Stat 88 lec 24

Warm up 2:00-2:10

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

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

a) Find $E(I_A) = 0 \cdot (1-P(A)) + 1 \cdot P(A) = P(A)$

b) Find $E(I_A^2)$ $0^2 \cdot (1-P(A)) + 1^2 \cdot P(A) = P(A)$

indicator for an event B

c) What is the distribution table for $I_A I_B$

	0	1
	$1-P(AB)$	$P(AB)$

d) Find $E(I_A I_B) = P(AB)$

$I_A I_B = I_{AB}$

Last time

Sec 7.1 Sums of independent random variables

If X_1, X_2 are independent RVs

$Var(X_1 + X_2) = Var(X_1) + Var(X_2)$

$X \sim$ Bernoulli (p)

$SD(X) = \sqrt{pq}$

$X \sim$ Binomial (n, p)

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

$X \sim$ Poisson (μ)

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

$X \sim$ Geometric (p)

$SD(X) = \frac{\sqrt{q}}{p}$

Today

1. Sec 7.2 Sampling without replacement
2. Sec 7.3 The Law of Averages

(3) Sec 7.2 Sampling without replacement

Whereas a binomial RV is a sum of independent indicators, a hypergeometric RV is a sum of dependent indicators. This makes it more complicated to find the formula for SD (hypergeometric).

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	$1 - P(A)$	$P(A)$

We saw in warmup:

if I_B is the indicator for event B

then $I_A I_B = I_{AB} = \begin{cases} 1 & \text{if } A \text{ and } B \text{ true} \\ 0 & \text{else} \end{cases}$ — $P(AB)$

and $E(I_A I_B) = P(AB)$

SD of Hypergeometric

Let $X \sim HG(N, G, n)$

X = # good in a SRS from pop size N w/ G good, $p = \frac{G}{N}$

$X = I_1 + \cdots + I_n$ where $I_2 = \begin{cases} 1 & \text{if } 2^{nd} \text{ draw good} \\ 0 & \text{else.} \end{cases}$

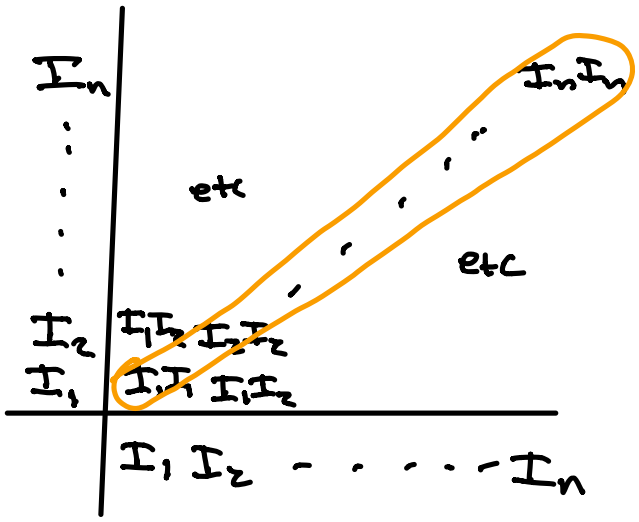
Recall

$E(X) = n\left(\frac{G}{N}\right)$

Find $Var(X)$

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

$X^2 = (I_1 + \cdots + I_n)^2$



$(I_1 + \cdots + I_n)^2$ = sum of diag terms + sum of off diag terms.

$$= \sum_{j=1}^{n} I_j^2 + \sum_{\substack{j=1 \\ j \neq k}}^{n} \sum_{k=1}^{n} I_j I_k$$

so

$$E(X^2) = E\left(\sum_{j=1}^{n} I_j^2\right) + E\left(\sum_{\substack{j=1 \\ j \neq k}}^{n} \sum_{k=1}^{n} I_j I_k\right)$$

$$= \sum_{j=1}^{n} E(I_j^2) + \sum_{\substack{j=1 \\ j \neq k}}^{n} \sum_{k=1}^{n} E(I_j I_k)$$

Symmetry

$$= nE(I_1^2) + n(n-1)E(I_1 I_2)$$

$P(I_1 = 1)$

$P(I_1 = 1 \text{ and } I_2 = 1) = P(I_1 = 1) \cdot P(I_2 = 1 \mid I_1 = 1)$

$$E(I_1^2) = \frac{6}{N}, \quad E(I_1 I_2) = \frac{6}{N} \cdot \frac{6-1}{N-1}$$

so,

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

Since,

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

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

boring algebra (see appendix to lecture)

$$Var(X) = n\frac{G}{N}\cdot\frac{\overbrace{N-G}^{B}}{N}\cdot\frac{N-n}{N-1}$$

$$= npq\cdot\left[\frac{N-n}{N-1}\right]$$

$$\Rightarrow SD(X) = \sqrt{npq}\cdot\sqrt{\frac{N-n}{N-1}}$$

$\sqrt{npq}$: SD of Binomial (n,p)

$\sqrt{\frac{N-n}{N-1}}$: called Finite Population Correction (FPC)

Ex Draw 5 cards from a deck.

