

Correllation:

Want to deviate from the mean with me?

~~"True friendship comes when the silence
between two people is comfortable."~~

Their random variables are correlated

CS 109
Lecture 17
May 4th, 2016

Review

Did The Impossible Just Happen?



STATE OF HAWAII DEPARTMENT OF HEALTH
CERTIFICATE OF LIVE BIRTH
HLS Form 151 61 10641

Child's First Name (Type or print)		Middle Name		Last Name			
BARACK		HUSSEIN		OBAMA, II			
Sex	A. This Birth	4. If Twin or Triplet, Was Child Born	Month	Day	Year	6h. Hours /	
Male	Single <input checked="" type="checkbox"/> Twin <input type="checkbox"/> Triplet <input type="checkbox"/>	1st <input type="checkbox"/> 2nd <input type="checkbox"/> 3rd <input type="checkbox"/>	August	4	1961	7:24 P.M.	
Place of Birth: City, Town or Rural Location						6h. Island	
Honolulu						Oahu	
Name of Hospital or Institution (If not in hospital or institution, give street address)						6d. In Place of Birth Inside City or Town Limited?	
Kapiolani Maternity & Gynecological Hospital						If yes, give judicial district	
Usual Residence of Mother: City, Town or Rural Location						7a. County and State or Foreign Country	
Honolulu						Oahu	Honolulu, Hawaii
12. Street Address						7b. In Residence Inside City or Town Limited?	
6085 Kalanianaʻolaha Highway						If yes, give judicial district	
Mother's Mailing Address						7c. In Residence on a Farm or Plantation?	
						Yes <input type="checkbox"/> No <input checked="" type="checkbox"/>	
Full Name of Father		Middle Name		Last Name			
BARACK		HUSSEIN		OBAMA			
13. Age of Father	11. Birthplace (State, Territory or Foreign Country)	13b. Usual Occupation		13c. Kind of Education or Industry			
25	Kenya, East Africa	Student		University			
Full Mother Name of Mother		Middle Name		Last Name			
STANLEY		ANN		DUNHAM			
15. Age of Mother	11. Birthplace (State, Territory or Foreign Country)	13b. Type of Occupation Outside Home During Pregnancy		13c. Date Last Worked			
18	White, Kansas	None					
I certify that the above stated information is true and correct to the best of my knowledge.						14a. State of Signature	
14b. Signature of Person or Other Informant						Parent <input checked="" type="checkbox"/> 8-7-61	
14c. Signature of Registrar						8-8-61	
14d. Date Accepted by Local Reg						8-8-61	
AUG - 8 1961						14e. Date Accepted by Reg General	
						AUG - 8 1961	



BREAKING NEWS



Last year, 1% chance of winning the Republican primary

Will The Unlikely Happen?



Now, according to betting markets: 27.2% of being President

Bhutan's Happiness

- You want to know the true mean and variance of happiness in Buthan
 - But you can't ask everyone.
 - Randomly sample 200 people.
 - Your data looks like this:



Happiness = {72, 85, 79, 91, 68, ... , 71}

- The mean of all of those numbers is 83. Is that the true average happiness of Bhutanese people?

Sample Mean

- Consider n I.I.D. random samples X_1, X_2, \dots, X_n

- Sample mean:

$$\bar{X} = \sum_{i=1}^n \frac{X_i}{n}$$

- Sample variance:

