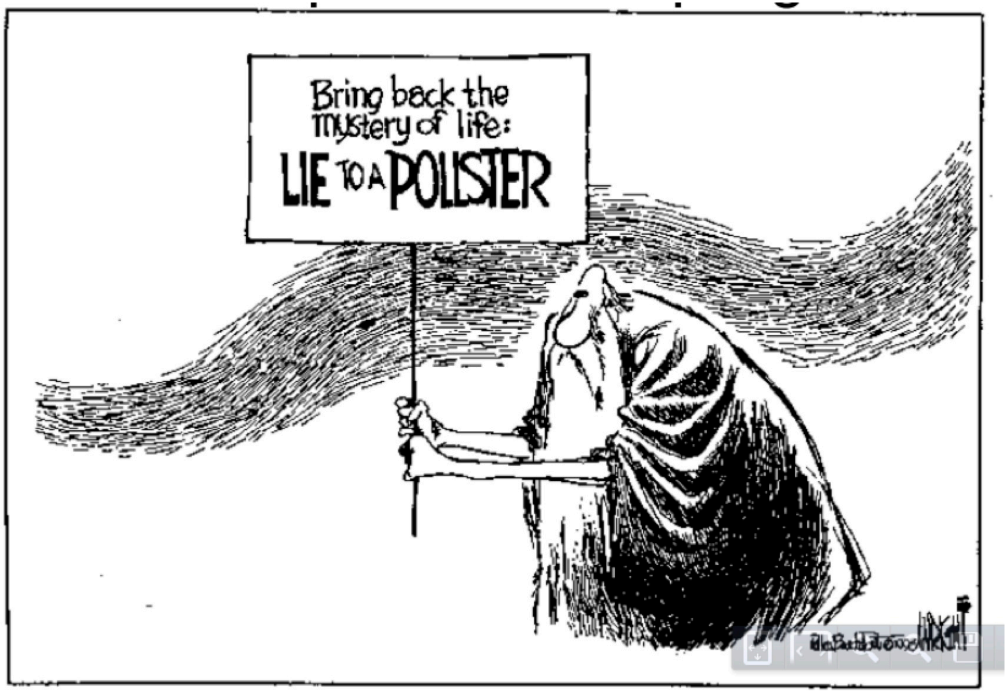
Stat 88: Probability & Mathematical Statistics in Data Science



Lecture 23: 3/15/2021

Sections 7.1, 7.2

Sums of RVs and sampling without replacement

Warm up

Let X be a non-negative random variable such that $E(X) = 100 = Var(X)$.

a) Can you find $E(X^2)$ exactly? If not, what can you say?

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

$$100 = E(X^2) - (100^2)$$

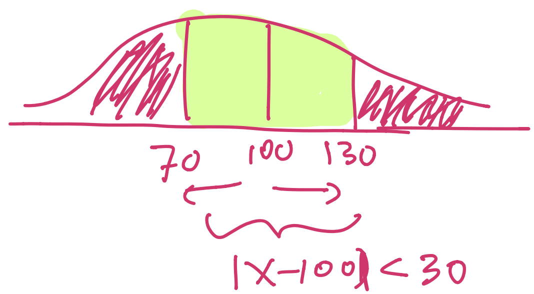
$$E(X^2) = 100 + 10{,}000 = 10100$$

$\mu_X = E(X)$ $\quad \sigma_X^2 = Var(X)$

b) Can you find $P(70 < X < 130)$ exactly? If not, what can you say?

Chebyshev $\rightarrow P(|X-\mu| \geq c) \leq \dfrac{Var(X)}{c^2}$

$$P(|X-100| \geq 30) \leq \frac{1\cancel{0}\cancel{0}}{9\cancel{0}\cancel{0}} = \frac{1}{9} \qquad \uparrow 30^2$$



$$P(|X-100| < 30) \geq 1 - \frac{1}{9} = \frac{8}{9}$$

7.1: Sums of Independent Random Variables

- Recall that expectation is additive, which we used many times.

$$(E(X+Y) = E(X) + E(Y))$$

- What about $Var(X+Y)$? Well, it depends.