X = # of hearts in your hand.

Find $E(X) = 5\frac{13}{52} = 1.25$

$SD(X) = \sqrt{5\frac{13}{52}\frac{39}{52}}\cdot\sqrt{\frac{52-5}{52-1}} = .93$

Size of the FPC

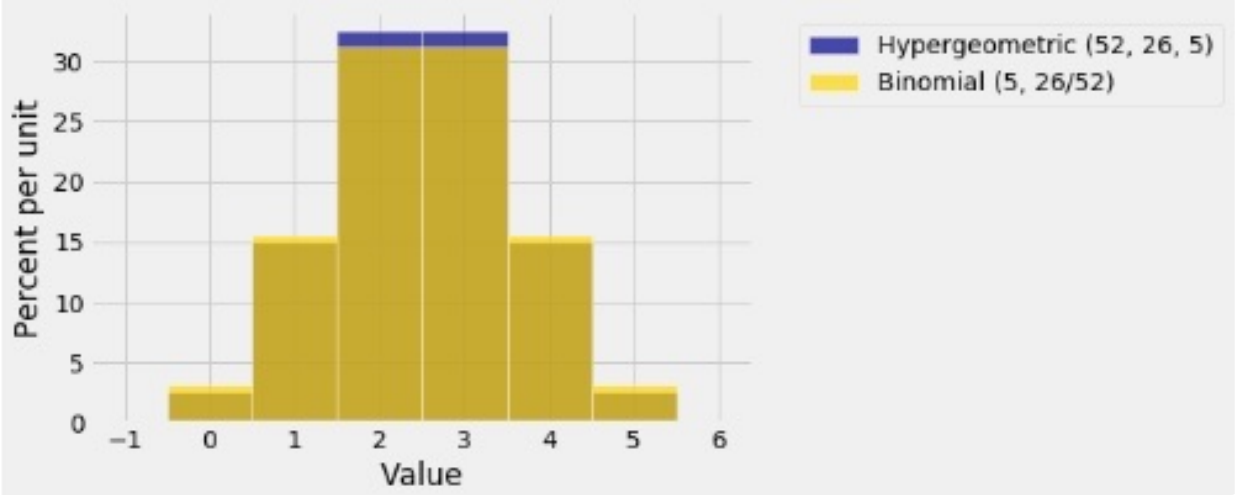
$$FPC = \sqrt{\frac{N-n}{N-1}} \leq 1$$

if $n << N$ (less than 1%) than like drawing with replacement and FPC = 1

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

says $SD(HG) \leq SD(\text{Binomial})$

$\equiv$



Polls are made without replacement because the count of # democrats is more accurate (i.e. smaller SD than polling with replacement).

Tinyurl.com/march16-pt1

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:

so FPC ≈ 1

a roughly the same in both states

b larger in California

c larger in New Mexico

.01% of 40M = .0001 (40 m) = 4000

.01% of 1m = .0001 (1m) = 100

SD (Dem CA) = $\sqrt{4000pq}$ ← larger

SD (Dem NM) = $\sqrt{100pq}$

If $n \ll N$ think of N as ∞

We see from this that SD depends on p, on n but not on N.

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:

a) roughly the same in both states

b) larger in CA

c) larger in NM

(2) The Law of Averages

The sample sum

You toss a fair coin n times

Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(\frac{1}{2})$

$S_n = X_1 + X_2 + \cdots + X_n$ is the sample sum

What would you put your money on:

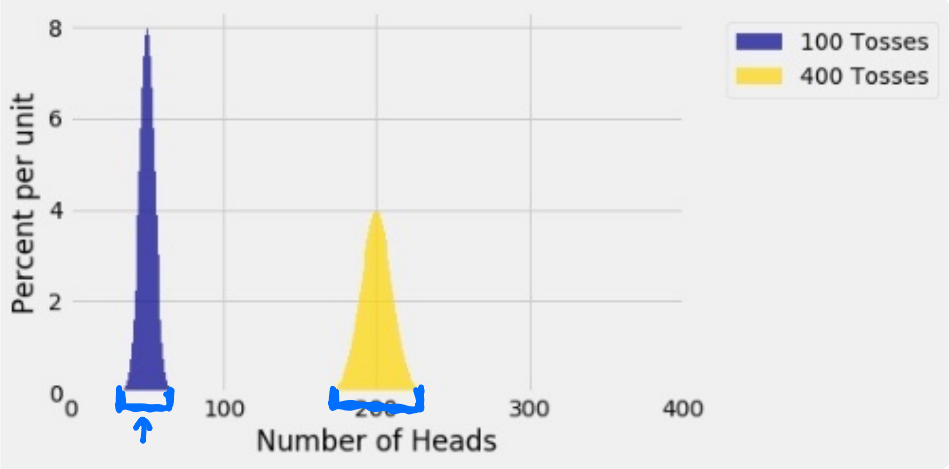
a) $S_{100} = 50$ or

b) $S_{400} = 200$?