$$S^2 = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n-1}$$

- They are both “unbiased” estimates

Variance of Sample Mean

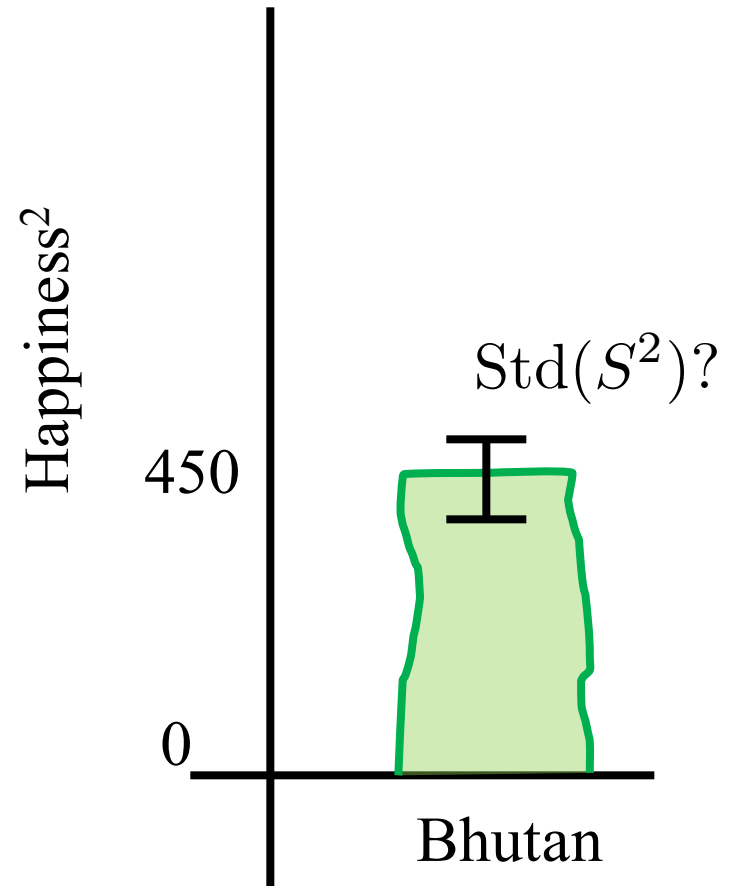
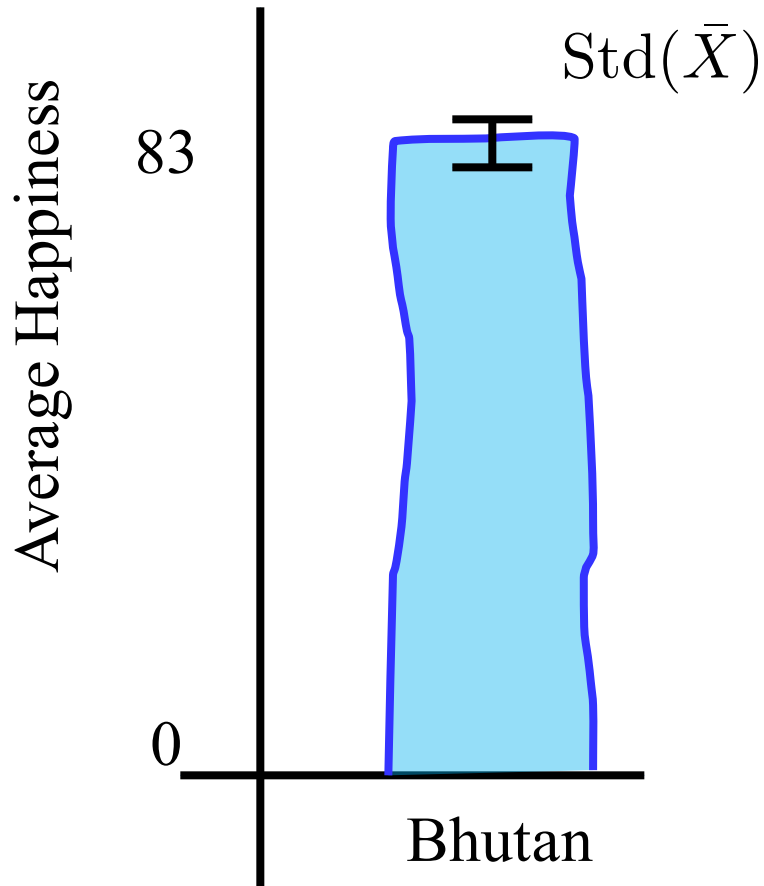
- Consider n I.I.D. random samples X_1, X_2, \dots, X_n
 - What is $\text{Var}(\bar{X})$?

$$\begin{aligned}\text{Var}(\bar{X}) &= \text{Var}\left(\sum_{i=1}^n \frac{X_i}{n}\right) = \left(\frac{1}{n}\right)^2 \text{Var}\left(\sum_{i=1}^n X_i\right) \\ &= \left(\frac{1}{n}\right)^2 \sum_{i=1}^n \text{Var}(X_i) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^n \sigma^2 = \left(\frac{1}{n}\right)^2 n\sigma^2 \\ &= \frac{\sigma^2}{n}\end{aligned}$$

