STAT 2857A – Lecture 13

Chapter 3 Summary Exercise

Introduction

During the COVID-19 pandemic, we all became familiar with reporting of how many people in Canada were infected at any specific point in time. How was this determined? The obvious way to do this is to sample a number of people from the population, test them for COVID, and then multiply the proportion of people in the sample by the total population size.

To put this in the context of what we have been studying, we have a population of size N of which M individuals are infected and $N - M$ are not. We know N (the total population of Canada), but we don't know M. To estimate M, we test n individuals from the population and find that x of them are infected. We then estimate M by[1]

$$\hat{M} = \frac{x}{n}N = \frac{xN}{n}.$$

This makes sense intuitively, but today we are going to look at where this estimate comes from statistically.

Sampling without Replacement

One of the first decisions we would be faced with is whether we should sample individuals with or without replacement. Once again, there is an intuitive answer. Sampling with replacement would risk testing the same individual twice (or possibly even more). It seems clear that this would provide less information than sampling without replacement. Today we will actually show that this is the case.

If we sample without replacement, then we are in a very familiar situation. We are sampling a fixed number of individuals without replacement from a population that has two types of

[1] The hat notation is commonly used to denote an estimate of a quantity. E.g., an estimate of the mean of X would be denoted by $\hat{\mu}_X$ and an estimate of the variance of X by $\widehat{\sigma^2_X}$.

individuals (infected and not). This is the setup for a hypergeometric distribution. Specifically, if we let X denote the number of infected individuals in our population then

$$X \sim \text{Hypergeometric}(n, M, N).$$

One way to estimate M is to set the observed value of X to the expected value and then to solve for M. Using our usual notation, we let the observed value of X be x. This gives us that

$$E(X) = \frac{nM}{N} = x$$

or that

$$\hat{M} = \frac{Nx}{n}.$$

This is exactly the same as the intuitive estimator!

Example

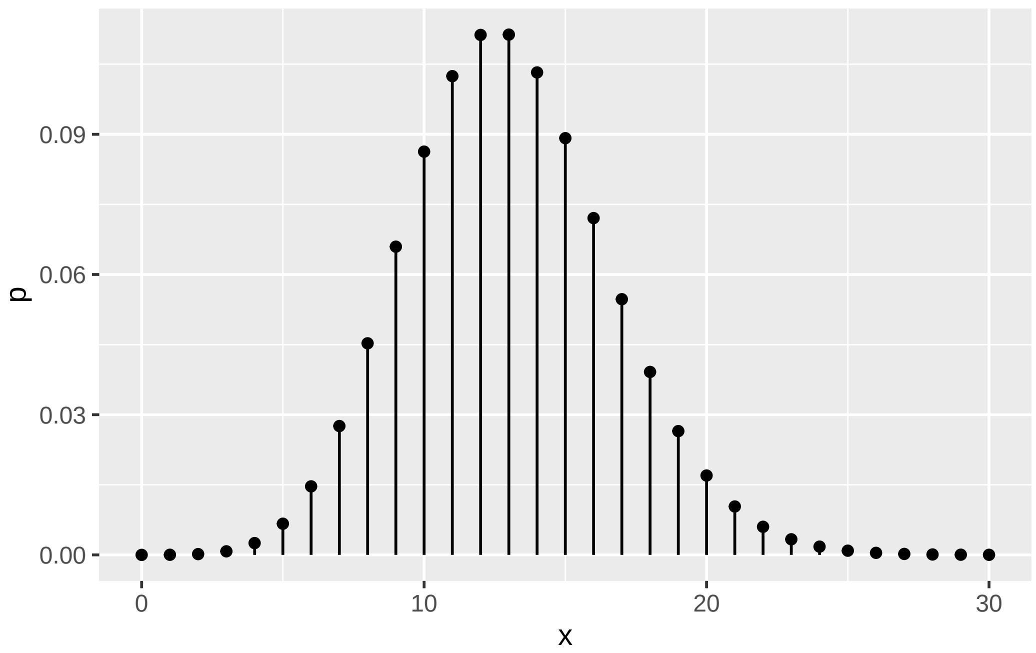
Let's put some real numbers to this. The population of London in 2021 was approximately $N = 422324$. Suppose $n = 10000$ randomly selected people were tested on January and $x = 13$ tested positive. The estimate of the total number of infected people would have been

$$\hat{M} = \frac{Nx}{n} = \frac{422324 \cdot 13}{10000} = 549.02.$$

Of course, we there can't have been 0.02 of an infected person, so this would be rounded to $\hat{M} = 549$.

Precision

One important thing to consider whenever we compute an estimate is that our estimate will rarely be exactly equal to the true value because of random error. Suppose that there were exactly 549 infected people in the population on January 1, 2021. Then the number of people infected follow a hypergeometric distribution with parameters $n = 10000$, $M = 549$, and $N = 422324$. The probability mass function looks like this



Although $x = 13$ is the most likely value, the probability of getting exactly this value is fairly small: $P(X = 13) = 0.1113357$. In fact, it's almost equally likely that $x = 12$ ($P(X = 12) = 0.1112698$) in which case we would estimate the number of infected people to be 507. The probability that the number of positive test is either 8 or $X = 17$ is also about half as large ($P(X = 8) = .0460$ and $P(X = 17) = .0547$) meaning that these values are reasonably likely to occur. The corresponding estimates of M would be $29,594$ and $62,886$ respectively.

