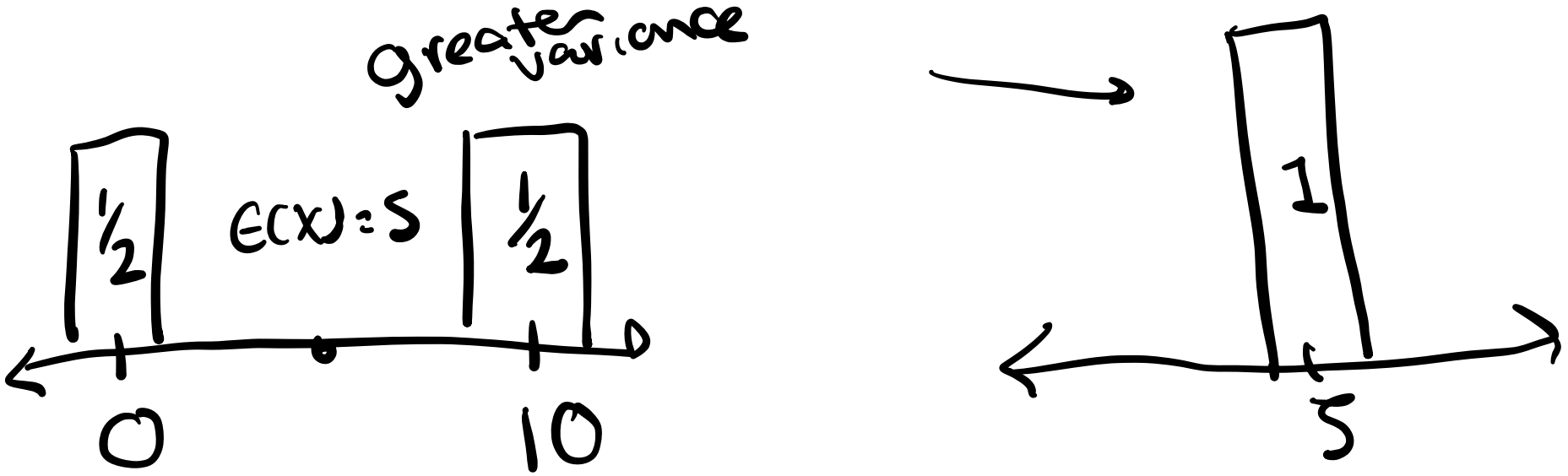
Quiz Day !!
↙
7pm (not Berkeley time)

Lecture 5C: Variance, Covariance & Correlation

UC Berkeley CS70
Summer 2023
Nikki Suzani

What is variance?

How much does a random variable deviate from its mean?



Variance (formally)

$E[cX] = cE[X]$

$E[c] = c$

Definition 16.1 (Variance). *For a r.v. X with expectation $\mathbb{E}[X] = \mu$, the variance of X is defined to be*

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mu)^2\right].$$

value

mean

squared deviation from the mean

The square root $\sigma(X) := \sqrt{\mathrm{Var}(X)}$ is called the standard deviation of X.

$$E[(X - E[X])^2] = E[X^2 - 2XE[X] + E[X]^2]$$

$$= E[X^2] - E[2XE[X]] + E[E[X]^2]$$

$$= E[X^2] - 2E[X]E[X] + E[X]^2$$

$$= E[X^2] - (E[X])^2$$

Variance of a Coin Toss

I flip a fair coin. What is the variance of ~~my coin toss~~ my number of heads?

X: # of heads on my flip

X { 1 w/ prob $\frac{1}{2}$; 0 w/ prob $\frac{1}{2}$

$E[X^2] - E[X]^2$

$E[X^2] = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{2}$

$\frac{1}{2} - (\frac{1}{2})^2 = \frac{1}{2}(1 - \frac{1}{2}) = \frac{1}{4}$

X^2	X	$P(X=x)$
0	0	$\frac{1}{2}$
1	1	$\frac{1}{2}$

Variance of a Dice Roll

I roll a single six-sided dice. What is the variance of my dice roll?