$(\frac{1}{2})^{50}, (\frac{1}{2})^{50}$

$P(S_{100} = 50) = \binom{100}{50}(\frac{1}{2})^{100} = .08$ ← larger

$P(S_{400} = 200) = \binom{400}{200}(\frac{1}{2})^{400} = .04$



Both histograms have area 1.

Yellow histogram is more spread out and hence $P(S_{400} = 200) < P(S_{100} = 50)$.

We can see this mathematically,

$$\begin{aligned} \mathrm{Var}(S_n) &= \mathrm{Var}(X_1 + \cdots + X_n) \\ &= \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) \\ &= n\,\mathrm{Var}(X_1) \\ &= \boxed{n\sigma^2} = n\tfrac{1}{4} \end{aligned}$$

$\mathrm{Var}(\text{Bernoulli}(\tfrac{1}{2})) = \tfrac{1}{2}\cdot\tfrac{1}{2} = \boxed{1/4}$

$$\Rightarrow SD(S_n) = \sqrt{n}\,\sigma = \boxed{\sqrt{n}\,\tfrac{1}{2}}$$

The sample average

You toss a fair coin n times

Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(\tfrac{1}{2})$

$S_n = X_1 + X_2 + \cdots + X_n$ is the sample sum

let $A_n = S_n/n$ be the sample average

Where do you put your money?

a) $A_{100} = 50$ or

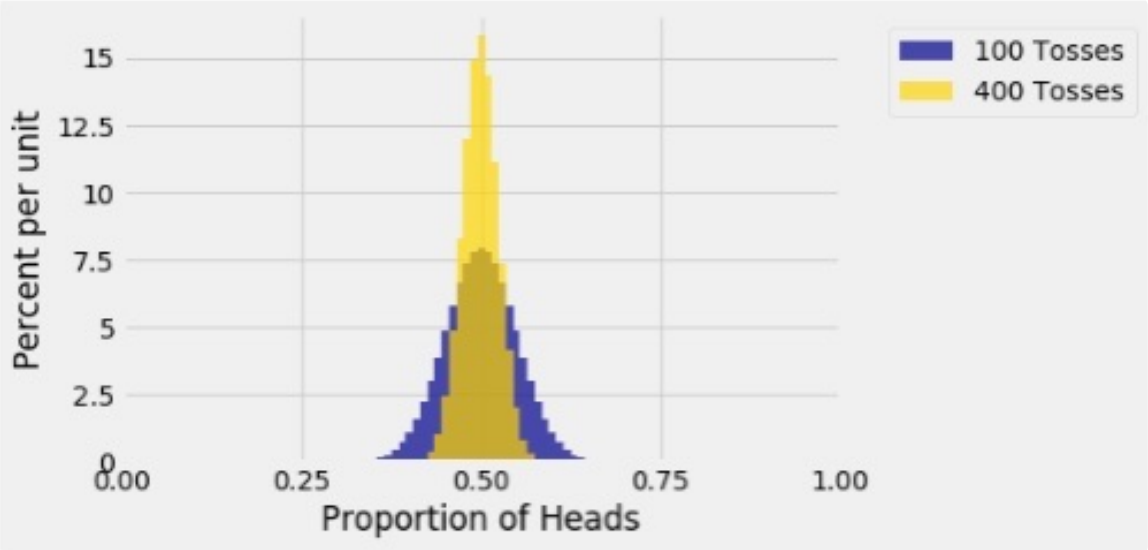
b) $A_{400} = 200$?

$$SD(A_n) = SD(S_n/n) = \frac{1}{n} SD(S_n) = \frac{1}{n}\sqrt{n}\,\sigma = \boxed{\frac{\sigma}{\sqrt{n}}}$$

$\sqrt{n}\,\sigma$

$$SD(A_{100}) = \frac{1/2}{10} = 1/20$$

$$SD(A_{400}) = \frac{1/2}{20} = 1/40 \leftarrow \text{less error}$$



In general, the larger the sample size n, the more likely it is that the sample average A_n will be close to the population average μ.

This is the "law of averages".

Appendix

Boring algebra details

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

pull out $\frac{nG}{N}$

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

common denom.

$$= \frac{nG}{N}\,\frac{(N-1)N + N(n-1)(G-1) - nG(N-1)}{N(N-1)}$$

Factor out

$$= \frac{nG}{N}\,\frac{N^2 - N + NnG - NG - nN + N - nGN + nG}{N(N-1)}$$

reverse FOIL

$$= \frac{nG}{N}\,\frac{(N-G)(N-n)}{N(N-1)}$$

$$= \boxed{n\frac{G}{N}\cdot\frac{N-G}{N}\cdot\frac{N-n}{N-1}}$$