Statistical Inference Chapter 3

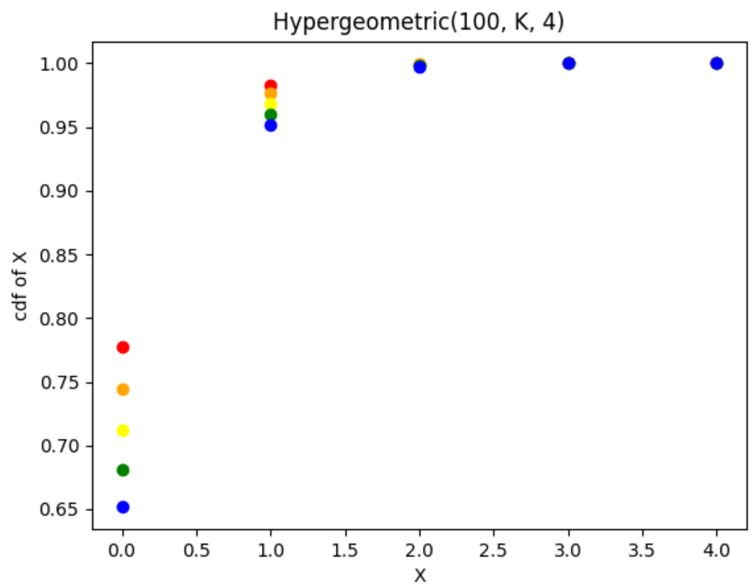
Gallant Tsao

July 21, 2024

1. We first note that the pmf of X is
$$p_X(x) = \frac{1}{N_1 - N_0 + 1},\ x \in \{N_0, N_0+1, ..., N_1\}.$$
Then we get the expectation to be
$$\begin{aligned}\mathbb{E}[X] &= \sum_{x=N_0}^{N_1} x\frac{1}{N_1 - N_0 + 1} \\ &= \frac{1}{N_1 - N_0 + 1} \cdot \frac{N_1 - N_0 + 1}{2}(2N_0 + (N_1 - N_0 + 1 - 1)) \\ &= \frac{N_1 + N_0}{2}.\end{aligned}$$
As for the variance, we get
$$\begin{aligned}\mathbb{E}[X^2] &= \sum_{x=N_0}^{N_1} x^2\frac{1}{N_1 - N_0 + 1} \\ &= \frac{1}{N_1 - N_0 + 1}\Big(\sum_{x=1}^{N_1} x^2 - \sum_{x=1}^{N_0-1} x^2\Big) \\ &= \frac{1}{N_1 - N_0 + 1}\Big(\frac{N_1(N_1+1)(N_1+2) - (N_0-1)(N_0)(2N_0-1)}{6}\Big)\end{aligned}$$
So that
$$\begin{aligned}Var(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\ &= 1\end{aligned}$$

2. Let X = number of defective parts in the sample. Then $X \sim \text{Hypergeometric}(100, n, K)$.
   (a) Firstly, we need $n = 6$ because for the same K, increasing n decreases the value of the Hypergeometric pmf (image shown at end of answer). Then with $n = 6$,
$$\begin{aligned}P(X=0) &= \frac{\binom{6}{0}\binom{94}{K}}{\binom{100}{K}} \\ &= \frac{(100-k)\cdots(100-K-5)}{100\cdots 95}.\end{aligned}$$

After some trial and error with the calculations, we have that when $K = 31$, $P(X = 0) = 0.10056$, but when $K = 32$, $P(X = 0) = 0.09182$. Therefore, the sample size must be at least 32.



(b) By the same reasoning above, we need $n = 6$. Then with this n,

$$P(X = 0 \text{ or } 1) = \frac{\binom{6}{0}\binom{94}{K}}{\binom{100}{K}} + \frac{\binom{6}{1}\binom{94}{K-1}}{\binom{100}{K}}.$$

Again, by trial and error, when $K = 50$, $P(X = 0 \text{ or } 1) = 0.10220$, but when $K = 51$, $P(X = 0 \text{ or } 1) = 0.09331$ hence the sample size must be at least 51.

3. During the three seconds that the person is crossing, there should be no cars passing. The probability of this happening is $(1 - p)^3$. The only possibility for the person to not wait exactly 4 seconds is when there is a car at the first second and no cars in the next 3 seconds. The probability of this happening is $p(1 - p)^3$. Since the times are independent, the probability that the pedestrian has to wait exactly 4 seconds is $[1 - p(1 - p)^3](1 - p)^3$.

4. (a) Let X be the number of trials. Then in this case $X \sim \text{Geom}(0.1)$. Therefore the mean number of trials is just $\frac{1}{0.1} = 10$.

   (b)

5. Let X = number of effective cases. Suppose the new drug is equally effective as the old drug. Then $X \sim \text{Binomial}(100, 0.8)$ if the cases are independent from each other, which is a reasonable assumption. We have

$$P(X \geq 85) = \sum_{k=85}^{100} \binom{100}{k} 0.8^k \cdot 0.2^{100-k} = 0.1285.$$

From this, the probability of getting 85 or more effective cases is not too small, hence we cannot directly make a conclusion that the new drug is superior.