X: Value of my dice roll

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

$$= \left(\frac{1+4+9+16+25+36}{6}\right) - \left(\frac{1+2+3+4+5+6}{6}\right)^2$$

$$= \frac{35}{12}$$

X^2	X	$P(X=x)$
1	1	1/6
4	2	1/6
9	3	1/6
16	4	1/6
25	5	1/6
36	6	1/6

Variance of a Uniform Random Variable

$$E(X) = \sum_{k=0}^{\infty} k P(X=k)$$

$$E(X^2) = \sum_{k=0}^{\infty} k^2 P(X=k)$$

$X \sim \text{Uniform}(n)$

Let X be a random variable that takes on the values 1, … , n with the same probability (1/n). What is Var(X)?

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

$$= \sum_{k=0}^{\infty} k^2 \underbrace{P(X=k)}_{=\frac{1}{n}} - \left(\sum_{k=0}^{\infty} k P(X=k)\right)^2$$

$$= \frac{1}{n}\sum_{k=0}^{\infty} k^2 - \left(\frac{1}{n}\sum_{k=0}^{\infty} k\right)^2$$

Variance of a Uniform Random Variable (cont.)

$$\frac{1}{n} \cdot \frac{n(n+1)(2n+1)}{6} - \left(\frac{1}{n} \frac{n(n+1)}{2}\right)^2$$

$$= \frac{n^2-1}{12}$$

Example: Variance of a Dice Roll (again)

Roll a fair six-sided dice. What is the variance?

$\{1, \ldots 6\}$
each w/ prob $\frac{1}{6}$

$X \sim \text{Uniform}(6)$

$$\text{Var}(X) = \frac{n^2-1}{12} = \frac{36-1}{12} = \frac{35}{12}$$

$SD = \sqrt{Var}$

$E[cX] = cE[X]$

Properties of Variance

$$\text{Var}(cX) = E[(cX)^2] - (E[cX])^2$$

$$SD(cX) = E[c^2X^2] - (cE[X])^2$$

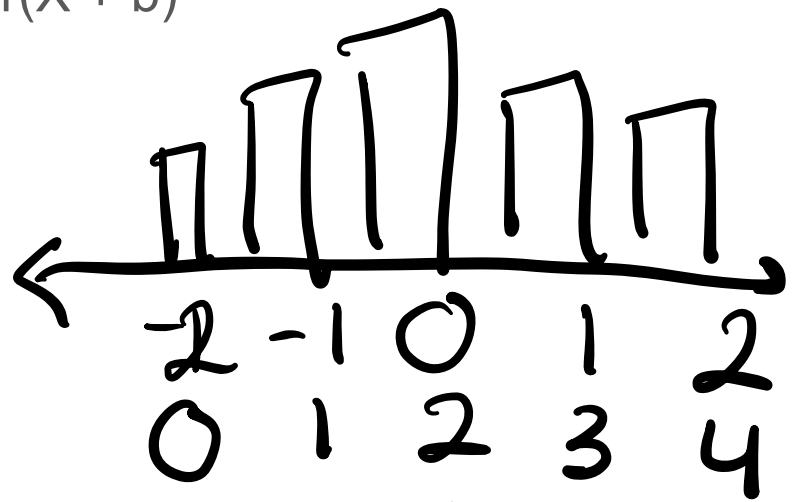
$$= c^2E[X^2] - c^2E[X]^2$$

$$= c^2(E[X^2] - E[X]^2)$$

$$= c^2(\text{Var}(X)) \rightarrow c\,SD(X)$$

Properties of Variance

Var(X + b)



Var(X + 2)

$$E((X+b)^2) - E(X+b)^2$$

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

$$= Var(X)$$

Properties of Variance

Var(X+b)

