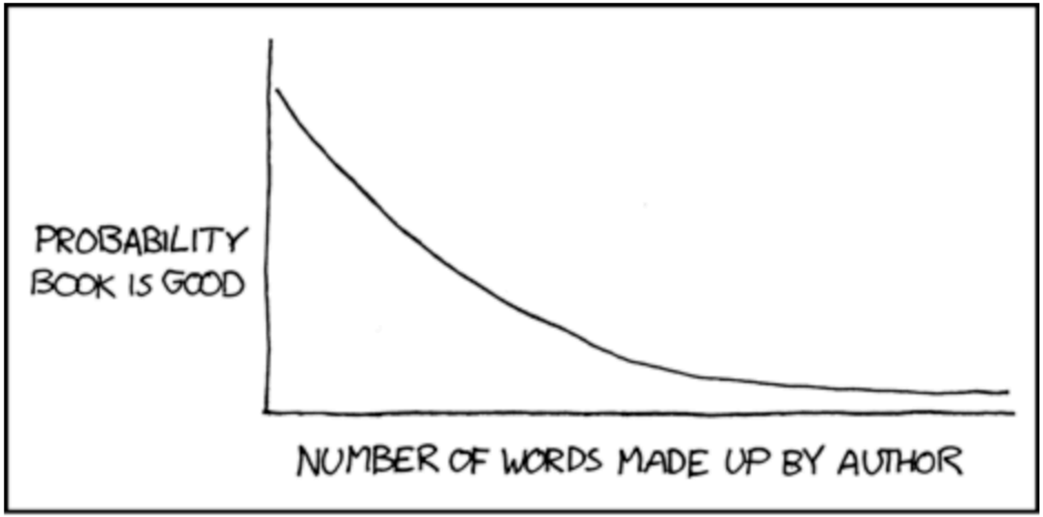
Stat 88: Probability & Mathematical Statistics in Data Science



"THE ELDERS, OR *FRAÁS*, GUARDED THE *FARMLINGS* (CHILDREN) WITH THEIR *KRYTOSES*, WHICH ARE LIKE SWORDS BUT *AWESOMER*..."

xkcd.com/483

Lecture 24: 3/17/2021

Sections 7.2, 7.3

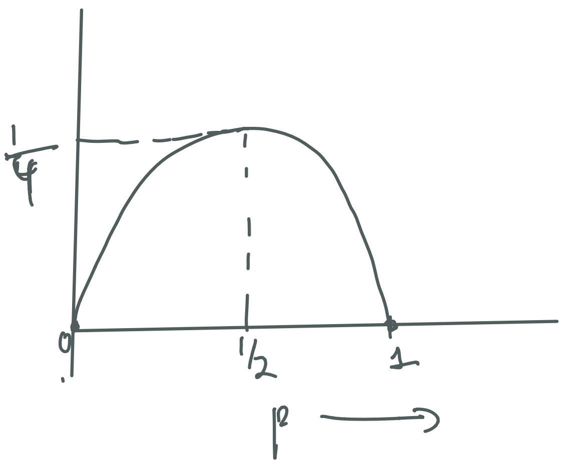
Sampling without replacement and the Law of Averages

Last time:

- $X \sim Bin(n, p),\ Var(X) = np(1-p) = npq,\ SD(X) = \sqrt{npq}$
- $X \sim Pois(\mu),\ E(X) = Var(X) = \mu,\ SD(X) = \sqrt{\mu}$
- $X \sim Geom(p),\ E(X) = \frac{1}{p}, Var(X) = \frac{1-p}{p^2},\ SD(X) = \frac{\sqrt{1-p}}{p}$
- Consider $X \sim Bernoulli(p),\ Var(X) = p(1-p)$. For what p is the variance highest?

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

$p - p^2$



Upper bound on variance of a Bernoulli r.v. is $\frac{1}{4}$

(upper for SD is $\frac{1}{2}$)

Variance of a hypergeometric random variable

sum of draws from a 0-1 population

- Let $X \sim HG(N, G, n)$, then can write $X = I_1 + I_2 + \cdots + I_n$, where I_k is the indicator of the event that the kth draw is good.

- We can compute the expectation of X using symmetry: $E(X) = \frac{nG}{N}$
- But what about variance?
- Since the indicators are not independent, we can't just add the variances
- Let's just use the formula: $Var(X) = E(X^2) - \left(\frac{nG}{N}\right)^2$
- $X^2 = \underbrace{(I_1 + I_2 + \cdots + I_n)^2}_{} = \underbrace{\sum_{k=1}^n I_k^2}_{} + \boxed{\sum_j \sum_k}{}_{k \neq j} I_j I_k$ # of pairs = $n(n-1)$

$$E(X^2) = nE(I_k^2) + n(n-1)E(I_j I_k) = n\frac{G}{N} + n(n-1)P(I_j = 1)P(I_k = 1 \mid I_j = 1)$$