Sampling

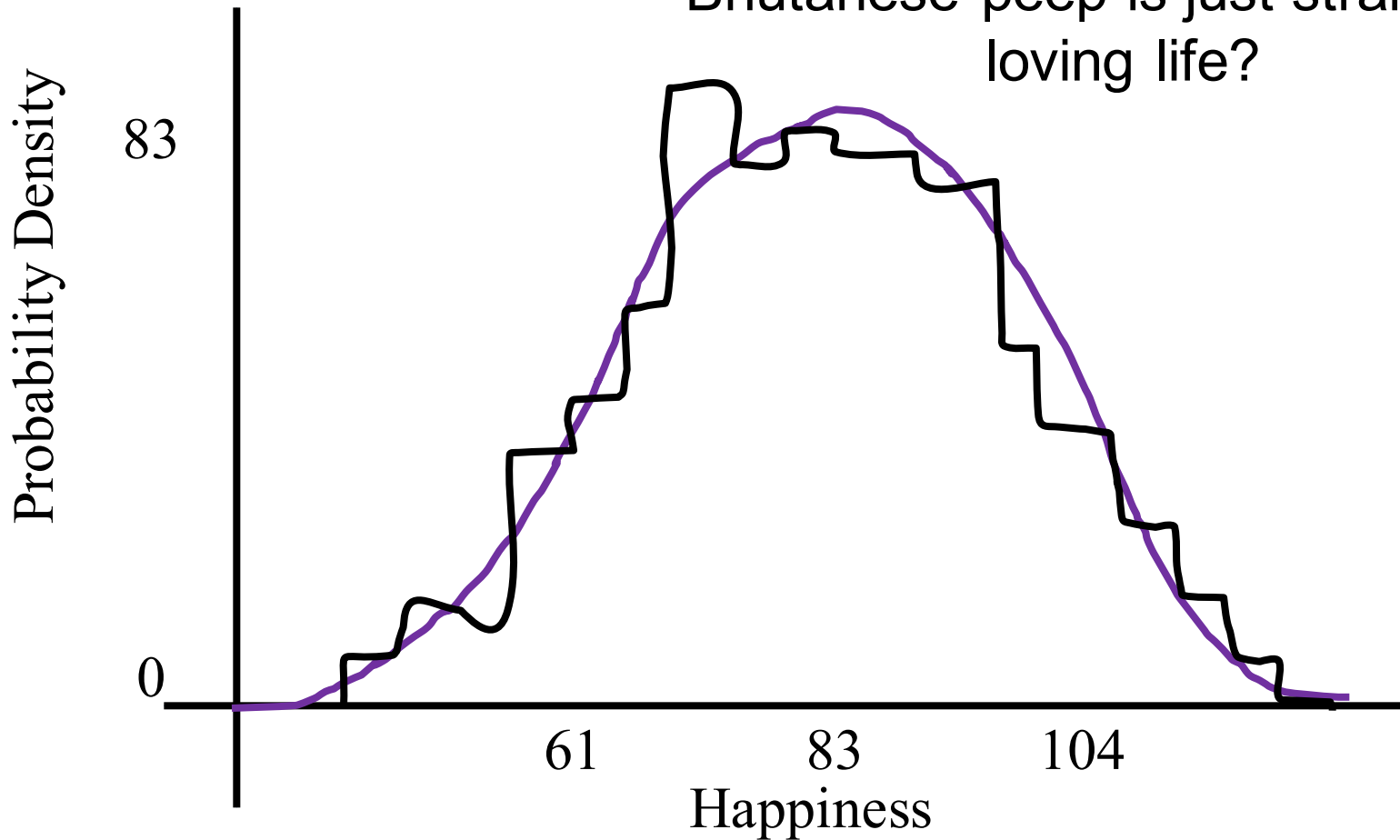
Sample mean: \bar{X}

Sample Variance: S^2



Happiness of Bhutan

What is the probability that a Bhutanese peep is just straight up loving life?



This ignores the variance of the sample mean
(and variance of the sample variance)

Case Study: Declaring Election

May 3

Indiana - 57 delegates



9% reporting

Delegates

Votes

Donald Trump (won)