$0 \cdot 1-p + 1 \cdot p = p$

Variance of the Bernoulli Distribution

X^2	X	$P(X=x)$
0	0	$1-p$
1	1	p

Let X ~ Bernoulli(p). What is Var(X)?

$$X: \begin{cases} 1 & \text{w/ probability } p \\ 0 & \text{w/ probability } 1-p \end{cases}$$

$$\text{Var}(X) = E(X^2) - E(X)^2 = \boxed{p(1-p)}$$

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

$$= E(X)(1 - E(X))$$

Making Serves

Nikki plays tennis (badly). Her serve ends up in the "legal zone" 20% of the time. What is the variance of the ~~number of times~~ one serve her serve is legal?

serve one time

$$X: \begin{cases} 1 & \text{w/ prob } 0.2 \\ 0 & \text{w/ prob } 0.8 \end{cases} \qquad X \sim \text{Bernoulli}(0.2)$$

$$\text{Var}(X) = p(1-p) = 0.2(1-0.2) = 0.2 \cdot 0.8 = 0.16$$

Variance of Independent Random Variables

If X and Y are **independent** random variables, Var(X + Y) = Var(X) + Var(Y)

possibly true when not independent

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$$

all independent trials

Variance of the Binomial Distribution

Let X ~ Binomial(n, p). What is Var(X)?

$X_i \sim \text{Bernoulli}(p)$

$$X = X_1 + X_2 + \dots X_n$$

$$\text{Var}(X) = \text{Var}(X_1 + X_2 + \dots X_n)$$

$$= \text{Var}(X_1) + \text{Var}(X_2) + \dots \text{Var}(X_n)$$

$$= p(1-p) + \underbrace{p(1-p) + \dots p(1-p)}_{= np(1-p)}$$

$P(1-P)$

$(1-P)P$

Example: Variance of 10 coin tosses

You toss a coin ten times that flips heads with probability ⅔. What is the variance of the number of heads?

X: # of heads $\quad X = X_1 + X_2 + \dots X_{10}$

$X_i \sim$ 1 if got a heads, 0 otherwise $\quad X_i \sim \text{Bernoulli}(2/3)$

$X \sim \text{Binomial}(10, 2/3) = 10 \cdot 2/3 (1 - 2/3)$

$= 20/9$

Example: More tennis

Carlos Alcaraz wins matches against Novak Djokovic 1/3 of the time. If they play five times next year, what is the variance of the number of matches Alcaraz wins?

Year	Event	Surface	RND	Winner	Result
2023	Wimbledon Great Britain	Outdoor Grass	F	Carlos Alcaraz	16 76^{6} 61 36 64
2023	Roland Garros France	Outdoor Clay	SF	Novak Djokovic	63 57 61 61
2022	ATP Masters 1000 Madrid Spain	Outdoor Clay	SF	Carlos Alcaraz	67^{5} 75 76^{5}

$X = X_1 + X_2 + \ldots X_5$ $\quad X_i \sim \text{Bernoulli}(1/3)$

$X \sim \text{Binomial}(5, 1/3)$ $\quad \text{Var}(X) = 5 \cdot 1/3 \cdot (1 - 1/3) = 10$

$= \frac{}{9}$

What is covariance?

Cov(X, Y) gives us information about **how related** X and Y are to each other.

Covariance is positive

when X is big, Y is big

Covariance is negative

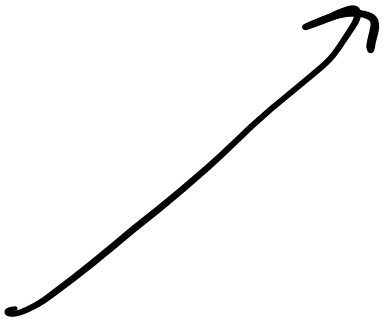
When X is small, Y is big

Covariance (Formally)

Definition 16.2 (Covariance). *The* covariance *of random variables X and Y, denoted* $\mathrm{Cov}(X,Y)$*, is defined as*

$$\mathrm{Cov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y],$$

where $\mu_X = \mathbb{E}[X]$ *and* $\mu_Y = \mathbb{E}[Y]$.