On the flip side, there are many possible value of M for which it is not unreasonable for a sample of 10000 tests to result in 13 positive cases. The probability that $X = 13$ is 0.1051886 when $M = 500$ and 0.1055476 when $M = 600$. Relatively speaking, the value $X = 13$ is almost as likely to occur in these cases as when $M = 549$.

Precision refers to the amount of uncertainty or variance in an estimate. In fact, the two are opposites. An estimate is more precise if it is less variable. One way to quantify the amount of uncertainty is to consider all possible values of the unknown parameter, M in this case, for which the probability of the observed data is above a certain threshold. For example, we might consider what values of M would assign probabilities greater than 5% to the value 13, $P(X = 13) > .05$. It turn out that this is true for all of the values between 380 and 761. This implies that any value of M between 380 and 761 is plausible, even thought $\hat{M} = 549$ is our best guess.

Sampling with Replacement

Suppose now that testing was conducted with replacement. Let Y be the number of people who test positive. Then

$$Y \sim \text{Binomial}(n, p)$$

where n is the sample size and p is the probability that a random sampled individual tests positive. If a total of M people are infected then $p = M/N$. Once again, we can estimate the total number of infected people by equating the expected and observed values of Y and solving for M. This yields

$$E(Y) = np = \frac{nM}{N} = y$$

so that

$$\hat{M} = \frac{Nx}{n}$$

exactly as above. Whether we sample with or without replacement, the estimate of M is the same.

Example

Suppose again that $n = 10000$ are tested with replacement and $y = 13$. The estimate of the total number of infected people would still

$$\hat{M} = \frac{Nx}{n} = \frac{422324 \cdot 13}{10000} = 549.02.$$

which gives $\hat{M} = 549$ rounded to the nearest whole person.

Precision

Once again, we can gauge the precision of the estimator by the range of values of M for which $P(Y = 13)$ is greater than some threshold. In this case, $P(Y = 13) > .05$ for all values of M between 380 and 762. If we compare the number of plausible values, then there are $761 - 380 = 381$ values of M for which $P(X = 13) > .05$ when sampling without replacement and $762 - 380 = 382$ when sampling with replacement.

Note that the estimate is very slightly more precise (the range of plausible values for M is slightly narrower) when sampling without replacement than sampling with replacement. This is the reason that sampling without replacement is better and reflects the fact that we get less information when sampling with replacement because there is a chance that we sample the same individuals multiple times.

Another way to compare the precision of the estimates is to compare their variances. The estimator computed when sampling with replacement,

$$\hat{M} = \frac{NX}{n},$$

is a linear transformation of the random variable X. Then

$$\text{Var}(\hat{M}) = \frac{N^2}{n}\text{Var}(X)$$

and since X is hypergeometric

$$\text{Var}(\hat{M}) = \frac{N^2}{n^2}\left(\frac{N-n}{N-1}\right)\left(\frac{nM}{N}\right)\left(1-\frac{M}{N}\right).$$

Similarly, if we sample without replacement then

$$\hat{M} = \frac{NY}{n},$$

is a linear transformation of the random variable Y. Then

$$\text{Var}(\hat{M}) = \frac{N^2}{n}\text{Var}(Y)$$

and since Y is binomial

$$\text{Var}(\hat{M}) = \frac{N^2}{n^2}n\left(\frac{M}{N}\right)\left(1-\frac{M}{N}\right).$$

The ratio of these two is

$$\frac{\frac{N^2}{n^2}\left(\frac{N-n}{N-1}\right)\left(\frac{nM}{N}\right)\left(1-\frac{M}{N}\right)}{\frac{N^2}{n^2}n\left(\frac{M}{N}\right)\left(1-\frac{M}{N}\right)} = \frac{N-n}{N-1}$$

which shows that the variance is lower by a factor of $(N-n)/(N-1)$ when sampling is conducted with replacement. Note that the relative variance does not depend on M. It only depends on N and n. It's always above 1, but gets closer and closer to 1 as the size of the population, N, gets bigger while the size of the sample, n, remains fixed.

In our example, the ratio of the variances would be

$$\frac{422324-10000}{422324-1} = 0.9763238.$$

Once again, this suggests that the estimate will be slightly more precise when sampling without replacement.

Sampling until a Fixed Number of Successes

For completeness, we will also consider the case in which individuals are tested until a specific number of diseased individuals are found. We will assume that individuals are sampled with replacement, though we've just showed that this is not sensible, because it produces a distribution we have studied.

Suppose that testing is conducted until r individuals test positive. Let Z be the number of individuals that test negative before the r-th positive test, and assume that individuals are sampled with replacement. Then,

$$Z \sim \text{Negative Binomial}(r, p)$$

where $p = M/N$ is the proportion of infected individuals in the population. Then

$$E(Z) = \frac{r(1-p)}{p} = r(N/M - 1).$$

Solving for M yields

$$M = \frac{N}{(E(Z)/r + 1)}$$

which provides the estimate

$$\hat{M} = \frac{N}{(z/r + 1)}$$

after replacing the expected value of Z, $E(Z)$, by the observed value, z.

Suppose, for example, that we tested until 13 individuals tested positive and that this was reached on the 10,000-th test. Then $r = 13$ and $z = 10000 - 13 = 9987$ so that

$$\hat{M} = \frac{422324}{(9987/13 - 1)} = 549.0212.$$

Note that this is exactly the same estimate that we had before. In fact, it doesn't really matter how the sampling is conducted. Provided that the population is homogeneous (every individual has the same probability of being infected) then all that matters is the size of the sample and how many people were infected.

In this case, however, the estimate is not a linear combination of the random variable Z. This makes it much more difficult to compute the variance of the new estimate, and so we will leave it there.