45

54.2%
79,031

Ted Cruz

0

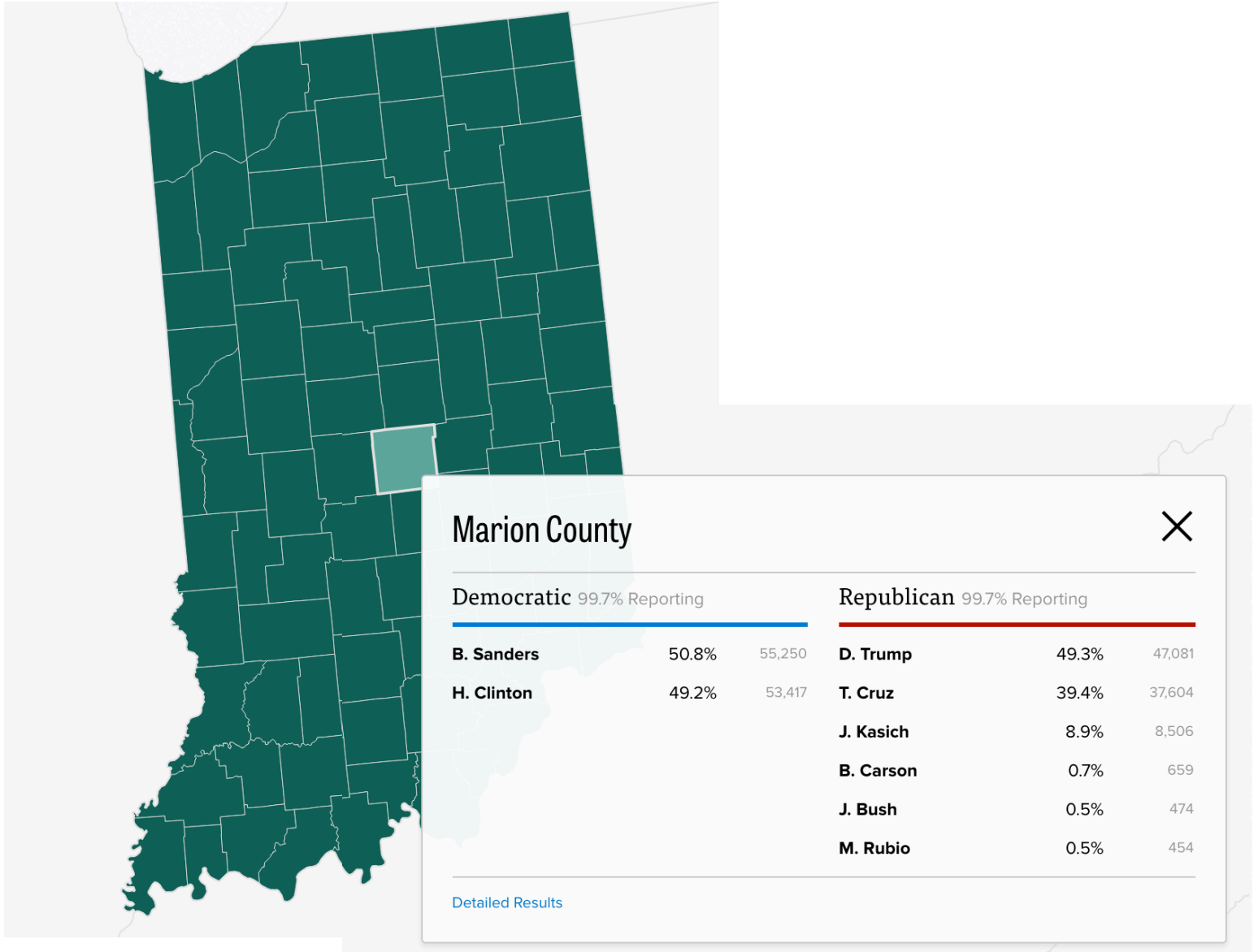
33.8%
49,360

John Kasich

0

9.1%
13,336

Indiana Counties



Case Study: Declaring Election

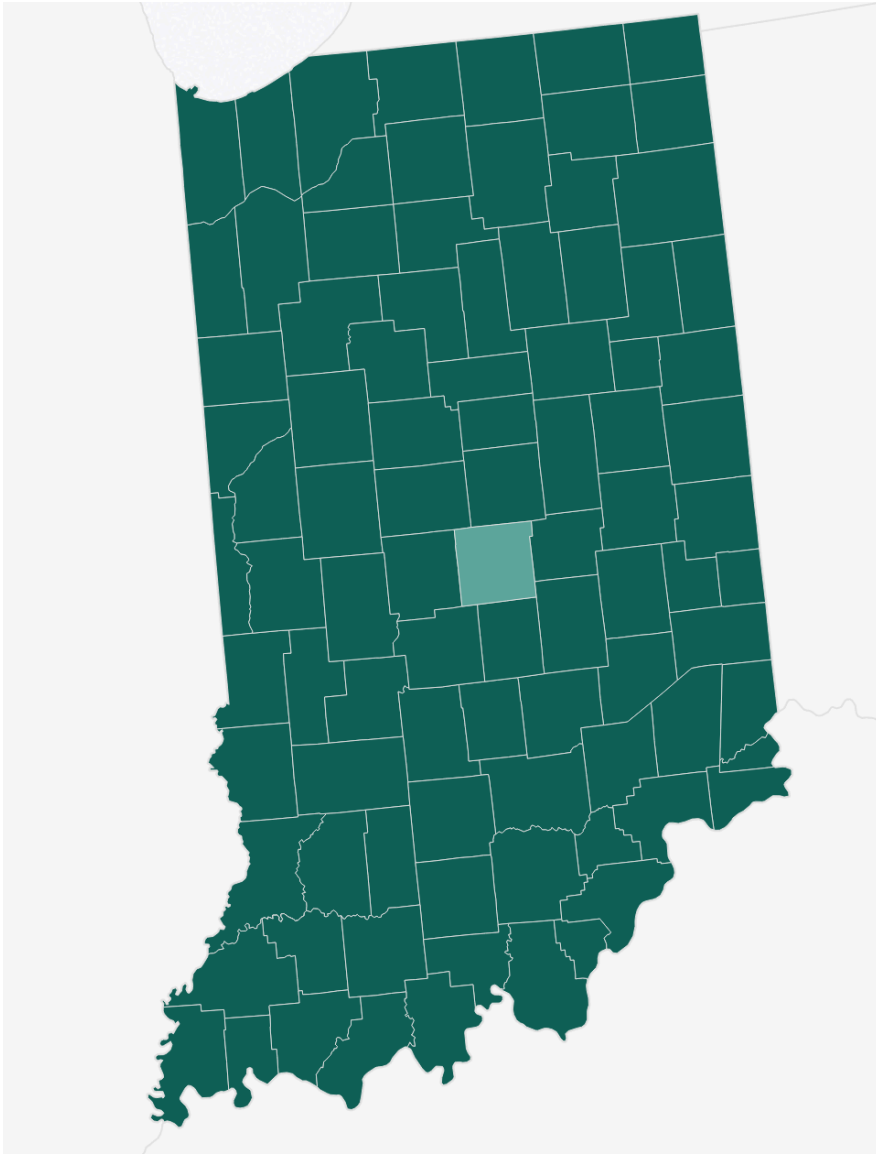
- Say X and Y are random variables:
 - X is the total number of votes that candidate 1 gets
 - Y is the total number of votes that candidate 2 gets
 - Calculate: $P(X > Y)$.
 - If that is high enough (say over 0.98), call the election.

$$P(X > Y) = P(X - Y > 0) = P(Y - X < 0)$$



Convolution of Y and $-X$

What is X?



Let X_i be a random variable that is the number of votes from county i

$$X = \sum_i X_i$$

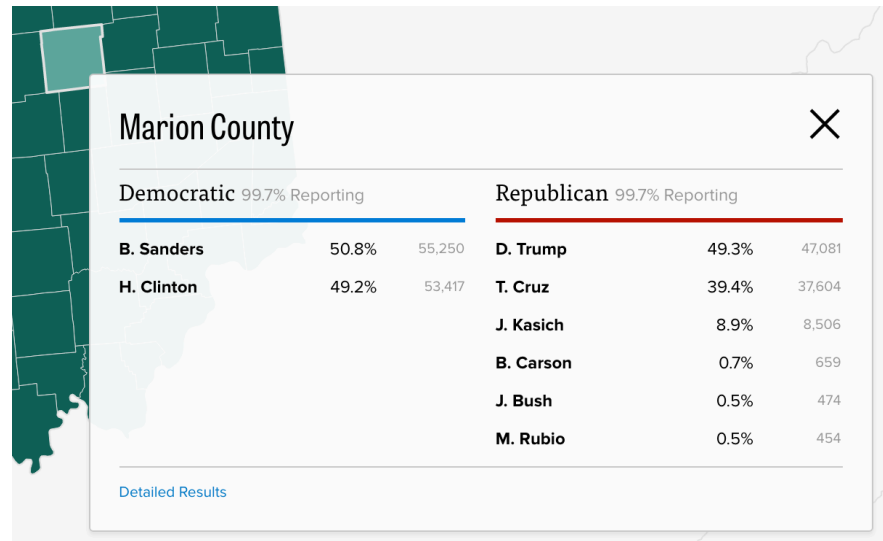
$$Y = \sum_i Y_i$$



ProTip: This means
for all i

What is X_i ?

Let X_i be a random variable that is the number of votes from county i



So far:

$$P(X > Y) = P(Y - X < 0)$$

$$X = \sum_i X_i$$

We don't know too much about X_i . We want it to convolve nicely.
Hopefully its normal.

What parameters to use for X_i ?

Let V_i be an indicator variable which is 1 if a voter in the county i votes for X : 9% of precincts reporting

Assume each reported voter in the county, Z_j , is an IID sample of V_i . Let n be the number of voters in the reporting precincts.

- Sample mean:

$$\bar{Z}_i = \sum_{j=1}^n \frac{Z_j}{n}$$

- Make sure we have enough:

$$\text{Var}(\bar{Z}_i) \quad \dots \text{Make sure the county is worth including}$$

$$P(V_i) = E[V_i] = \bar{Z}_i$$

Like estimating happiness in Bhutan



What parameters to use for X_i ?

