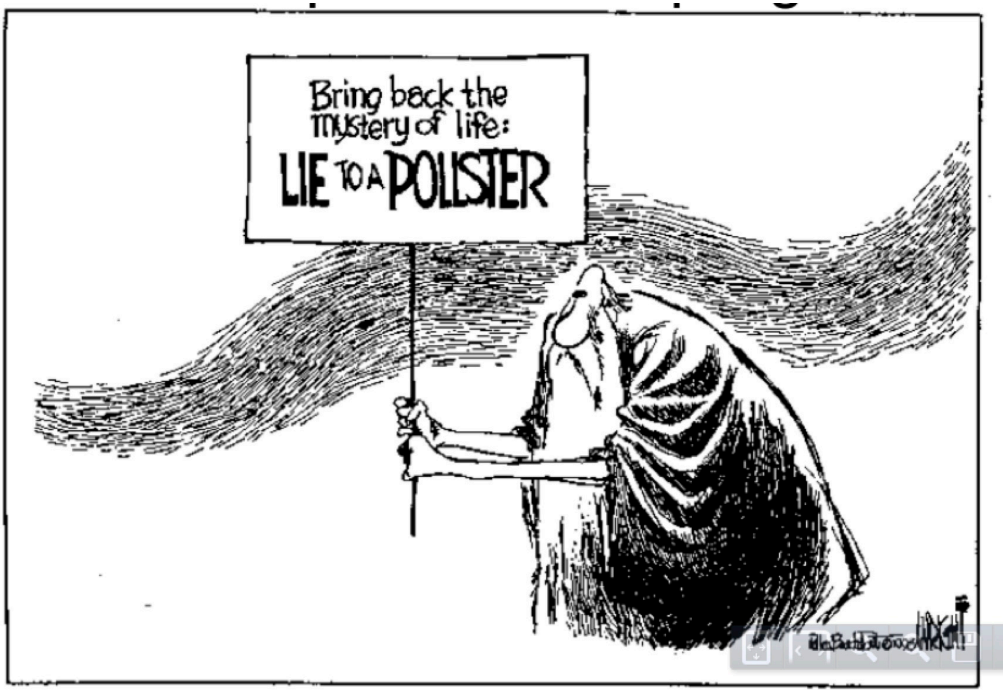
Stat 88: Prob. & Math. Statistics in Data Science



Lecture 18: 3/29/2022

Sampling without replacement, the law of averages, distribution of a sample sum

7.2, 7.3, 8.1

Review: variance, SD, inequalities when we don't know dsn

$E(X) = \mu$ $\quad E((X-\mu)^2) = E(X^2) - \mu^2$

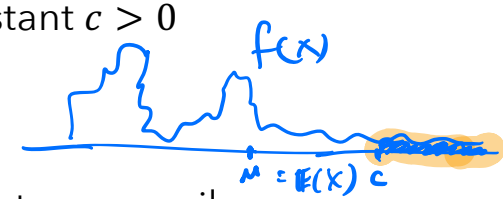
1. The variance of a rv X: $Var(X) = \sigma^2 = E[(X - E(X))^2] = E(X^2) - (E(X))^2$
2. The SD or standard deviation is given by $SD(X) = \sigma = \sqrt{Var(X)} \geq 0$

If X and Y are ***independent***, then

$Var(X_1 + X_2 + \cdots + X_n) = Var(X_1) + Var(X_2) + \cdots + Var(X_n)$

3. $Var(X + Y) = Var(X) + Var(Y)$ & $SD(X + Y) = \sqrt{Var(X) + Var(Y)}$
4. **Markov's Inequality**: For a nonnegative rv X, and constant $c > 0$