“Mean of the product minus the product of the means”

Covariance is Bilinear

$$\text{Cov}(aX + bY, aX + bY) = \text{Cov}(aX, aX) + \text{Cov}(aX, bY) + \text{Cov}(bY, aX) + \text{Cov}(bY, bY)$$

$$a\,\text{Cov}(X, aX) = a^2\,\text{Cov}(X, X)$$

$$ab\,\text{Cov}(X, Y) = ab\,\text{Cov}(Y, X)$$

$$b^2\,\text{Cov}(Y, Y)$$

If X & Y are independent, they are uncorrelated

→ $Cov(X,Y) = 0$

E[XY] = E[X]E[Y] if X and Y are independent

$$E[XY] = \sum_x \sum_y xy\,P(X=x \wedge Y=y)$$

$$= \sum_x \sum_y xy\,P(X=x)P(Y=y)$$

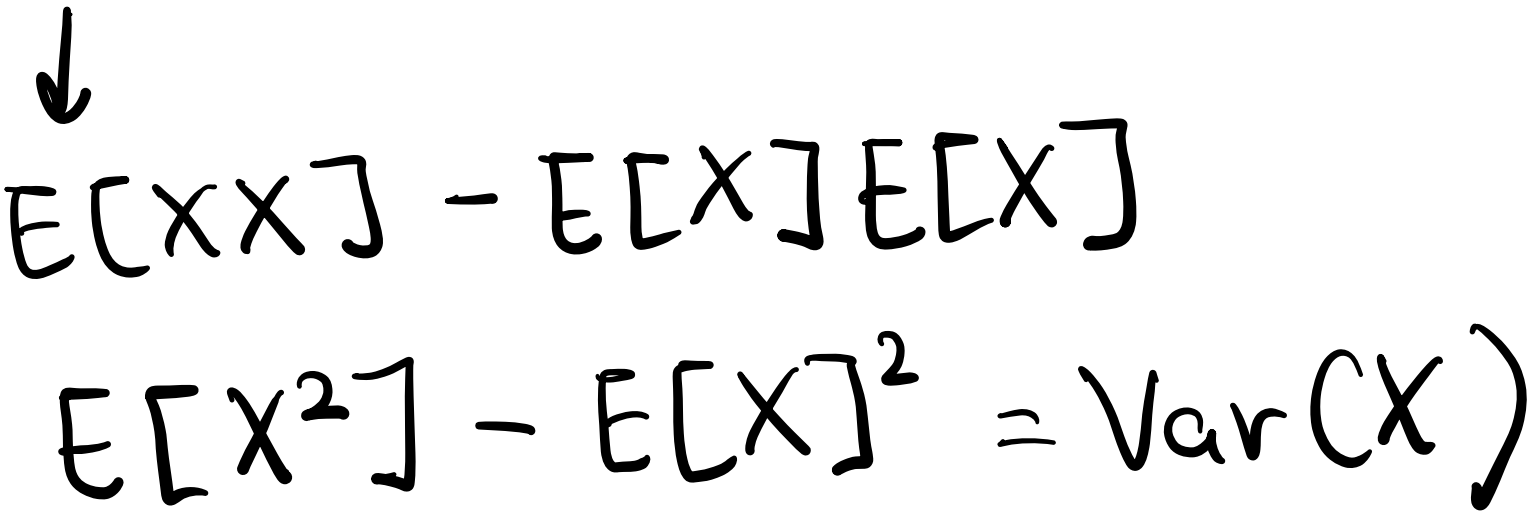
$$= \left(\sum_x x\,P(X=x)\right)\left(\sum_y y\,P(Y=y)\right)$$

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

NOTE: If two events are uncorrelated, this doesn't mean they're independent.

