Extra Credit 2: A Hypergeometric variable related to the Covid study Due: Tuesday April 29 (9pm)

Michael J. Neely
University of Southern California

I. Analysis

The paper [1] gives data on rapid antigen testing for covid in asymptomatic people in Canada during 2021. A group of $N = 1322$ people tested positive. Of these, $n = 1103$ volunteered for a more extensive PCR test that can be viewed as "ground truth" for determining if they truly have covid. Let K be the (unknown) number within the group of 1322 people who tested positive and who truly had covid. Define the (unknown) parameter p:

$$p = \frac{K}{N} = \frac{K}{1322}$$

Assume that the $n = 1103$ people in the smaller group were selected with all $\binom{N}{n}$ choices equally likely. *In this problem we treat N, n, K as fixed and nonrandom values* ($N = 1322, n = 1103$, while K is unknown). Define X as the random number of people within the smaller group of size $n = 1103$ who actually have covid. The randomness of X arises purely because we randomly choose a subset of n people from the larger set of N people. We want to understand the *confidence* we can have about the approximation:

$$\frac{X}{1103} \approx \frac{K}{1322} = p$$

a) Explain why $X \in \{\max\{0, n+K-N\}, \ldots, \min\{n, K\}\}$. The random variable X is a *hypergeometric random variable*. In terms of general constants N, K, n, compute

$$P[X = x] = \frac{?}{\binom{N}{n}} \quad \forall x \in \{\max\{0, n+K-N\}, \ldots, \min\{n, K\}\}$$

State what this is (in terms of K) for the case $N = 1322, n = 1103$. While this is the exact PMF, it can be difficult to work with.

b) For $i \in \{1, \ldots, N\}$, let $k_i \in \{0, 1\}$ be the (nonrandom) binary value that is 1 if person i has covid, and 0 else; let $C_i \in \{0, 1\}$ be the Bernoulli random variable that is 1 if and only if person i is chosen for the smaller group of size n. Explain why:

$$\begin{aligned} K &= \textstyle\sum_{i=1}^N k_i \\ n &= \textstyle\sum_{i=1}^N C_i \\ X &= \textstyle\sum_{i=1}^N k_i C_i \end{aligned}$$

Define $q = \mathbb{E}[C_i]$ (by symmetry it is the same for all $i \in \{1, \ldots, N\}$). Take expectations of the above equalities to find the value of q and to prove that $\mathbb{E}[X] = np$. The estimate $\frac{X}{n}$ is said to be an *unbiased estimator* of p because its expectation is exactly p.

c) Argue that:

$$\begin{aligned} K^2 &= \left(\textstyle\sum_{i=1}^N k_i\right)\left(\textstyle\sum_{j=1}^N k_j\right) = K + \textstyle\sum_{i \neq j} k_i k_j \\ n^2 &= \left(\textstyle\sum_{i=1}^N C_i\right)\left(\textstyle\sum_{j=1}^N C_j\right) = n + \textstyle\sum_{i \neq j} C_i C_j \\ X^2 &= \left(\textstyle\sum_{i=1}^N k_i C_i\right)\left(\textstyle\sum_{j=1}^N k_j C_j\right) = X + \textstyle\sum_{i \neq j} k_i k_j C_i C_j \end{aligned}$$

Define $\lambda = \mathbb{E}\left[C_i C_j\right]$ for $i \neq j$ (by symmetry it is the same for all $i \neq j$). Use this to prove

$$\begin{aligned}\mathbb{E}\left[X^2\right] &= np + \frac{np(n-1)(pN-1)}{(N-1)} \\ Var(X) &= \frac{p(1-p)n(N-n)}{N-1} \\ Var(X/n) &= \frac{p(1-p)(N-n)}{n(N-1)}\end{aligned}$$

d) For our case $N = 1322, n = 1103$, compute $Var(X/n)$ and prove that

$$Var(X/n) \leq \frac{N-n}{4n(N-1)} = \frac{219}{5828252} \approx 0.000037575588701380791359$$

e) Use the Chebyshev inequality to find a value $c > 0$ such that

$$P\left[\left|\tfrac{X}{n} - p\right| \geq c\right] \leq 0.05$$

The value c defines the *confidence interval* $\left[\frac{X}{n} - c, \frac{X}{n} + c\right]$, so that we have 95% confidence that the value of p lies within this interval. Specifically, here p is nonrandom but unknown, the interval endpoints are random, and $P\left[p \in \left[\frac{X}{n} - c, \frac{X}{n} + c\right]\right] = P[\frac{X}{n} \in [p-c, p+c]] \geq 0.95$.