$H \sim Bin(10, \frac{1}{2})$, $T \sim Bin(10, \frac{1}{2})$

- Consider tossing a fair coin 10 times. Let H be the number of heads and T be the number of tails in 10 tosses. Then $H+T=10$. Note that $Var(H), Var(T) \neq 0$, but $Var(H+T) = Var(10) = 0$

$Var(H+T) = 0 \neq Var(H) + Var(T)$

$H_1 \sim Bin(5, \frac{1}{2})$

- But now let H_1 be the number of heads in the first 5 tosses, and H_2 the number of heads in the last 5 tosses. Will we have that $Var(H_1 + H_2) = 0$?

$Var(H_1 + H_2) \neq 0$, b/c $H_1 + H_2$ can vary.

$H_2 \sim Bin(5, \frac{1}{2})$

- It turns out that if X and Y are ***independent***, then we have that

$$Var(X+Y) = Var(X) + Var(Y)$$

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

IF $X_1, \dots, X_n$ ARE INDEPENDENT

Sums of iid random variables

- Let $X_1, X_2, \dots, X_n$ be independent and identically distributed random variables with mean μ and variance σ^2. Define S_n to be their sum:

$$S_n = X_1 + X_2 + \cdots + X_n.$$

- We already know that $E(S_n) = \sum_{k=1}^{n} E(X_k) = n\mu.$
- Now we can further say that:

Variance of the sum = sum of the variances.

$$Var(S_n) = Var(X_1 + X_2 + \cdots + X_n) = Var(X_1) + \cdots + Var(X_n) = n\sigma^2$$

$$SD(S_n) = \sqrt{n}\,\sigma = \sqrt{Var(S_n)}$$

- Notice that the expected value grows as n, but the sd grows as $\sqrt{n}$.

$\mathbb{E}(X_1) = \mu$ $\quad Var(X_1) = \sigma^2$ $\quad SD(X_1) = \sigma$

$\mathbb{E}(X_1+X_2) = 2\mu$ $\quad Var(X_1+X_2) = 2\sigma^2$ $\quad SD(X_1+X_2) = \sqrt{2\sigma^2} = \sqrt{2}\,\sigma$

$\mathbb{E}(X_1+X_2+X_3) = 3\mu$ $\quad Var(X_1+X_2+X_3) = 3\sigma^2$ $\quad SD(X_1+X_2+X_3) = \sqrt{3\sigma^2} = \sqrt{3}\,\sigma$

$\mathbb{E}(X_1+X_2+X_3+X_4) = 4\mu$ $\quad Var(X_1+X_2+X_3+X_4) = 4\sigma^2$, $\quad SD(X_1+X_2+X_3+X_4) = 2\sigma = \sqrt{4}\,\sigma$

Variance of the Binomial distribution

- Recall that a binomial random variable $X \sim Bin(n, p)$ is the sum of n iid Bernoulli(p) random variables $I_1, I_2, \ldots, I_n$ where I_k is the indicator of success on the kth trial.
- What are the mean and variance of I_k? And therefore, what are the mean and variance of X? For what p will this variance be maximum?

$X \sim Bin(n,p) \qquad X = I_1 + I_2 + \cdots + I_n$

$$I_k = \begin{cases} 1 & \text{if } k^{th} \text{ trial is a success} \\ 0 & \text{o/w.} \end{cases}$$

$\mathbb{E}(I_k) = P(I_k = 1) = p$

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

$= p - p^2$

$= p(1-p)$

$\boxed{Var(I_k) = \mathbb{E}(I_k^2) - (\mathbb{E}(I_k))^2}$

$\mathbb{E}(X) = n \cdot p$

$Var(X) = n \cdot p(1-p)$

$SD(X) = \sqrt{np(1-p)}$ if we define $q = 1-p$

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

Variance of Poisson (μ) and geometric(p)

- Recall that one way to get the Poisson rv is by approximating the Binomial(n, p) distribution when n is large and p is small. ($\mu = np$)