We can estimate the probability that a voter in county i votes for a candidate

$$P(V_i) = E[V_i] = \bar{Z}_i$$

There are m_i expected voters in the county

Large n . And reasonable p

Binomial

$$X_i \sim N(m_i \bar{Z}_i, m_i \bar{Z}_i (1 - \bar{Z}_i))$$

Putting it all together

X, Y are the total number of votes that candidates gets

$$P(X > Y) = P(Y - X < 0)$$

Let X_i be a random variable that is the number of votes from county i

$$X = \sum_i X_i \quad Y = \sum_i Y_i$$

Assume voters from reporting precincts make up a sample of an indicator variable:

$$X_i \sim N(m_i \bar{Z}_i, m_i \bar{Z}_i (1 - \bar{Z}_i))$$

$$X \sim N \left(\sum_i m_i \bar{Z}_i, \sum_i m_i \bar{Z}_i (1 - \bar{Z}_i) \right)$$

$$Y \sim N \left(\sum_i m_i \bar{W}_i, \sum_i m_i \bar{W}_i (1 - \bar{W}_i) \right)$$

Bringing it Home Like Were E.T.

$$X \sim N \left(\sum_i m_i \bar{Z}_i, \sum_i m_i \bar{Z}_i (1 - \bar{Z}_i) \right)$$
$$Y \sim N \left(\sum_i m_i \bar{W}_i, \sum_i m_i \bar{W}_i (1 - \bar{W}_i) \right)$$

Now let's calculate $P(X > Y)$

More convolution...

$$Y - X \sim N \left(\sum_i m_i \bar{W}_i - \sum_i m_i \bar{Z}_i, \sum_i m_i \bar{W}_i (1 - \bar{Z}_i) + \sum_i m_i \bar{Z}_i (1 - \bar{W}_i) \right)$$

By CDF of normal

$$P(X > Y) = \phi \left(\frac{0 - \sum_i m_i \bar{W}_i - \sum_i m_i \bar{Z}_i}{\sqrt{\sum_i m_i \bar{W}_i (1 - \bar{Z}_i) + \sum_i m_i \bar{Z}_i (1 - \bar{W}_i)}} \right)$$

Case Study: Declaring Election

May 3

Indiana - 57 delegates



9% reporting

Delegates

Votes

Donald Trump (won)

45

54.2%
79,031

Ted Cruz

0

33.8%
49,360

John Kasich

0

9.1%
13,336

Great Question



Missing at random

Review

The Dance of the Covariance

- Say X and Y are arbitrary random variables
- Covariance of X and Y :

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

x	y	$(x - E[X])(y - E[Y])p(x,y)$
Above mean	Above mean	Positive
Bellow mean	Bellow mean	Positive
Bellow mean	Above mean	Negative
Above mean	Bellow mean	Negative

The Dance of the Covariance

- Say X and Y are arbitrary random variables

- Covariance of X and Y :

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

- Equivalently:

$$\text{Cov}(X, Y) = E[XY - E[X]Y - XE[Y] + E[Y]E[X]]$$

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

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

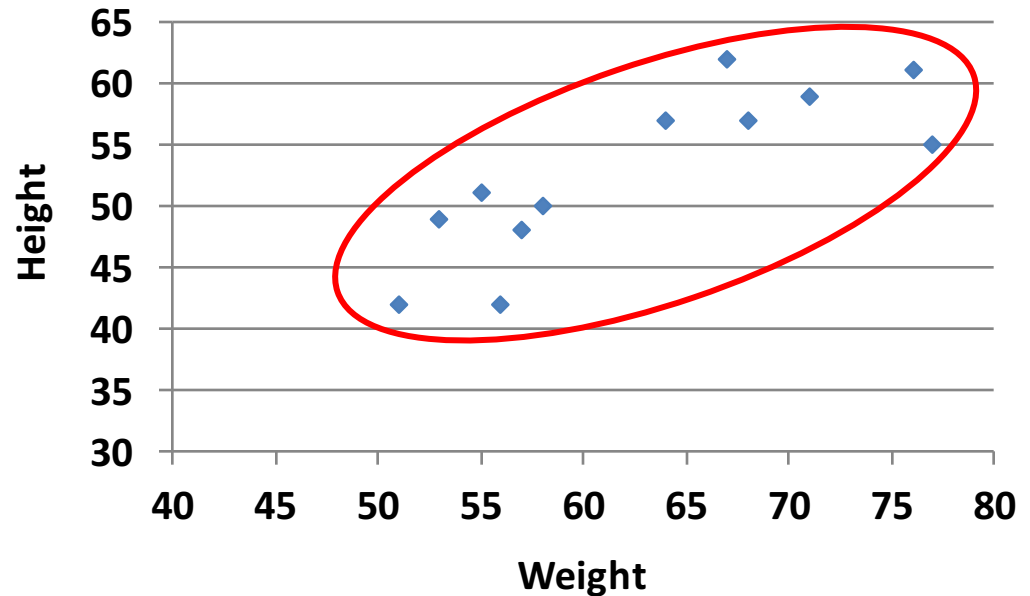
- X and Y independent, $E[XY] = E[X]E[Y] \rightarrow \text{Cov}(X, Y) = 0$
- But $\text{Cov}(X, Y) = 0$ does **not** imply X and Y independent!

