this condition is satisfied.

(ii) and $Var(X_i) = \sigma^2$ meaning that our variance is finite. Then Z_N in distribution to a normal distribution of N(0,1). And because our $Var(X_i) = p(1-p)$ our variance.

(iii)

With the Central Limit Theorem even though the population of Chicago might not be normal we can extend the Single Sample Hypothesis test for normally distributed populations to those that are not normally distributed. We have a sample of size n, where n is sufficiently large. The null hypothesis is: $H_0 : p = p_0$.

The alternative hypothesis is $H_1 : p > p_0$

When we assume the null hypothesis, we know from CLT that the sample mean has a normal distribution.

We will assume that the level of significance is α. Because our sample takes on a bernoulli distribution we know that $X_N \sim$ Binomial (p, $\frac{p(1-p)}{n}$)

From b) we know that our variables and sample satisfy CLT assumptions and can therefore conclude that our

$$\frac{X_N - p}{\sqrt{\frac{p(1-p)}{n}}}$$

converges in distribution to a standard normal distribution N(0,1).

Now lets evaluate our null hypotheses H_0 which states that

$$\frac{X_N - p_0}{\sqrt{\frac{p_0(1-p_)}{n}}}$$

converges in distribution to a standard normal distribution N(0,1).

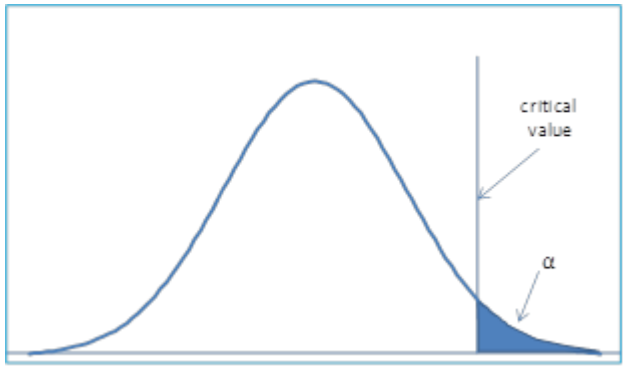
Because the alternative hypothesis contains an greater than inequality we will use a right hand test. To find the critical value we must find the region that is greater than $Z_{1-\alpha}$. The rejection region is

$$Z_{1-\alpha} < \frac{X_N - p_0}{\sqrt{\frac{p_0(1-p_)}{n}}}$$

- this is our test statistic.

Solving for our sample mean. Our official rejection region is:

$$X_n > \frac{Z_{1-\alpha}\sqrt{p_0(1-p_0)}}{\sqrt{n}} + p_0$$



(d) Without the asymptomatic results we cannot use CLT. We must use the fact that our sample has a bernoulli distribution. This means that X_n are binomial like we established in c). $X_N \sim$ Binomial (p, $\frac{p(1-p)}{n}$)

We reject the null hypothesis when $X_n > C_b$ assuming that the following condition is met:

$$\sum_{k=c_b}^{n} \binom{n}{k} p_0^k (1-p_0)^{n-k} \leq \alpha$$

(e)

In this question we need to solve for the number of test kits here. In order to do that we need to find an n (number of test kits) such that the critical value for the type I and the type II errors are equal to each other. The type I errors are those that fall under H_0 and the type II errors are those that fall under H_1

$$\frac{Z_{1-\alpha}\sqrt{p_0(1-p_0)}}{\sqrt{n}} + p_0$$

=

$$\frac{Z_{1-\alpha}\sqrt{p_1(1-p_1)}}{\sqrt{n}} + p_1$$

Now we solve for n and we get:

$$\frac{Z_{1-\alpha}\sqrt{p_0(1-p_0)} - \sqrt{p_1(1-p_1)}}{p_1 - p_2})^2 = N$$

(f)