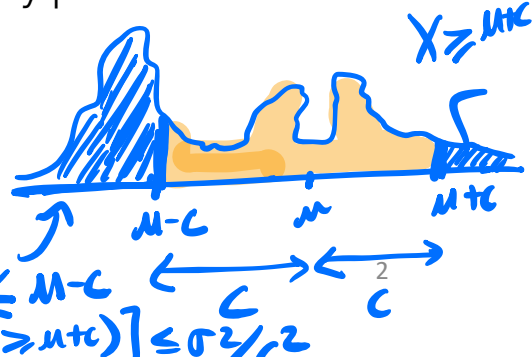
$$P(X \geq c) \leq \frac{E(X)}{c}$$



5. **Chebyshev's inequality**: For any random variable X (not necessarily non-negative), with mean μ and standard deviation σ, for any positive constant $c > 0$, we have:

$$P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2} = \frac{Var(X)}{c^2}$$

dist b/w X & μ



$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$

$P(\mu-c < X < \mu+c) = 1 - P(X \leq \mu-c \text{ or } X \geq \mu+c)$

$\geq 1 - \sigma^2/c^2$

Example

A list of non negative numbers has an average of 1 and an SD of 2. Let p be the proportion of numbers over 4. To get an upper bound for p, you should:

a) ~~Assume a binomial distribution~~

b) Use Markov's inequality.

c) Use Chebyshev's inequality

d) ~~None of the above~~.

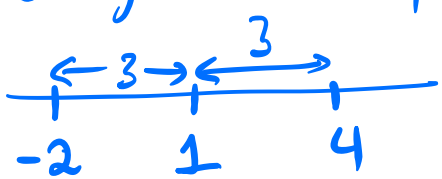
p = proportion of #s over 4

If X is a # picked at random from list. $E(X) = 1$, $SD(X) = 2$

(b) $P(X > 4) \leq \frac{1}{4} \leftarrow E(X)$

$\uparrow c$

(c) Chebyshev's inequality



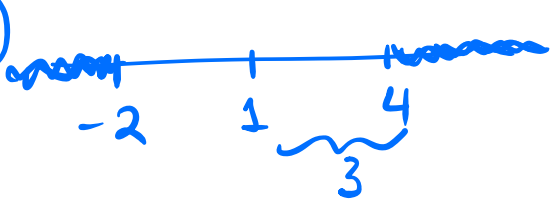
$P(|X-1| \geq 3) = P(X \leq -2) + P(X \geq 4) \geq P(X \geq 4)$

$\uparrow c$

$P(X \geq 4) \leq [P(X \leq -2) + P(X \geq 4)] \leq \frac{\sigma^2}{c^2} = \frac{2^2}{3^2} = \frac{4}{9}$

Chebyshev's ineq $P(|X-\mu| \geq c) \leq \frac{\sigma^2}{c^2}$

$P(X>4)$ $P(|X-1| \geq c)$ $c=3$



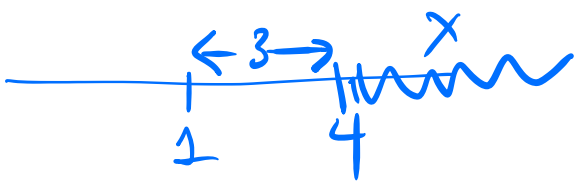
$P(|X-1| \geq 3)$

$= P(X \leq -2) + P(X \geq 4)$ ($P(X \leq -2) = 0$)

$\geq P(X \geq 4) \geq P(X > 4)$

$P(X>4) \leq P(X \geq 4) \leq \frac{\sigma^2}{c^2} = \frac{2^2}{3^2} = \frac{4}{9}$

Want $\rightarrow P(|X-1| \geq 3)$



Answer: (b) Markov's inequality gives a better bound.

Recap: sums of iid random variables

- Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with mean μ and variance σ^2. Define $S_n = X_1 + X_2 + \cdots + X_n$
- $E(S_n) = \sum E(X_k) = n\mu$ & $Var(S_n) = Var(X_1) + \cdots + Var(X_n) = n\sigma^2$
- $SD(S_n) = \sqrt{n}\,\sigma$ (Square root law: sd grows by a factor of $\sqrt{n}$, **not** n)

- $X \sim Bin(n,p)$: $E(X) = np, Var(X) = np(1-p), SD(X) = \sqrt{np(1-p)}$
- $X \sim$ Poisson(μ): $E(X) = Var(X) = \mu, SD = \sqrt{\mu}$
- $X \sim$ Geometric(p) distribution: $E(X) = \frac{1}{p}$, $Var(X) = \frac{1-p}{p^2}$

7.3: Sampling without replacement

Draw tickets from a box without replacement

- When we have a simple random sample (SRS), the draws are without replacement (like drawing cards from a deck).
- The random variables are no longer independent
- So, how do we compute the variance of the sum of draws of a SRS?
- To begin with, let's look at the squares and products of indicators
- If I_A and I_B are indicator functions, what can we say about I_A^2 and $I_A I_B$?