Covariance Relation to Variance

Cov(X, X) = Var(X)



Covariance is Symmetric

Cov(X, Y) = Cov(Y, X)

$E[XY] - E[X]E[Y]$

$E[YX] - E[Y]E[X]$

equivalent

What does this mean about **Var(X+Y)?**

$$Cov(X+Y, X+Y) = Cov(X,X) + Cov(X,Y) + Cov(Y,X) + Cov(Y,Y)$$

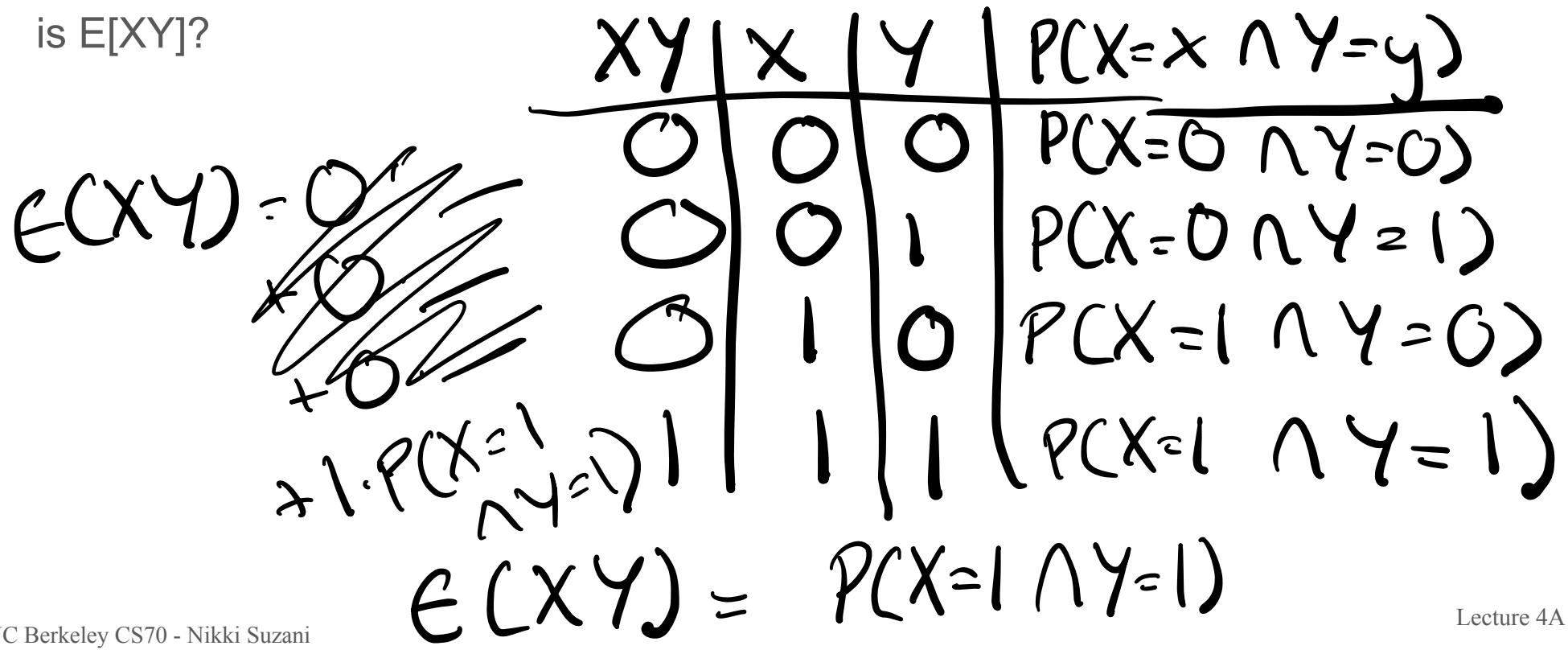
$$= Cov(X,X) + 2Cov(X,Y) + Cov(Y,Y)$$

$$= \mathrm{Var}(X) + 2\cancel{\mathrm{Cov}(X,Y)} + \mathrm{Var}(Y)$$

$$\overset{\text{if } X \text{ \& } Y \text{ are independent}}{=} \mathrm{Var}(X) + \mathrm{Var}(Y)$$

Covariance of Two indicator variables

Let X and Y be two not-necessarily-independent Bernoulli random variables. What is E[XY]?