Another Example of Covariance

- Consider the following data:

Weight	Height	Weight * Height
64	57	3648
71	59	4189
53	49	2597
67	62	4154
55	51	2805
58	50	2900
77	55	4235
57	48	2736
56	42	2352
51	42	2142
76	61	4636
68	57	3876

$$\begin{aligned} E[W] &= 62.75 & E[H] &= 52.75 & E[W*H] &= 3355.83 \\ = 62.75 & & = 52.75 & & = 3355.83 & \end{aligned}$$



$$\begin{aligned} \text{Cov}(W, H) &= E[W*H] - E[W]E[H] \\ &= 3355.83 - (62.75)(52.75) \\ &= 45.77 \end{aligned}$$

End Review

Correlation

Viva La Correlación

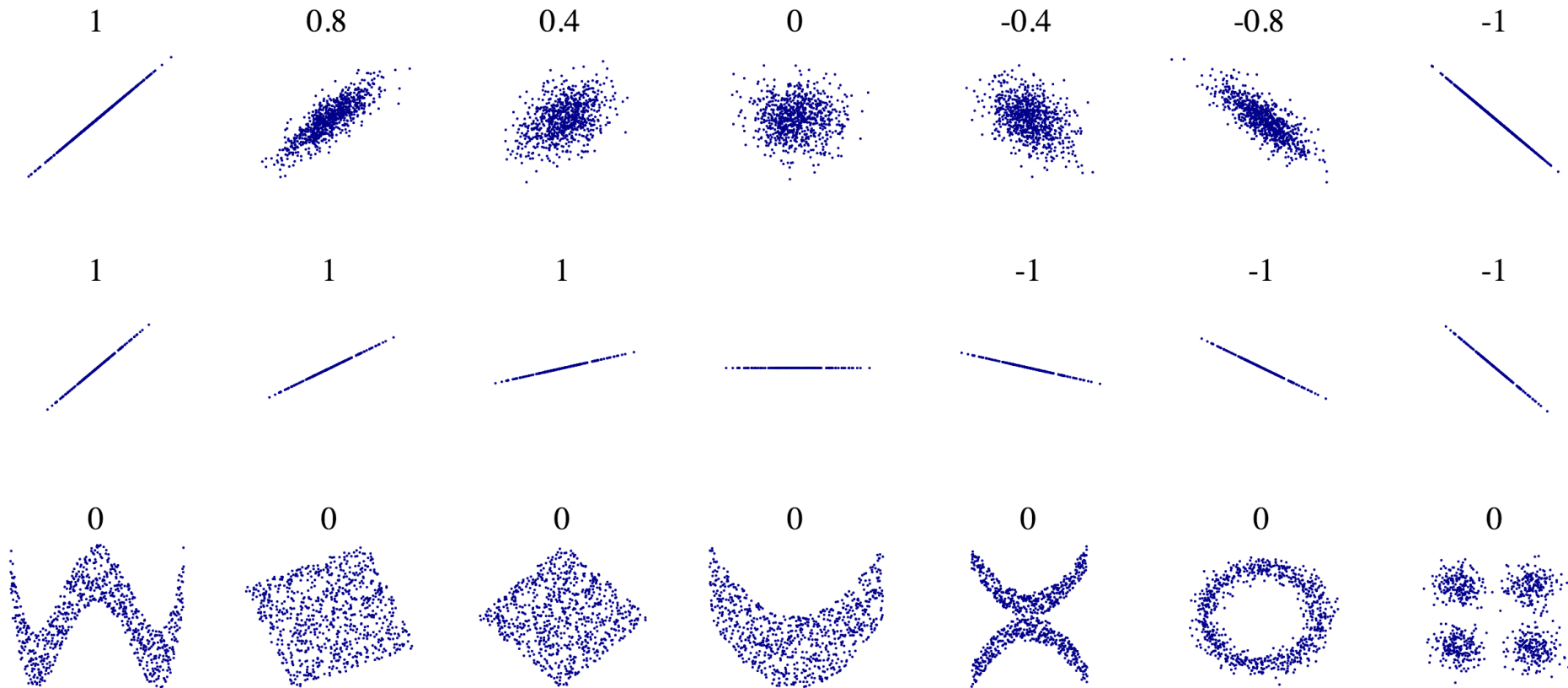
- Say X and Y are arbitrary random variables