I_A is an indicator for event A, $E(I_A) = P(A)$

I_B " " " B,

$E(I_B) = P(B)$

$I_A = \begin{cases} 1 & \text{if A true} \\ 0 & \text{if A not true.} \end{cases}$

$I_A^2 = \begin{cases} 1^2, & \text{A true} \\ 0^2, & \text{if A false.} \end{cases}$

$E(I_A^2) = 1 \cdot P(A) + 0 \cdot P(A^c)$

$= P(A)$

$I_A \cdot I_B = \begin{cases} 1, & \text{if BOTH A \& B are true} \\ 0, & \text{o/w} \end{cases}$

$E(I_A I_B) = P(A \cap B)$

$E(I_j I_k) = P(I_j = 1 \ \& \ I_k = 1) = P(I_j = 1)\, P(I_k = 1 \mid I_j = 1)$

Variance of a hypergeometric random variable

total popn. # of "Good" tickets, n = sample size

- 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.

$P(I_k = 1) = \frac{G}{N}$ ←

$E(X) = E(I_1) + E(I_2) + \cdots = \frac{G}{N} + \frac{G}{N} + \cdots + \frac{G}{N} = \frac{nG}{N} = \frac{G}{N} + \frac{G}{N} + \cdots \frac{G}{N}$

- 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: $Var(X) = E(X^2) - \big(E(X)\big)^2 = E(X^2) - \left(\frac{nG}{N}\right)^2$

$$X^2 = (I_1 + I_2 + \cdots + I_n)^2 = \sum_{k=1}^{n} I_k^2 + \sum_{j} \sum_{k, k \neq j} I_j I_k$$

$$E(X^2) = nE\big(I_k^2\big) + n(n-1)E\big(I_j I_k\big)$$

$$= n\frac{G}{N} + n(n-1)P\big(I_j = 1\big)P\big(I_k = 1 \mid I_j = 1\big)$$

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

$X = I_1 + I_2 + I_3 + \cdots + I_n$, $\mathbb{E}(X) = n \cdot \frac{G}{N}$

$\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \mathbb{E}(X^2) - \left(n\frac{G}{N}\right)^2$

$X^2 = (I_1 + I_2 + \cdots + I_n)^2$

For example, if $n=2$: $(I_1+I_2)^2 = I_1^2 + I_2^2 + I_1 I_2 + I_2 I_1$

$X^2 = \underbrace{I_1^2 + I_2^2 + \cdots + I_n^2}_{n} + \underbrace{\sum_{j=1}^{n}\sum_{\substack{k=1 \\ j\neq k}}^{n} I_j I_k}_{j\neq k}$

$n=3$: $(I_1+I_2+I_3)(I_1+I_2+I_3) = I_1^2 + I_1I_2 + I_1I_3 + I_2I_1 + I_2^2 + I_2I_3 + I_3I_1 + I_3I_2 + I_3^2$

$\mathbb{E}(X^2) = \mathbb{E}\left(I_1^2 + I_2^2 + \cdots + I_n^2\right) + \mathbb{E}\left(\sum_{j=1}^{n}\sum_{\substack{k=1 \\ j\neq k}}^{n} I_j I_k\right)$

$\mathbb{E}(I_1^2) = \mathbb{E}(I_1) = \frac{G}{N}$

$= n \cdot \mathbb{E}(I_1^2) + \sum_{j=1}^{n}\sum_{\substack{k=1 \\ j\neq k}}^{n} \mathbb{E}(I_j I_k)$