- SD of the binomial distribution is $\sqrt{np(1-p)}$. $= \sqrt{npq}$, $q = 1-p$
- Note that if p is small, $(1-p) \approx 1$, and we can say that $np(1-p) \approx np$.
- This gives us that the SD of the Poisson(μ) distribution is $\sqrt{\mu}$

If $X \sim Pois(\mu)$, $SD(X) = \sqrt{\mu}$

$E(X) = \mu = Var(X)$

- Geometric($1/p$) distribution: **Fact**: the variance of the geometric distribution is $\frac{1-p}{p^2}$

$X \sim Bin(n, p)$, $Var(X) = np(1-p)$

$X \sim Pois(\mu)$ $var(X) = \mu$

$X \sim Geom(p)$ $Var(X) = \frac{1-p}{p^2}$

- Ex: (Waiting till the 10th success) Suppose you roll a die until the 10th success. Let R be the number of rolls required. Find $SD(R)$. $P(S) = \frac{1}{6}$

$R = X_1 + X_2 + \cdots + X_{10}$, $X_k \sim Geom(\frac{1}{6})$

$\underbrace{- - - - S_1}\ \underbrace{- - - - - S_2}\ \underbrace{- - - - - S_3}$

T_1 Geom$(\frac{1}{6})$ $T_1 =$ Geom$(\frac{1}{6})$ Geom$(\frac{1}{6})$

$$Var(R) = \sum_{i=1}^{10} Var(X_k) = \frac{10(1-p)}{p^2}$$

$$= \frac{10 \cdot 5/6}{1/36} = 10 \cdot 30 = 300.$$

Exercise 7.4.5

The number of typos on the cover page of an exam has a distribution given by

x	0	1
$P(X = x)$	0.8	0.2

T_c # of typos on cover page is Bernoulli(0.2)

T_R # of typos in rest of exam ~ Poisson(3)

Total # of typos = X, $E(X) = 0.2 + 3$
$= 3.2$

$X = T_c + T_R$

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.

Var(X) = sum of variances.
$= (0.2)(0.8) + 3$
$= 0.16 + 3 = 3.16.$

Variance of Bernoulli
$= p(1-p)$
(Bin. w/ $n=1$)

$$E(X) = E(T_c) + E(T_R)$$

$$Var(X) = Var(T_c) + Var(T_R) = 3.16$$

$$SD(X) = \sqrt{Var(X)} = \sqrt{3.16}$$

Sampling 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 not independent any more.
- 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 $I_A I_B$?

A, B be events, $P(A) \neq 0$, $P(B) \neq 0$

$I_A = \begin{cases} 1, & \text{if } A \text{ true} \\ 0, & \text{o/w} \end{cases}$, I_B indicator of B.

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

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

$I_A I_B = \begin{cases} 1, & I_A = 1 \text{ AND } I_B = 1 \\ & \text{if } AB \text{ is true} \\ 0, & \text{otherwise} \end{cases}$

$\mathbb{E}(I_A I_B) = P(AB)$

Variance of a hypergeometric random variable

Hypergeometric

- 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.

$I_k = \begin{cases} 1, & \text{if } k^{th} \text{ draw is } S \\ 0, & \text{if } k^{th} \text{ draw is } F \end{cases}$

N tkts $\boxed{S}$ $\boxed{F}$ → n tkts w/o repl.

$\boxed{1}$ $\boxed{0}$

- We can compute the expectation of X using symmetry: $E(X)$ $= \mathbb{E}(I_1 + I_2 + \cdots + I_n)$

$= \sum_{k=1}^{n} \mathbb{E}(I_k)$

$= n \cdot \frac{G}{N}$ ← symmetry

- But what about variance?

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

$$\mathbb{E}(X^2) = \mathbb{E}\left[(I_1 + I_2 + \cdots + I_n)^2\right] =$$

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

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

Variance of a hypergeometric random variable

$$\mathbb{E}(X^2) = n\,\mathbb{E}(I_j^2) + n \cdot (n-1)\,\mathbb{E}(I_j I_k)$$

$$\mathbb{E}(I_j I_k) = P(I_j \cdot I_k = 1)$$