f) In [1] the value $p = \frac{K}{N}$ represents an empirical value for $P[Have|Positive]$. The paper found $X = 1103 - 462 = 641$. For this random realization of X, give the interval $[A, B]$ such that $P[p \in [A, B]] \geq 0.95$.

g) Chapter 6, equation (6.41) in [2] provides the *fourth centralized moment* of a hypergeometric random variable X with parameters N, K, n with $p = K/N$ and $N \geq 4$: Defining $\mu_4 = \mathbb{E}\left[(X - np)^4\right]$ yields

$$\mu_4 = \frac{p(1-p)n(N-n)}{(N-1)(N-2)(N-3)}\left[N^2 + N - 6n(N-n) + 3p(1-p)n(N-n)(N+6) - 6p(1-p)N^2\right]$$

Prove that for any $c > 0$ we have:

$$P[|X/n - p| \geq c] \leq \frac{\mu_4}{n^4 c^4}$$

Use this to get an *improved* 95% confidence interval for (e) and (f).

h) Search the web to find other bounds that can be used to get confidence intervals for this problem (such as exponential tail bounds for the hypergeometric random variable). Compare them to the bounds above. Can we find any better bounds than the ones given above? What kind of bounds are they? Are they solid bounds or just approximations?

II. Simulation

Simulate M independent experiments (where M is a suitably big number to see interesting results, say, $M = 20000$): Each experiment has $N = 1322$ people, the first $K = 768$ have the disease and the remaining people do not. Thus, in this simulation we have $p = 768/1322 = 0.58093797$. We randomly choose $n = 1103$ people (uniformly over all choices). Let X_m be the number of chosen people who have the disease on experiment $m \in \{1, ..., M\}$. You may want to use a random sampling command (without replacement). In MATLAB it is randsample(1322,1103).

a) Plot $\frac{1}{j}\sum_{m=1}^{j}\frac{X_m}{1103}$ versus $j \in \{1, ..., M\}$ and compare with the value p. What does $\frac{1}{j}\sum_{m=1}^{j}\frac{X_m}{1103}$ have to do with the value p and why do we want to consider large j? What is the mean and variance of $\frac{1}{j}\sum_{m=1}^{j}\frac{X_m}{1103}$?

b) Plot $\frac{1}{j}\sum_{m=1}^{j}\frac{X_m^2}{1103^2}$ versus $j \in \{1, ..., M\}$ and compare with $\mathbb{E}\left[\frac{X^2}{1103^2}\right]$ from problem I(c).

c) Plot $\frac{1}{j}\sum_{m=1}^{j}1_{\{|\frac{X_m}{1103} - p\} \geq c\}}$ versus $j \in \{1, ..., M\}$ for your c value of problem I(g) and compare to the upper bound on $P[|\frac{X}{1103} - p| \geq c]$ of 0.05. What is the mean and variance of $\frac{1}{j}\sum_{m=1}^{j}1_{\{|\frac{X_m}{1103} - p\} \geq c\}}$?

References

[1] Joshua S. Gans, Avi Goldfarb, Ajay K. Agrawal, Sonia Sennik, Janice Stein, and Laura Rosella. False-Positive Results in Rapid Antigen Tests for SARS-CoV-2. *JAMA*, 327(5):485–486, 02 2022.

[2] N. L. Johnson, A. W. Kemp, and S. Kotz. *Univariate Discrete Distributions, 3rd ed.* John Wiley & Sons, Inc., 2005.