$= n \cdot \mathbb{E}(I_1^2) + n(n-1)\,\mathbb{E}(I_j I_k)$ — all the same

$= n \cdot \frac{G}{N} + n(n-1)\,P(I_j=1)\,P(I_k=1 \mid I_j=1)$

$P(I_j=1) = \frac{G}{N}$, $P(I_k=1 \mid I_j=1) = \frac{G-1}{N-1}$

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

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

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

$E(I_j I_k) = P(I_j=1 \ \& \ I_k=1) = P(I_j=1)\,P(I_k=1 \mid I_j=1)$

Variance of a hypergeometric random variable

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

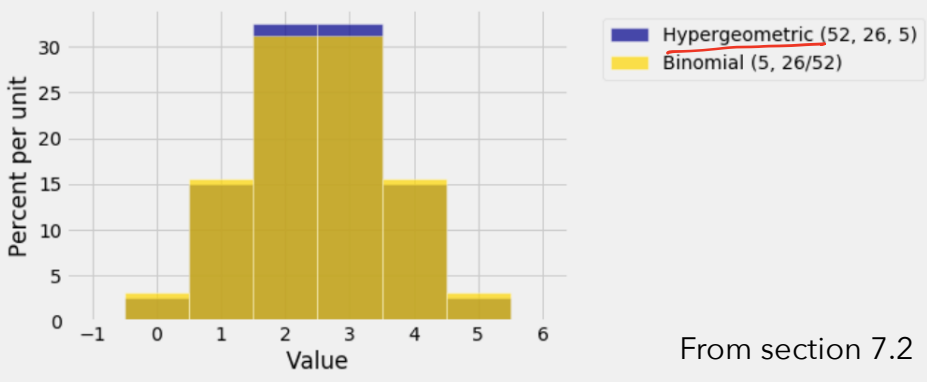
$$= \frac{nG}{N}\left[\frac{N(N-1) + (nN-N)(G-1) - nGN + nG}{N(N-1)}\right]$$

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

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

$$= \frac{nG}{N}\cdot\frac{(N-n)(N-G)}{N(N-1)} = n\underbrace{\left(\frac{G}{N}\right)}_{P(S)}\underbrace{\left(\frac{N-G}{N}\right)}_{P(F)}\left[\frac{N-n}{N-1}\right]$$

The finite population correction & the accuracy of SRS



From section 7.2

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

Note that fpc ≤ 1

So $SD(HG) \leq SD(Bin)$

SD is less when we don't have independence

In general we have that the :

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

$$SD(HG) = SD(\text{binomial}) \times \sqrt{\frac{N-n}{N-1}}$$

$$Var(HG) = Var(Bin) \cdot \left(\frac{N-n}{N-1}\right)$$

"Try this at home": Plug in different values of N, n & see what is fpc.

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.

False

Try fpc with $N = 121485$, $n = 625$

$N = 4 \times 10^6$, $n = 625$

Accuracy depends on SD & fpc

If fpc ≈ 1, then only depends on SD.

Example (adapted 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 2022 Oscars. 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.

Exercise

Law of Averages

- Essentially a statement that you are already familiar with: If you toss a fair coin many times, roughly half the tosses will land heads.

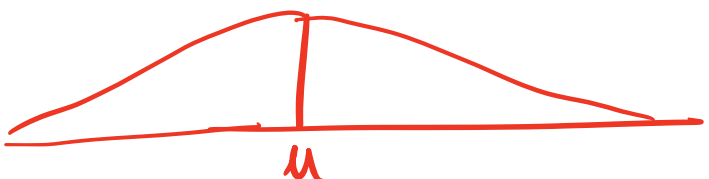
- We are going to consider sample sums and sample means of iid random variables $X_1, X_2, \ldots, X_n$ where the mean of each X_k is μ and the variance of each X_k is σ^2.

$X_k \sim \mu, \sigma^2$

- Recall the ***sample sum*** $S_n = X_1 + X_2 + \cdots + X_n$, with $E(S_n) = n\mu$, $Var(S_n) = n\sigma^2, SD(X_n) = \sqrt{n}\sigma$

- We see here, as we take more and more draws, the variability of the sum keeps increasing, which means the values get more and more dispersed around the mean ($n\mu$).

