1 The problem

In a project, I faced a problem of sampling a certain number of data points for human to tag. We want to find a particular property of the data, that is very rare, if even exists.

There two questions here essentially need answers:

1. If this property does not exist, how many we should sample to be confident it does not exists within our total population.
2. If the answer to the previous question becomes too large for human review. Then with a given number of samples (that is small enough for human review), and our observation that there are $\hat{x}$ positives (and $\hat{x}$ can be zero) within the samples, how confident are we with the population within the total population.

This describes a familiy of problems. For example, imagine the dataset you have is the millions of trades from a bank, you want to find potnetial rogue trading behaviour, which you are not sure if it is exists.

The general setup of this problem can be described as following:

1. We have n total data points (n is large), and then we need to sample k from these points for expensive observation (e.g. manual inspection).
2. There are x data points, within n total population we have, are positive (i.e. having the property that we have).
3. Within k samples, we observe $\hat{x}$ positive data points.
4. Given observation we have $\hat{x} \in \{0, 1, \cdots, k\}$ (note we included 0 to indicate thre can be no such property), what is the probability distribution of x?

Note: This problem is slightly different from the standard problem that we were facing when talking about confidence interval. you have a random variable follow certain distribution (say, the closest is Bernoulli distribution) and it tells you how many sample you should draw from it (thus there's no concept of total population). On the contrary, in our problem we care about drawing sample from a set total population.

2 The thought

As I am very Bayesian person, so I decided to approach this problem in a Bayesian way, the above question to me is equivalent to solve the following:

$$\mathbb{P}(X_n = x | \hat{X}_k = \hat{x}) = \frac{\mathbb{P}(\hat{X}_k = \hat{x} | X_n = x)\mathbb{P}(X_n = x)}{\mathbb{P}(\hat{X}_k = \hat{x})} \tag{1}$$

where

- X_n is the random variable where given n total population, you observed X positives.
- $\hat{X}_k$ is the random variable where given k sample, you observed $\hat{X}$ positives.
- $\mathbb{P}(X_n = x|\hat{X}_k = \hat{x})$ is the probability distribution of observing X number of positives within total n population, given we have observed $\hat{x}$ within our sample.

Naturally, we start with uninformative prior $\mathbb{P}(X_n) = \frac{1}{n+1}$ (notice the $n+1$ here is to include the possibility of zero positives), the normalising constant is

$$\mathbb{P}(\hat{X}_k = \hat{x}) = \sum_{i=0}^{n} \mathbb{P}(\hat{X}_k = \hat{x}|X_n = i) \tag{2}$$

The interesting bit lies in the calculation of the posterior distribution. Consider an instance here:

$$\mathbb{P}(\hat{X}_2 = 0|X_3 = 2) \tag{3}$$

with a total population $n = 3$, and we have 1 positive, what is the probability of seeing 0 positive from a random sample of 2?

Let's go back to basics, assume each datapoint is A, B and C, and when sample 2, we have a combination of AB, AC, BC, and without loss of generality, assume that one postive data point is C. Then it means we have 1/3 of chance to observe 0 positive if we sample 2.

What about 4? We have AB, AC, AD, BC, BD, CD, we have 1/2. Now we generalise this into a formula:

$$\mathbb{P}(\hat{X}_k = 0|X_n = x) = \frac{C_k^{n-x}}{C_k^n} \tag{4}$$

$$= \frac{(n-x)!}{k!(n-x-k)!}\frac{k!(n-k)!}{n!} \tag{5}$$

$$= \frac{\Pi_{i=0}^{x-1}(n-k-i)}{\Pi_{i=0}^{x-1}(n-i)} \tag{6}$$

but this is not general enough.