ECMA 31320 Pset 1

Chris Liao

Noah Sobel-Lewin

April 4, 2022

1

a

Our sequence is $\overline{X}_{(N)} = \frac{1}{N}\sum_{i=1}^{N} X_i$. As N increases in size, then by the weak law of large numbers, $\overline{X}_{(N)} \xrightarrow{p} \mathbb{E}[X] = p$

b

The conditions of the "classical" Central Limit Theorem are that $(X_1, \cdots, X_N)$ are iid and that X_i has finite mean and variance. We know that each X_i is Bernoulli(p) , and that the infected statuses of each individual are mutually independent so our sample $(X_1, \cdots, X_N)$ is iid. Further, because X_i are Bernoulli(p), they have finite mean (p) and variance ($p(1-p)$).

c

First, here are some basic facts

$$\mathbb{E}[X_i] = p \quad \mathbb{E}[\overline{X}_{(N)}] = p$$

$$\text{Var}[X_i] = p(1-p) \quad \text{Var}[\overline{X}_{(N)}] = \frac{p(1-p)}{n}$$

Next, since $(X_1, \cdots, X_N)$ is iid, by the central limit theorem,

$$\frac{\overline{X}_{(N)} - p_0}{\sqrt{\frac{p_0(1-p_0)}{N}}} \equiv Z_N \xrightarrow{d} \mathcal{N}(0,1)$$

Therefore we reject the null hypothesis if our test statistic

$$\frac{\overline{X}_{(N)} - p_0}{\sqrt{\frac{p_0(1-p_0)}{N}}} \equiv Z_N > \Phi^{-1}(1-\alpha)$$

which is equivalent to

$$\overline{X}_{(N)} > p_0 + \sqrt{\frac{p_0(1-p_0)}{N}}\Phi^{-1}(1-\alpha)$$

Thus, the rejection region associated with this test is $(p_0 + \sqrt{\frac{p_0(1-p_0)}{N}}\Phi^{-1}(1-\alpha), 1]$ since $X_{(N)}$ is bounded above by 1.

d

Define

$$k = \sum_{i=1}^{N} \mathbb{I}[X_i = 1]$$

where k is the number of infected people in our sample. We know that the distribution of the number of infected people is distributed Binomial(N, p). Our goal therefore is to find the smallest value k such that

$$\sum_{i=k}^{N} \binom{N}{i} p_0^i (1-p_0)^{N-i} < \alpha$$

Define

$$k_{min} = \min\left\{\sum_{i=k}^{N} \binom{N}{i} p_0^i (1-p_0)^{N-i} < \alpha \mid k \in \mathcal{N}\right\}$$

Then, we reject the null hypothesis if $k > k_{min}$. We also know that this is true:

$$\overline{X_{(N)}} = \frac{1}{N}\sum_{i=1}^{N} X_i = \frac{k}{N} > \frac{k_{min}}{N}$$

Thus, we can rewrte the statement as the following - we reject the null hypothesis if

$$\overline{X_{(N)}} > \frac{k_{min}}{N}$$

As such, the rejection region associated with this test is $(\frac{k_{min}}{N}, 1]$ as $X_{(N)}$ is bounded above by 1.

e

Under the assumption that the null hypothesis is false and that $p = p_1$ is true, the definition of having a type II error of β is that

$$\beta \leq \mathbb{P}\left(\overline{X_{(N)}} < p_0 + \sqrt{\frac{p_0(1-p_0)}{N}}\Phi^{-1}(1-\alpha)\right)$$

or that the probability our test statistic does not fall in the rejection region is less than or equal to beta, when the null hyopthesis is false. Since $(X_1, \cdots, X_N)$ is iid, by the central limit theorem then we can do the following

$$\mathbb{P}\left(\overline{X_{(N)}} < p_0 + \sqrt{\frac{p_0(1-p_0)}{N}}\Phi^{-1}(1-\alpha)\right) = \beta \iff$$

$$\mathbb{P}\left(\frac{\overline{X_{(N)}} - p_1}{\sqrt{\frac{p_1(1-p_1)}{N}}} < \frac{p_0 - p_1 + \sqrt{\frac{p_0(1-p_0)}{N}}\Phi^{-1}(1-\alpha)}{\sqrt{\frac{p_1(1-p_1)}{N}}}\right) = \beta \iff$$

where we know $\frac{\overline{X_{(N)}} - p_1}{\sqrt{\frac{p_1(1-p_1)}{N}}} \sim \mathcal{N}(0,1)$ so the next line follows by the CLT

$$\Phi\left(\frac{p_0 - p_1 + \sqrt{\frac{p_0(1-p_0)}{N}}\Phi^{-1}(1-\alpha)}{\sqrt{\frac{p_1(1-p_1)}{N}}}\right) = \beta \iff$$

$$\sqrt{\frac{p_0(1-p_0)}{N}}\Phi^{-1}(1-\alpha) - \Phi^{-1}(\beta)\sqrt{\frac{p_1(1-p_1)}{N}} = p_1 - p_0 \iff$$

$$\sqrt{N} = \frac{\sqrt{p_0(1-p_0)}\Phi^{-1}(1-\alpha) - \Phi^{-1}(\beta)\sqrt{p_1(1-p_1)}}{p_1 - p_0} \iff$$

$$N = \left(\frac{\sqrt{p_0(1-p_0)}\Phi^{-1}(1-\alpha) - \Phi^{-1}(\beta)\sqrt{p_1(1-p_1)}}{p_1 - p_0}\right)^2$$

