STAT 88: Lecture 23

Contents
Section 7.2: Sampling Without Replacement

Warm up: (Exercise 7.4.5) The number of typos on the cover page of an exam has a distribution given by

value	0	1
probability	0.8	0.2

The number of misprints in the rest of the exam has the Poisson(3) distribution, independently of the cover page. Find the expectation and SD of the total number of misprints on the exam.

Last time

Sums of independent random variables: If X and Y are independent random variables, then

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

If $X \sim \mathrm{Bernoulli}(p)$,

$$\mathrm{SD}(X) = \sqrt{p(1-p)}.$$

If $X \sim \mathrm{Binomial}(n, p)$,

$$\mathrm{SD}(X) = \sqrt{np(1-p)}.$$

If $X \sim \mathrm{Poisson}(\mu)$,

$$\mathrm{SD}(X) = \sqrt{\mu}.$$

If $X \sim \mathrm{Geom}(p)$,

$$\mathrm{SD}(X) = \frac{\sqrt{1-p}}{p}.$$

Example: (Exercise 7.4.10) A non-negative integer valued random variable has expectation 50 and SD 10. Could the random variable have a binomial distribution?

7.2. Sums of Independent Random Variables

The draws in a SRS are not independent of each other and this makes computing the SD of the hypergeometric distribution more complicated than for binomial distribution.

Squares and products of indicators

Let I_A be the indicator of the event A. Then the distribution of I_A is given by

value	0	1
probability	$1 - P(A)$	$P(A)$

Find $E(I_A)$ and $\mathrm{E}(I_A^2)$.

Now let I_B be the indicator for event B. What values does $I_A I_B$ take?

Find $E(I_A I_B)$.

SD of Hypergeometric

Let $X \sim \text{HG}(N, G, n)$. So X measures the number of good elements in a simple random sample of size n drawn from a population of N elements of which G are good. Then we can write

$$X = I_1 + \cdots + I_n,$$

where

$$I_j = \begin{cases} 1 & \text{if } j\text{th draw is good} \\ 0 & \text{else} \end{cases} \qquad p = \frac{G}{N}$$

Recall $E(X) = n\frac{G}{N}$. We want to find $\text{Var}(X) = E(X^2) - (EX)^2$.

We start with the following equation:

$$X^2 = (I_1 + \cdots + I_n)^2 = \sum_{j=1}^{n} I_j^2 + \sum_{j \neq k} I_j I_k.$$

Taking expectaion on both side,

$$\begin{aligned} E(X^2) &= E(\sum_{j=1}^{n} I_j^2) + E(\sum_{j \neq k} I_j I_k) \\ &\overset{\text{Additivity}}{=} \sum_{j=1}^{n} E(I_j^2) + \sum_{j \neq k} E(I_j I_k) \\ &\overset{\text{Symmetry}}{=} nE(I_1^2) + n(n-1)E(I_1 I_2). \end{aligned}$$

We know from our calculation above that $E(I_1^2) =$ and $E(I_1 I_2) =$. So,

$$E(X^2) = n\frac{G}{N} + n(n-1)\frac{G}{N} \cdot \frac{G-1}{N-1}.$$

Since $\text{Var}(X) = E(X^2) - (EX)^2$, it follows that

$$\text{Var}(X) = \underbrace{n\frac{G}{N} + n(n-1)\frac{G}{N} \cdot \frac{G-1}{N-1}}_{E(X^2)} - \underbrace{\left(n\frac{G}{N}\right)^2}_{(EX)^2}.$$

After some boring algebra (see the last page of the note), we can simplify it to

$$\text{Var}(X) = n\frac{G}{N} \cdot \frac{N-G}{N} \cdot \frac{N-n}{N-1}, \text{ and } \text{SD}(X) = \sqrt{n\frac{G}{N} \cdot \frac{N-G}{N} \cdot \frac{N-n}{N-1}}.$$

$= n \cdot p \cdot (1-p) \cdot \left(\frac{N-n}{N-1}\right)$ $= \sqrt{np(1-p)}\sqrt{\frac{N-n}{N-1}}$ (SD of Binomial (n,p))

Example: Draw 5 cards from a deck. Let X be the number of hearts in your hand. Find $E(X)$ and $\text{SD}(X)$.

The Size of the FPC

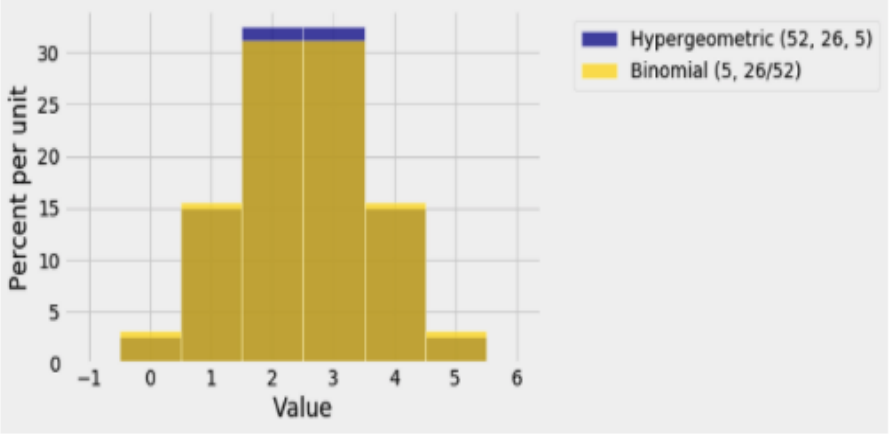
Finite population correction or FPC is given by

$$\text{fpc} = \sqrt{\frac{N-n}{N-1}}.$$

We saw that

$$\text{SD(HG)} = \text{SD(Binomial)} \cdot \text{fpc},$$

so SD(HG) < SD(Binomial).



Sampling with and without replacement are essentially the same when the sample size is small relative to the population size.

Read "The Accuracy of Simple Random Samples" in Ch 7.2 of the textbook.

Example: New Mexico has a population of 1M and California a population of 40M. The two states have the same proportion of Democrats. A random sample of size 0.01% of the population is taken. The SD for the number of democrats in the sample is:

1. roughly the same in both states
2. larger in California
3. larger in New Mexico

Example: Follow up question: suppose in each state a random sample of 500 is taken. The SD of the number of democrats in the poll is:

1. roughly the same in both states
2. larger in California
3. larger in New Mexico

Algebra details:

$$
\begin{aligned}
\mathrm{Var}(X) &= n\frac{G}{N} + n(n-1)\frac{G}{N}\cdot\frac{G-1}{N-1} - \left(n\frac{G}{N}\right)^2 \\
&= n\frac{G}{N}\left(1 + (n-1)\cdot\frac{G-1}{N-1} - n\frac{G}{N}\right) \\
&= n\frac{G}{N}\cdot\frac{N(N-1) + N(n-1)(G-1) - nG(N-1)}{N(N-1)} \\
&= n\frac{G}{N}\cdot\frac{N^2 - N + nNG - nN - NG + N - nNG + nG}{N(N-1)} \\
&= n\frac{G}{N}\cdot\frac{N^2 - nN - NG + nG}{N(N-1)} \\
&= n\frac{G}{N}\cdot\frac{(N-G)(N-n)}{N(N-1)} \\
&= n\frac{G}{N}\cdot\frac{N-G}{N}\cdot\frac{N-n}{N-1}.
\end{aligned}
$$