S_n $\quad Var(S_n) = n\sigma^2$

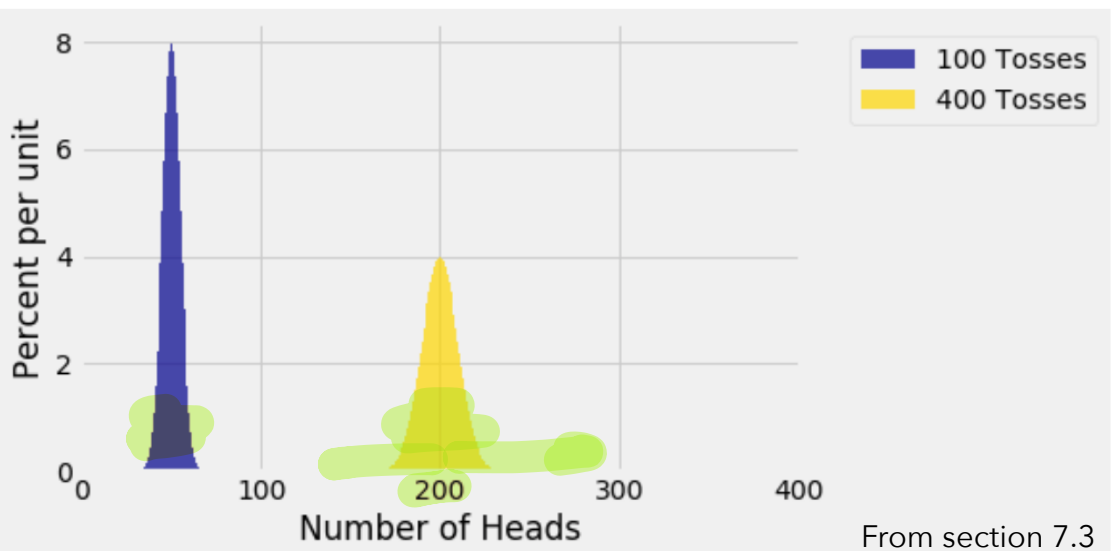


Coin tosses

- Consider a fair coin, toss it 100 times & 400 times, count the number of H

Expect in first case, roughly 50 H, and in second, roughly 200 H.

- So do you think chance of 50 H in 100 tosses and 200 H in 400 tosses should be the same?



From section 7.3

Total area = 1 in both cases
so spread about the mean in 400 tosses is greater.

Example: Coin toss

$S_{100} \sim Bin(100, \frac{1}{2})$

$S_{400} \sim Bin(400, \frac{1}{2})$

- $SD(S_{100}) = \sqrt{npq} = \sqrt{100 \cdot \frac{1}{2} \cdot \frac{1}{2}} = 10 \cdot \frac{1}{2} = 5$
- $SD(S_{400}) = \sqrt{npq} = \sqrt{400 \cdot \frac{1}{2} \cdot \frac{1}{2}} = 20 \cdot \frac{1}{2} = 10$

- P(200 H in 400 tosses) $= P(S_{400} = 200) = \binom{400}{200}\left(\frac{1}{2}\right)^{400} \approx 0.04$

- P(50 H in 100 tosses) $P(S_{100} = 50) = \binom{100}{50}\left(\frac{1}{2}\right)^{100} \approx 0.08$

Law of Averages for a fair coin

SD increase
error increases

- Notice that as the number of tosses of a fair coin increases, the *observed error* (number of heads – half the number of tosses) increases. This is governed by the standard error.

- The *percentage* of heads observed comes very close to 50%

- *Law of averages*: The long run *proportion* of heads is very close to 50%.

sample mean ⟶ prob of H as n gets large.

Sample sum, sample average, and the square root law

- $S_n = X_1 + X_2 + \cdots + X_n$ A_n = sample mean $= \frac{S_n}{n}$
- Let $A_n = {}^{S_n}/_{n}$, so A_n is the average of the sample (or sample mean).
- If the X_k are indicators, then A_n is a proportion (proportion of successes)

- Note that $E(A_n) = \mu$ and $SD(A_n) = ??$ $SD\left(\frac{S_n}{n}\right) = \frac{\sqrt{n}\cdot\sigma}{n} = \frac{\sigma}{\sqrt{n}}$

as n gets large SD(Sn) gets large, SD(An) gets small.

- **The square root law:** the *accuracy* of an estimator is measured by its SD, the ***smaller*** the SD, the ***more accurate*** the estimator, but if you multiply the sample size by a factor, the accuracy only goes up by the **square root** of the factor.
- In our earlier example, we _____ the accuracy by quadrupling the size.