Covariance Method with Identically Distributed Variables

Find Var($X_1 + X_2 + \dots X_n$) where all the X_i are identically distributed variables.

n options $\quad n-1$

$$\text{Cov}(X_1 + X_2 + \dots X_n,\ X_1 + X_2 + \dots X_n)$$

$$= n\,\text{Cov}(X_i, X_i) + n(n-1)\,\text{Cov}(X_i, X_j)$$

$$= n\left(E[X_i^2] - E[X_i]^2\right) + n(n-1)\left(E[X_i X_j] - E[X_i]E[X_j]\right)$$

$$= n\left(P(X_i=1) - (P(X_i=1))^2\right) + n(n-1)\left(P(X_i=1 \cap X_j=1) - (P(X_i=1))^2\right)$$

Example: Envelopes Problem

You have n envelopes and n letters. What is the variance of the number of letters that end up in the correct envelope?

Example: Envelopes Problem (cont.)

Envelopes Problem (another approach)

Let $X = X_1 + X_2 + \ldots + X_n$

$$E(X^2) = \sum_i E(X_i^2) + \sum_{i \neq j} E(X_i X_j).$$

Variance of the Hypergeometric Distribution

$\frac{10}{20}$

Suppose you have a box with **10** red balls and **10** ~~blue~~ green balls. You draw 10 times randomly **without replacement**. What is the variance of the number of red balls you pick?

X: # of red balls

$X = X_1 + X_2 + \dots X_{10}$

$X_i = \begin{cases} 1 & \text{if you draw a red ball on draw } i \\ 0 & \text{otherwise} \end{cases}$

$\text{Var}(X_1 + X_2 + \dots X_{10})$

$\text{Cov}(X_1 + \dots X_{10}, X_1 + \dots X_{10})$

$X_i \sim \text{Bernoulli}(\frac{1}{2})$

Variance of the Hypergeometric Distribution (cont.)

$$= 10\,Cov(X_i, X_i) + 10 \cdot 9 \cdot Cov(X_i, X_j)$$

$$10\,Var(X_i) + 10 \cdot 9 \cdot Cov(X_i, X_j)$$ prob.

$$10 \cdot \tfrac{1}{2}\left(1 - \tfrac{1}{2}\right) + 10 \cdot 9 \cdot \left(E[X_i X_j] - E[X_i]\,E[X_j]\right)$$

$$10 \cdot \tfrac{1}{2}\left(\tfrac{1}{2}\right) + 10 \cdot 9 \cdot \left(\frac{10 \times 9}{20 \times 19} - \frac{1}{4}\right)$$

