# Lecture 13 Notes

Simon Bonner

## Setup

- We are going to consider the problem of estimating the size of a subpopulation.
- Suppose, for example, that we have a population of known size, and we want to estimate the number of individuals infected with some disease.
- To make this concrete, I have a bag containing 140 mints of which an unknown number are “diseased” (mint chocolate).
- Out goal is to estimate the number of “diseased’ ’ candies in the entire population.
- How can we do this?

## Activity

### Sampling without Replacement

#### 1. Estimation

- Students sample 20 candies without replacement. Let $x$ be the number of diseased candies in the sample.
- Ask them to estimate the number of diseased candies in the entire population. The obvious value is
$$\hat{M} = \frac{140x}{20}.$$
- Ask them why?
- Provide mathematics:
  - Given that we are sampling without replacement, $X \sim \text{Hypergeometric}(20, M, 140)$. Then
  $$E(X) = \frac{20M}{140}.$$
  - If we rearrange and solve for $M$ this provides
  $$M = \frac{140E(X)}{20}.$$

– Replacing $E(X)$ with the observed value yields

$$\hat{M} = \frac{140x}{20}.$$

- Compute estimate for the specific sample obtained by the class.
  – Round to nearest integer.

## 2. Precision

- Before continuing, enter value of $x$ in `Slide/lecture_13_slides.Rnw` and recompile.
- Ask: "Who believe that there are exactly $\hat{M}$ diseased candies in the bag?"
- Even though $\hat{M}$ may be our best guess, and it may be right, there is uncertainty in this experiment. If we repeated the experiment, then we would likely draw a different number of diseased candies in our sample. This would lead to different estimates.
- More importantly, there are different values of $M$ that could lead to the same value of $x$.
  – Use calculator at stattrek to compute $P(X = x|M = \hat{M})$ and $P(X = x|M = \hat{M}+1)$.
  – The probabilities should be similar. This implies that the true value of $M$ could easily be $\hat{M}$ or $\hat{M} + 1$.
  – In fact, there is a range of values for which $P(X = x|M)$ is close to $P(X = x|M = \hat{M})$.
- How big this range is tells us about the precision of the estimate. The precision is the opposite of variance. If the precision is big then the range of possible values is small and we can be confident that the true value is close to $\hat{M}$. On the other hand, if the range is big then the precision is small and we will have less confidence in our estimate.
- Show plot of $P(X = x|M)$ vs $M$.
  – One way to measure the precision is to consider the values of $M$ for which $P(X = x|M)$ is above a certain value.
  – We'll consider the values for which $P(X = x|M) > .05$. (Show next plot with line.)
  – For our sample, the values are… see following table for the values given the observed $x$.

## Sampling with Replacement

## 1. Estimation

- Why did we sample without replacement?
  – The answer seems intuitive. If we sample with replacement then we risk sampling the same indivdiuals multiple time, which would give us less information about the population. How can we see this?

- Suppose that we were to sample with replacement. Let $Y$ be the number of diseased candies sampled.
  - If we sample with replacement, $Y \sim \text{Binomial}(20, M/140)$. Then

$$E(Y) = \frac{20M}{140}.$$

  - If we rearrange and solve for $M$ this provides

$$M = \frac{140E(X)}{20}.$$

  - Replacing $E(Y)$ with the observed value yields exactly the same estimator

$$\hat{M} = \frac{140y}{20}.$$

- This means that if we observe the same number of diseased candies then our estimate will be the same regardless of whether or not we sample with replacement.
  - So, where does sampling with replacement matter?

## 2. Precision

- Sampling with replacement affects the precision of the estimate.
- Suppose that we consider the values of $M$ for which $P(Y = y|M) > .05$. (Show plot)
- In this case, the range of plausible values is... see Table 2 for values.
- The relative precision is...see Table 2 for values.
  - The range of values is slightly wider when we sample without replacement than when we sample with replacement.
  - This is why sampling without replacement is better.

## Variances

- Computing the range of values such that $P(X = x|M) > .05$ or $P(Y = y|M) > .05$ is not trivial. However, there is a better way to measure the precision.
- The precision measures the uncertainty of the estimate, and uncertainty is measured by the variance. In fact, precision in statistics is defined to be the inverse of the variance.
- Consider the case of the hypergeometric:

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

– 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 $N$ gets bigger.

- In our example, the ratio of the variances would be

$$\frac{140-20}{140-1} = 0.8633094.$$

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

### Sampling without Replacement

The following table presents the values required for each possible value of $X$ – the number of "diseased" candies sampled. The columns are:

- $x$ – the possible values (from 0 to $n$)

| x | P | Estimate | Variance | Phat | Lower | Upper | Precision |
|---|---|---|---|---|---|---|---|
| 0 | 0.0006480 | 0 | 0.0000000 | 1.0000000 | 0 | 18 | 18 |
| 1 | 0.0063998 | 7 | 0.8201439 | 0.4067804 | 1 | 27 | 26 |
| 2 | 0.0289161 | 14 | 1.5539568 | 0.3077208 | 3 | 36 | 33 |
| 3 | 0.0794321 | 21 | 2.2014388 | 0.2621072 | 7 | 43 | 36 |
| 4 | 0.1486987 | 28 | 2.7625899 | 0.2355596 | 12 | 51 | 39 |
| 5 | 0.2015305 | 35 | 3.2374101 | 0.2184487 | 17 | 58 | 41 |
| 6 | 0.2050456 | 42 | 3.6258993 | 0.2069165 | 23 | 66 | 43 |
| 7 | 0.1602655 | 49 | 3.9280576 | 0.1991090 | 29 | 72 | 43 |
| 8 | 0.0976618 | 56 | 4.1438849 | 0.1940435 | 35 | 79 | 44 |
| 9 | 0.0468191 | 63 | 4.2733813 | 0.1911835 | 41 | 86 | 45 |
| 10 | 0.0177393 | 70 | 4.3165468 | 0.1902579 | 47 | 93 | 46 |
| 11 | 0.0053165 | 77 | 4.2733813 | 0.1911835 | 54 | 99 | 45 |
| 12 | 0.0012569 | 84 | 4.1438849 | 0.1940435 | 61 | 105 | 44 |
| 13 | 0.0002329 | 91 | 3.9280576 | 0.1991090 | 68 | 111 | 43 |
| 14 | 0.0000334 | 98 | 3.6258993 | 0.2069165 | 74 | 117 | 43 |
| 15 | 0.0000037 | 105 | 3.2374101 | 0.2184487 | 82 | 123 | 41 |
| 16 | 0.0000003 | 112 | 2.7625899 | 0.2355596 | 89 | 128 | 39 |
| 17 | 0.0000000 | 119 | 2.2014388 | 0.2621072 | 97 | 133 | 36 |
| 18 | 0.0000000 | 126 | 1.5539568 | 0.3077208 | 104 | 137 | 33 |
| 19 | 0.0000000 | 133 | 0.8201439 | 0.4067804 | 113 | 139 | 26 |
| 20 | 0.0000000 | 140 | 0.0000000 | 1.0000000 | 122 | 140 | 18 |

## Sampling with Replacement

| y | P | Estimate | Variance | Phat | Lower | Upper | Precision | Rel_Precision | Rel_Variance |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.0011952 | 0 | 0.00 | 1.0000000 | 0 | 19 | 19 | 1.0555556 | NaN |
| 1 | 0.0095616 | 7 | 0.95 | 0.3773536 | 1 | 29 | 28 | 1.0769231 | 1.158333 |
| 2 | 0.0363340 | 14 | 1.80 | 0.2851798 | 3 | 37 | 34 | 1.0303030 | 1.158333 |
| 3 | 0.0872015 | 21 | 2.55 | 0.2428289 | 7 | 45 | 38 | 1.0555556 | 1.158333 |
| 4 | 0.1482426 | 28 | 3.20 | 0.2181994 | 12 | 52 | 40 | 1.0256410 | 1.158333 |
| 5 | 0.1897505 | 35 | 3.75 | 0.2023312 | 17 | 60 | 43 | 1.0487805 | 1.158333 |
| 6 | 0.1897505 | 42 | 4.20 | 0.1916390 | 22 | 67 | 45 | 1.0465116 | 1.158333 |
| 7 | 0.1518004 | 49 | 4.55 | 0.1844012 | 28 | 74 | 46 | 1.0697674 | 1.158333 |
| 8 | 0.0986703 | 56 | 4.80 | 0.1797058 | 34 | 80 | 46 | 1.0454545 | 1.158333 |
| 9 | 0.0526241 | 63 | 4.95 | 0.1770550 | 40 | 87 | 47 | 1.0444444 | 1.158333 |
| 10 | 0.0231546 | 70 | 5.00 | 0.1761971 | 46 | 94 | 48 | 1.0434783 | 1.158333 |
| 11 | 0.0084199 | 77 | 4.95 | 0.1770550 | 53 | 100 | 47 | 1.0444444 | 1.158333 |
| 12 | 0.0025260 | 84 | 4.80 | 0.1797058 | 60 | 106 | 46 | 1.0454545 | 1.158333 |
| 13 | 0.0006218 | 91 | 4.55 | 0.1844012 | 66 | 112 | 46 | 1.0697674 | 1.158333 |
| 14 | 0.0001244 | 98 | 4.20 | 0.1916390 | 73 | 118 | 45 | 1.0465116 | 1.158333 |
| 15 | 0.0000199 | 105 | 3.75 | 0.2023312 | 80 | 123 | 43 | 1.0487805 | 1.158333 |
| 16 | 0.0000025 | 112 | 3.20 | 0.2181994 | 88 | 125 | 37 | 0.9487179 | 1.158333 |
| 17 | 0.0000002 | 119 | 2.55 | 0.2428289 | 95 | 125 | 30 | 0.8333333 | 1.158333 |
| 18 | 0.0000000 | 126 | 1.80 | 0.2851798 | 103 | 125 | 22 | 0.6666667 | 1.158333 |
| 19 | 0.0000000 | 133 | 0.95 | 0.3773536 | 111 | 125 | 14 | 0.5384615 | 1.158333 |
| 20 | 0.0000000 | 140 | 0.00 | 1.0000000 | 121 | 125 | 4 | 0.2222222 | NaN |