Thus, if $H_1 : p = p_1 > p_0$ were true then I would purchase N test kits where

$$N = \left\lceil \left(\frac{\sqrt{p_0(1-p_0)}\Phi^{-1}(1-\alpha) - \Phi^{-1}(\beta)\sqrt{p_1(1-p_1)}}{p_1 - p_0}\right)^2 \right\rceil$$

f

First, recall from part *d* that for every N there is some k_{min} where $(\frac{k_{min}}{N}, 1]$ is the rejection region. Separately, by the definition of type II error, we also want that

$$\sum_{i=1}^{k_{min}} \binom{N}{i} p_1^i (1-p_1)^{N-i} \leq \beta$$

Thus, our goal is to find the smallest N such that

$$\sum_{i=1}^{k_{min}} \binom{N}{i} p_1^i (1-p_1)^{N-i} \leq \beta$$

where $k_{min} = \min\{\sum_{i=k}^{N} \binom{N}{i} p_0^i (1-p_0)^{N-i} < \alpha \mid k \in \mathcal{N}\}$. This problem devolves into solving a two-system set of equations with two unknowns, k_{min} and N.

g

Here is how we computationally identified N. First, we looped through a set of potential N from 1 to 300. Then, for each N, we drew 1000 samples of size N from a Bernoulli distribution of p_0 and found the sample mean for each sample. Then, using our collection of sample means we found the lower bound of the rejection region - the (1-α)-th percentile of our collection of sample means, henceforth known as $\overline{X_{(\alpha)}}$. Next, we drew 1000 samples of size N from a Bernoulli distribution of p_1 and took the sample mean. Using our collection of new sample means, we found that smallest value of the left-hand side of a rejection region satisfying a type II error of β, henceforth known as $\overline{X_{(\beta)}}$. Our value of N would be valid if $\overline{X_{(\beta)}} > \overline{X_{(\alpha)}}$ as that means that at least $1-\beta$ of our observations were rejected. Our aim is to find the smallest valid N such that we can construct a threshold that will guarantee type I error is less than α and type II error is less than β given p_0 and p_1.
See section *h* with a table with all our results

h

Table 1

	α	β	p_0	p_1	Monte Carlo	Binomial (No Asymptotic)	CLT
1	0.050	0.100	0.001	0.100	22	22	20
2	0.050	0.100	0.001	0.150	15	15	12
3	0.050	0.100	0.001	0.200	11	11	9
4	0.050	0.100	0.050	0.100	232	233	221
5	0.050	0.100	0.050	0.150	76	77	67
6	0.050	0.100	0.050	0.200	38	38	34
7	0.050	0.200	0.001	0.100	15	16	10
8	0.050	0.200	0.001	0.150	10	10	6
9	0.050	0.200	0.001	0.200	8	8	4
10	0.050	0.200	0.050	0.100	158	169	150
11	0.050	0.200	0.050	0.150	52	52	44
12	0.050	0.200	0.050	0.200	27	27	22

In general, the comparisons are quite similar, especially across smaller estimates of N. The Binomial and Monte Carlo estimates are especially similar for estimates $N < 60$. However, the CLT estimates tend to consistently be less than the Binomial estimates and the difference between them grows as beta grows. In general, the Monte Carlo and Binomial results are pretty similar, although for large N and large β then the difference between them also grows.

Problem 2

a

β_{OLS} solves the best linear prediction problem of Y given X. Hence,

$$\beta_{\text{OLS}} = \arg\min_b \mathbb{E}[(Y - X'b)^2]$$

By the first order conditions of this minimization problem we know:

$$\begin{aligned} 0 &= \frac{\partial}{\partial \beta}\mathbb{E}[(Y - X'\beta)^2] \\ &= \mathbb{E}[\frac{\partial}{\partial \beta}(Y - X'\beta)^2] \\ &= 2\mathbb{E}[XY - XX'\beta] \end{aligned}$$

$$= 2(\mathbb{E}[XY] - \mathbb{E}[XX']\beta)$$

Where the first equality follows by first order conditions. The second equality from the dominated convergence theorem. The third equality from taking the derivative. The fourth from the linearity of expectation.

$$0 = 2(\mathbb{E}[XY] - \mathbb{E}[XX']\beta) \iff \mathbb{E}[XY] = \mathbb{E}[XX']\beta$$

b

The gram matrix, $G \equiv \mathbb{E}[XX']$, is a linear map mapping from $\mathbb{R}^p \to \mathbb{R}^p$. G is not invertible, so therefore $\text{Rank}(G) < p$. By the Rank-nullity theorem, therefore $\text{Nullity}(G) > 0$. This implies that the set of vectors, S, for which each $v \in S$ satisfies $\mathbb{E}[XX']v = 0$, is infinitely large because the dimension of the space of vectors for which this is the case is non-zero.