Using the rejection region from d)

$$\sum_{k=c_b}^{n} \binom{n}{k} p_0^k (1-p_0)^{n-k} \leq \alpha$$

We need to minimize N number of test kits such that

$$\sum_{k=c_b}^{n} \binom{n}{k} p_1^k (1-p_1)^{n-k} \leq \beta$$

(g)

```r
N <- 1:1000
p0 = c(0.001,0.001,0.001,0.05,0.05,0.05,0.001,0.001,0.001,0.05,0.05,0.05)
p1 = c(0.1, 0.15, 0.2,0.1,0.15,0.2,0.1,0.15,0.2,0.1,0.15,0.2)
beta = c(0.1,0.1,0.1,0.1,0.1,0.1,0.2,0.2,0.2,0.2,0.2,0.2)

# MC model for generating the number of test kits needed
## given certain permutations
g_mc_model = function(alpha, beta, p0, p1) {
  # defining the replication function with a random sample
  # also ensuring that it has a bernoulli distribution
  r_sample <- replicate(10000, mean(rbinom(N, 1, p0)))
  # Evaluating the test statistic critical value
  mean <- mean(r_sample)
  sd <- sd(r_sample)
  critv <- qnorm(1-alpha, mean, sd)
  p1Beta <- pnorm(critv, p1, sqrt(p1*(1-p1)/N))
  result <- min(which(p1Beta <= beta))
  return(result)
}

for (i in 1:12) {
```

```
  print(g_mc_model(0.05, beta[i], p0[i], p1[i]))
  }
```

```
## [1] 16
## [1] 10
## [1] 7
## [1] 100
## [1] 27
## [1] 14
## [1] 7
## [1] 5
## [1] 3
## [1] 43
## [1] 12
## [1] 6
```

(h)

```
#let n := number of test kits purchase
# in particular, the number of test kits determines the sample size
# we will use the smallest n that satisfies alpha < 0.05 and beta < 0.1 (or 0.2 depending on the question)
# starting from n = 1.

#function binomial probability
# pbinom(infected, n, p0)
# pbinom is giving me issues
p_atleast_m <- function(m, n, p0) {
    prob <- 0
    for (j in m:n) {
        prob <- prob + (choose(n, j) * (p0 ^ j) * ((1 - p0) ^ (n - j)))
    }
```

```
    return(prob)
}

simulation <- function(p0, p1, beta, alpha) {
    #start from n = 1
    n = 1
    while (TRUE) {
        #   determine the threshold to reject null
        #   threshold is determined by j < n
        #   such that the probability of observing
        #   at most j infected < alpha
        for (thres in 0:n) {
            if (p_atleast_m(thres, n, p0) > alpha){
                next
            } else {
                break
            }
        }
        threshold <- thres - 1

        #run the simulation 1000 times for sample size n
        infected <- c()
        for (i in 1:1000) {
            samp <- rbinom(n, 1, p1)
            if (sum(samp) > threshold) {
                infected <- c(infected, 1)
            } else {
                infected <- c(infected, 0)
            }
        }

        #type 2 error is bounded by beta
```

```
        # thus if we reject the null better than 1-beta
        # then we are done
        if (sum(infected) / 1000 >= 1 - beta) {
            print(c(threshold, n))
            return (n)
        } else {
            n <- n + 1
        }
    }
}
```

```
## Simulation results
for (i in 1:12) {
    print(simulation(p0[i], p1[i], beta[i], 0.05))
}
```

Results for e:

1) 24
2) 15
3) 10
4) 17
5) 231
6) 7
7) 77
8) 4

9) 37

10) 16

11) 11

12) 7

Results for f:

1) 22

2) 15

3) 11

4) 226

5) 76

6) 39

7) 16

8) 11

9) 8

10) 154

11) 52

12) 27

Note we also collaborated with Zach Yung, Dila Sasmaz, Benamin Jacobs, Anahita Gogia, Armand Dang, and Rahma Safraoui on this question. We talked through the theory behind the question together.