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

$$Var(X) = E(X^2) - (E(X))^2$$

$\sum_{j=1}^{n} \sum_{\substack{k=1 \\ k \neq j}}^{n} I_j \cdot I_k$

$(I_1 + I_2)(I_1 + I_2)$

$(I_1 + I_2 + I_3)^2 = I_1^2 + I_2^2 + I_1 I_2 + I_2 I_1 + I_3^2$
$I_1 I_3 + I_2 I_3 + I_3 I_1 + I_3 I_2$

$I_1(I_2+I_3) + I_2(I_1+I_3) + I_3(I_1+I_2)$

Variance of a hypergeometric random variable

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$$

$$= \frac{nG}{N} + n(n-1)\frac{G}{N}\cdot\frac{G-1}{N-1} - \left(\frac{nG}{N}\right)^2$$

$$= \frac{nG}{N}\left[1 + \frac{(n-1)(G-1)}{N-1} - \frac{nG}{N}\right]$$

$$= \frac{nG}{N}\left[\frac{N(N-1) + N(n-1)(G-1) - nG(N-1)}{N(N-1)}\right]$$

$$= \frac{nG}{N}\left[\frac{N^2 - \cancel{N} + n\cancel{NG} - NG - nN + \cancel{N} - \cancel{nGN} + nG}{N(N-1)}\right]$$

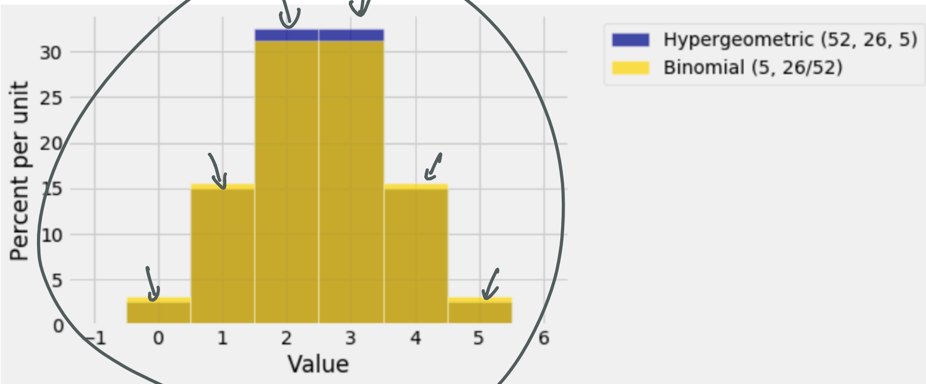
$$= \frac{nG}{N}\left[\frac{N(N-G) - n(N-G)}{N(N-1)}\right] = \frac{nG}{N}\cdot\frac{N-G}{N}\cdot\frac{N-n}{N-1}$$

$$\mathrm{Var}(X) = \underbrace{n}_{\text{Sample size}}\cdot\underbrace{\frac{G}{N}}_{P(S)}\cdot\underbrace{\frac{N-G}{N}}_{P(F)}\cdot\boxed{\frac{N-n}{N-1}}$$

square of Finite popn correction

Binomial (n,p) $\quad n \cdot p \cdot (1-p) \cdot \text{fpc}$

The finite population correction (fpc) & the accuracy of SRS



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

Note that fpc ≤ 1

So SD(HG) ≤ SD(Bin)

In general we have that the :

bigger than SD (w/o repl)

SD of sum of an SRS = SD of sum WITH repl. × fpc

Exercise. Plug in values of N, n in your calculator & see what $\sqrt{\frac{N-n}{N-1}}$ will be. , $N = 10^6$, $n = 1000$

$$\sqrt{\frac{10^6 - 10^3}{10^6 - 1}} = 0.999 \approx 1$$

Accuracy of samples

Simple random samples of the same size of 625 people are taken in Berkeley (population: 121,485) and Los Angeles (population: 4 million). True or false, and explain your choice: The results from the Los Angeles poll will be substantially more accurate than those for Berkeley.

Accuracy is governed by the SD.

$\sqrt{\frac{N-n}{N-1}} \leftarrow$

fpc

$n = 625$

$\rightarrow N_1 = 121485 \Leftarrow$ fp $\sqrt{\frac{121485-625}{121484}}$

0.9974

$N_2 = 4 \times 10^6$

$N_2 \approx 1$

Accuracy depends on Sample size

Example (from *Statistics*, by Freedman, Pisani, and Purves)

A survey organization wants to take an SRS in order to estimate the percentage of people who watched the 2021 Grammys. To keep costs down, they want to take as small a sample as possible, but their client will only tolerate a random error of 1 percentage point or so in the estimate. Should they use a sample size of 100, 2500, or 10000? The population is very large and the fpc is about 1. ← you can pretend that sampling w/ replacement.

Don't know p. so

X = percentage of 1's in sample

Want $SD(X) \leq 0.01$

X = sum of draws

$SD(X) = \frac{\sqrt{npq}}{n}$

Avg of draws = $\frac{\text{sum of draws}}{n}$.

Use the upper bound on variance to solve this