c

Yes. Consider $v_1 \neq 0_p \in S$. Let $\beta_1 \equiv \beta_0 + v_1$. By construction, $\beta_1 \neq \beta_0$. Now,

$$\mathbb{E}[XX']\beta_1 = \mathbb{E}[XX'](\beta_0 + v_1) = \mathbb{E}[XX']\beta_0 + \mathbb{E}[XX']v_1 = \mathbb{E}[XX']\beta_0 = \mathbb{E}[XY]$$

Where the third equality follows from the fact that $v_1 \in S$. We have found two coefficients, $\beta_1 \neq \beta_0$, that satisfies the equation.

d

1. We are interested in the statistical relationship between an individual's wages (Y) and whether or not each parent attended college for modern-day individuals.

2. Let,

   $$X \equiv \begin{bmatrix} 1 & X_1 & X_2 \end{bmatrix}'$$

   Where X_1 is an indicator if an individual's mother went to college and X_2 is an indicator if an individual's father went to college. Because of matching, it is likely that college-educated individuals tend to marry college-educated individuals and non-college-educated individuals tend to marry non-college-educated individuals. Therefore, in a small sample, it is not implausible to find that $\boldsymbol{x_1} = \boldsymbol{x_2}$ in the data — where $\boldsymbol{x_i}$ is a vector of realizations of X_i especially if we consider the modern era where an increasing proportion of the population is college educated.

3. We are interested in the education of each parent separately. Possibly, maternal or paternal education might have a differing impact on a child's wages because one's relationship with the mother is frequently different from their relationship with their father and different patents interact with the child differently, which creates variation in future wages. We want to include a constant because we are interested in the expected value of one's wage if neither of their parents went to college and because we expect that this value is nonzero.

4. If one had unlimited data on X and Y, they could determine the population regression coefficient on X because $\mathbb{E}[XX']$ is invertible because, while highly correlated, X_1 is not a linear function of X_2. Therefore, using the equation above,

$$\beta = \mathbb{E}[XY]\mathbb{E}[XX']^{-1}.$$

3

a

$$\begin{aligned}
\boldsymbol{X'X} &= \boldsymbol{X'}\begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{X'}x_1 & \cdots & \boldsymbol{X'}x_n \end{bmatrix} \\
&= \begin{bmatrix} \begin{bmatrix} x_1'x_1 \\ \vdots \\ x_n'x_1 \end{bmatrix} & \cdots & \begin{bmatrix} x_1'x_n \\ \vdots \\ x_n'x_n \end{bmatrix} \end{bmatrix} \\
&= \begin{bmatrix} \sum_{i=1}^N x_{1i}x_{1i} & \dots & \sum_{i=1}^N x_{1i}x_{Ni} \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^N x_{Ni}x_{1i} & \dots & \sum_{i=1}^N x_{Ni}x_{Ni} \end{bmatrix} \\
&= \sum_{i=1}^N \begin{bmatrix} x_{1i}x_{1i} & \dots & x_{1i}x_{Ni} \\ \vdots & \ddots & \vdots \\ x_{Ni}x_{1i} & \dots & x_{Ni}x_{Ni} \end{bmatrix} \\
&= \sum_{i=1}^N x_i x_i'
\end{aligned}$$

b

Let $\boldsymbol{X}$ be a data matrix $\in \mathbb{R}^{T\times K}$ with $x_i' \in \mathbb{R}^K$ for $i \in \{1, \ldots, T\}$

$$\boldsymbol{X} \equiv \begin{bmatrix} x_1 \\ \vdots \\ x_T \end{bmatrix}$$

Now, define $X \in \mathbb{R}^K$ being realizations of X - note that X is the population analog of x_i'.

$$\mathbb{E}[XX'] = \mathbb{E}\begin{bmatrix} x_1x_1 & \ldots & x_1x_K \\ \vdots & \ddots & \vdots \\ x_Kx_1 & \ldots & x_Kx_K \end{bmatrix}$$

Note that in the expression above, x_i refer to components of X in the population. The sample analogue for the second expression is naturally

$$\frac{1}{T}\sum_{i=1}^{T}\begin{bmatrix} x_{1i}x_{1i} & \ldots & x_{1i}x_{Ti} \\ \vdots & \ddots & \vdots \\ x_{Ti}x_{1i} & \ldots & x_{Ti}x_{Ti} \end{bmatrix} = \frac{1}{T}\sum_{i=1}^{T} x_i'x_i$$

We have shown above that $\sum_{i=1}^{T} x_i'x_i = \boldsymbol{X'X}$ so the expression above equals

$$\frac{1}{T}\boldsymbol{X'X}$$

Therefore the sample analogue of $\mathbb{E}[XX']$ is $\frac{1}{T}\boldsymbol{X'X}$.