$E[X_i] = 1/2$

$E[X_j] = 1/2$

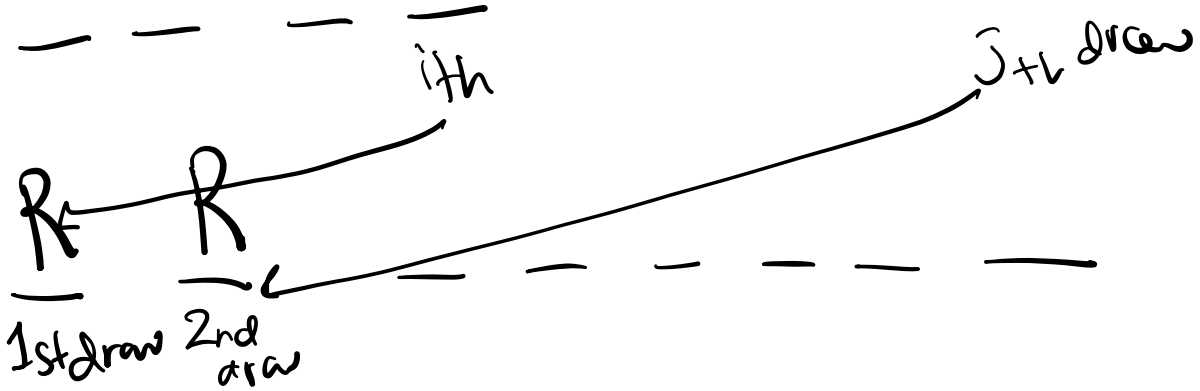
$E[X_i X_j] = P(X_i = 1 \cap X_j = 1)$

$= P(X_1 = 1 \cap X_2 = 1)$

$\frac{10}{20} \times \frac{9}{19}$

→ on some draw i we got red & on draw j we got red

$E[X_1 X_2]$



$$10 \cdot \frac{1}{4} + 10 \cdot 9 \cdot \left(-\frac{1}{76}\right)$$

$$\frac{10}{4} - \frac{90}{76}$$

$$\frac{190}{76} - \frac{90}{76} = \frac{100}{76}$$

Variance of the Hypergeometric (cont.)

What would the variance be if it was **with replacement?**

$X = X_1 + \dots X_{10}$ $\quad X_i \sim \text{Bernoulli}(1/2)$

independent

$\text{Binomial}(10, 1/2)$

$\text{Var}(X) = 10 \cdot 1/2 \cdot (1 - 1/2)$

$= \frac{190}{76}$

(Formally) Variance of Hypergeometric

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

Example of Hypergeometric: Tennis Balls

There are **three** ~~blue~~ black tennis balls, **four** red tennis balls, and **five** green tennis balls in a box. If you draw three times randomly without replacement, what is the variance of the number of red tennis balls that you draw?

$n: 3 \quad G: 4 \quad N: 3+4+5$

$\text{Hypergeometric}(12, 4, 3)$

$3 \cdot \left(\frac{4}{12}\right)\left(\frac{8}{12}\right)\left(\frac{12-3}{11}\right)$

Correlation (Formally)

It's hard to compare Cov(X, Y) and Cov(A, B) if A & B have different units than X & Y – numbers can vary pretty heavily based on units. **Correlation** helps us standardize units, in order to cross-compare covariances.

Definition 16.3 (Correlation). *Suppose X and Y are random variables with $\sigma(X) > 0$ and $\sigma(Y) > 0$. Then, the correlation of X and Y is defined as*

$$\text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma(X)\sigma(Y)}.$$

$$\sigma(X) = \sqrt{\text{Var}(X)}$$

$$\sigma(Y) = \sqrt{\text{Var}(Y)}$$

$$-1 \leq \text{Corr}(X,Y) \leq 1$$

(If time, derivation) Variance of the Poisson Distribution

Let X ~ Poisson(λ). What is Var(X)?

$$Var(X) = \lambda$$

(If time, derivation) Variance of the Geometric Distribution

Let X ~ Geometric(p). What is Var(X)?

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

$$E(X^2) = p + 4p(1-p) + 9p(1-p)^2 + \dots$$

$$-(1-p)E(X^2) = p(1-p) + 4p(1-p)^2 + 9p(1-p)^3 + \dots$$

$$pE(X^2) = 2E(X) - 1$$

$$pE(X^2) = 2\left(\frac{1}{p}\right) - 1$$

$$pE(X^2) = \frac{2}{p} - 1$$

$$E(X^2) = \frac{\left(\frac{2-p}{p}\right)}{p} = \frac{2-p}{p^2}$$

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

$$= \frac{2-p}{p^2} - \left(\frac{1}{p}\right)^2 = \frac{2-p-1}{p^2} = \frac{1-p}{p^2}$$

Recap

- Discussed the topics of variance, covariance, and correlation
  - Properties of Variance
  - Properties of Covariance
  - Properties of Correlation

- Discussed the variance of popular distributions and their derivations