6. (a) $X \sim \text{Binomial}(2000, 0.01)$.

   (b)

$$\sum_{k=0}^{99} \binom{2000}{k} 0.01^k \cdot 0.99^{2000-k}.$$

   (c) In our problem, $n = 2000, p = 0.01, q = 0.99$. Since $np, nq > 5$, we can use normal approximation here. The normal approximation is $Y \sim N(\mu, \sigma^2)$, where

$$\mu = np = 20, \sigma^2 = npq = 19.8.$$

   Then we get

$$P(X < 100) \approx P(Z < 17.979) = 1.$$

7. Let X be the number of chocolate chips in the cookie. Then $X \sim \text{Poisson}(\lambda)$. We want that

$$P(X \geq 2) = 1 - P(X \leq 1) > 0.99 \implies P(X \leq 1) = e^{-\lambda} + \lambda e^{-\lambda} < 0.01.$$

   Solving the above numerically, we get that $\lambda = 6.6384$.

8. (a) Let X be the number of customers in the theater. Then $X \sim \text{Binomial}(1000, \frac{1}{2})$. We want

$$P(X > N) = \sum_{k=N+1}^{1000} \binom{1000}{k} \left(\frac{1}{2}\right)^k \left(1 - \frac{1}{2}\right)^{1000-k} < 0.01.$$

   In other words, we are solving the smallest N such that

$$\left(\frac{1}{2}\right)^{1000} \sum_{k=N+1}^{1000} \binom{1000}{k} < 0.01.$$

   By looping over N, we eventually get that $N = 537$.

   (b) $n = 1000, p = q = \frac{1}{2}$. Therefore the parameters for the normal approximation are $\mu = np = 500, \sigma^2 = npq = 250$. Then we are solving for

$$P(X > N) \approx P(Z > \frac{N - 500}{\sqrt{250}}) < 0.01.$$

   Using R, we get that

$$\frac{N - 500}{\sqrt{250}} = 2.326 \implies N \approx 537,$$

   which is the same as our answer in part (a).

9. (a) Let $X \sim$ Binomial as depicted in the question.

$$\begin{aligned} P(X \geq 5) &= 1 - P(X \leq 4) \\ &= 1 - \sum_{k=0}^{4} \binom{60}{k} \left(\frac{1}{90}\right)^k \left(1 - \frac{1}{90}\right)^{60-k} \\ &\approx 0.0006, \end{aligned}$$

which I think is rare enough to be on the news.

(b) Let X be the number of schools in New York state with 5 or more sets of twins. Then $X \sim \text{Binomial}(360, 0.0006)$. We have that

$$P(X \geq 1) = 1 - P(X = 0) \approx 0.1698.$$

(c) Let X be the number of states in the past 10 years having 5 or more sets of twins. Then $X \sim \text{Binomial}(500, 0.1698)$. We have that

$$P(X \geq 1) = 1 - P(X = 0) = 1.$$

Therefore this event becomes almost certain as we broaden the time scope.

10. (a) Let X be the number of packets of cocaine from the first draw, and let Y be the number of noncocaine packets from the second draw. Then we have that $X \sim \text{Hypergeometric}(N + M, N, 4)$ and $Y \sim \text{Hypergeometric}(N + M - 4, M, 2)$. Then the probability that the defendant is innocent is

$$P(X = 4)P(Y = 2) = \frac{\binom{N}{4}\binom{M}{0}}{\binom{N+M}{4}} \frac{\binom{M}{2}\binom{N-4}{0}}{\binom{N+M-4}{2}} = \frac{\binom{N}{4}\binom{M}{2}}{\binom{N+M}{4}\binom{N+M-4}{2}}.$$

(b) Since the denominator from part (a) is a constant, we just have to find the maximizer of the numerator, which is just $\binom{N}{4}\binom{496-N}{2}$. After some calculus, the local maximizer is about 330.834, hence the maximum is attained at $N = 331, M = 165$, with value about 0.022.

11. (a)

12. Consider a sequence of independent Bernoulli(p) random variables. We define $X =$ Number of successes in n trials, and $Y =$ Number of failures until the rth success. Then X, Y have the specified distributions in the questions. Then

$$\begin{aligned} F_X(r-1) &= P(X \leq r - 1) \\ &= P(r\text{th success on } (n+1)\text{th or later trial}) \\ &= P(\text{At least } (n + 1 - r) \text{ failures before the } r \text{ th success}) \\ &= P(Y \geq n - r + 1) \\ &= 1 - P(Y \leq n - r) \\ &= 1 - F_Y(n - r). \end{aligned}$$

13. Firstly, note that we can find the expectation and variance of the truncated distribution for a general discrete random variable ranging from 0, then we can plug in the values:

$$\begin{aligned}
\mathbb{E}[X_T] &= \sum_{k=1}^{\infty} kP(X_T = k) \\
&= \sum_{k=1}^{\infty} k\frac{P(X = k)}{P(X > 0)} \\
&= \frac{1}{P(X > 0)} \sum_{k=1}^{\infty} kP(X = k) \\
&= \frac{\mathbb{E}[X]}{P(X > 0)}.
\end{aligned}$$

(a)