- Correlation of X and Y , denoted $\rho(X, Y)$:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

- Note: $-1 \leq \rho(X, Y) \leq 1$
- Correlation measures linearity between X and Y

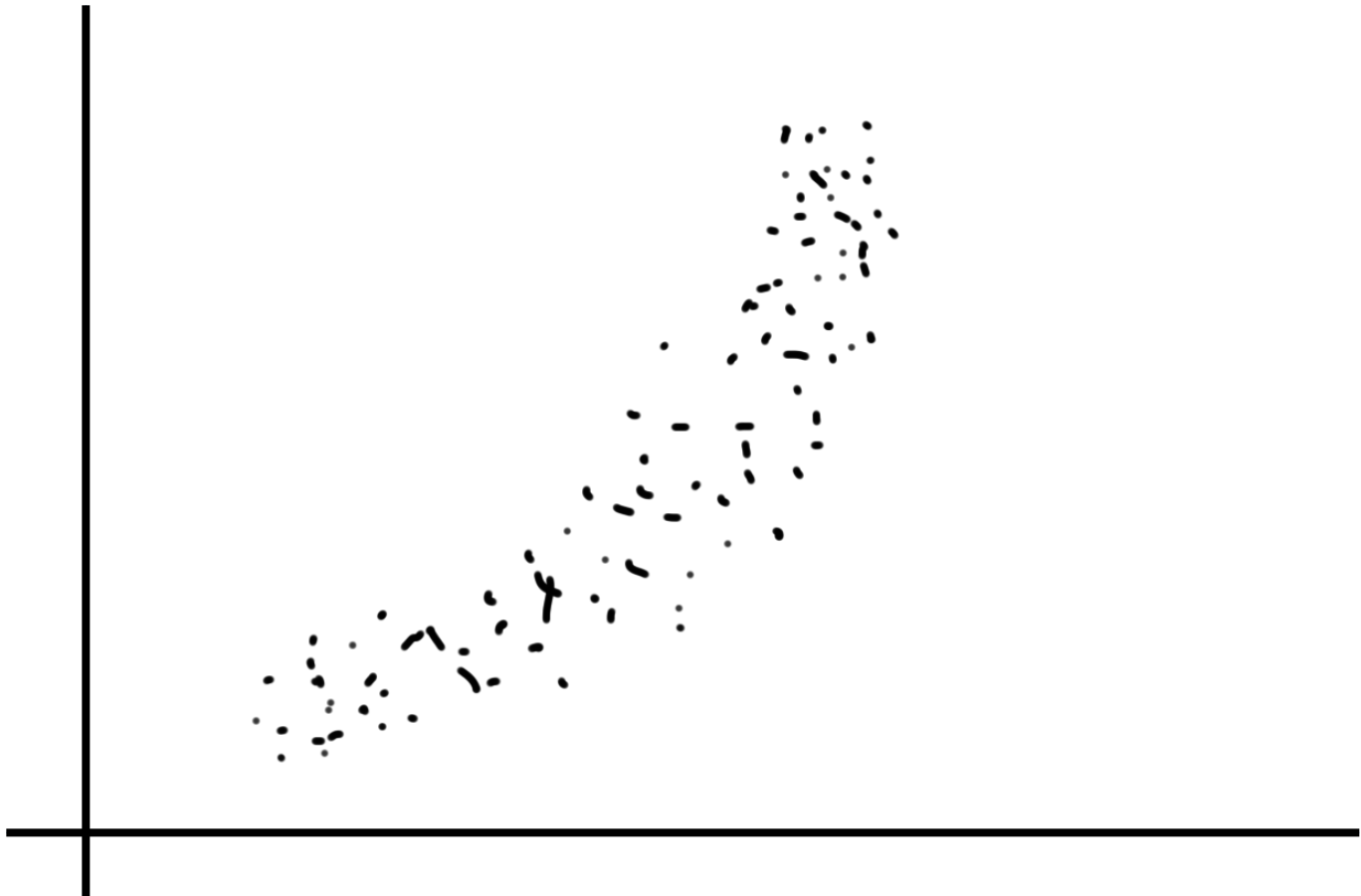
Pearson Correlation



*If someone just says “Correlation” they mean Pearson Correlation

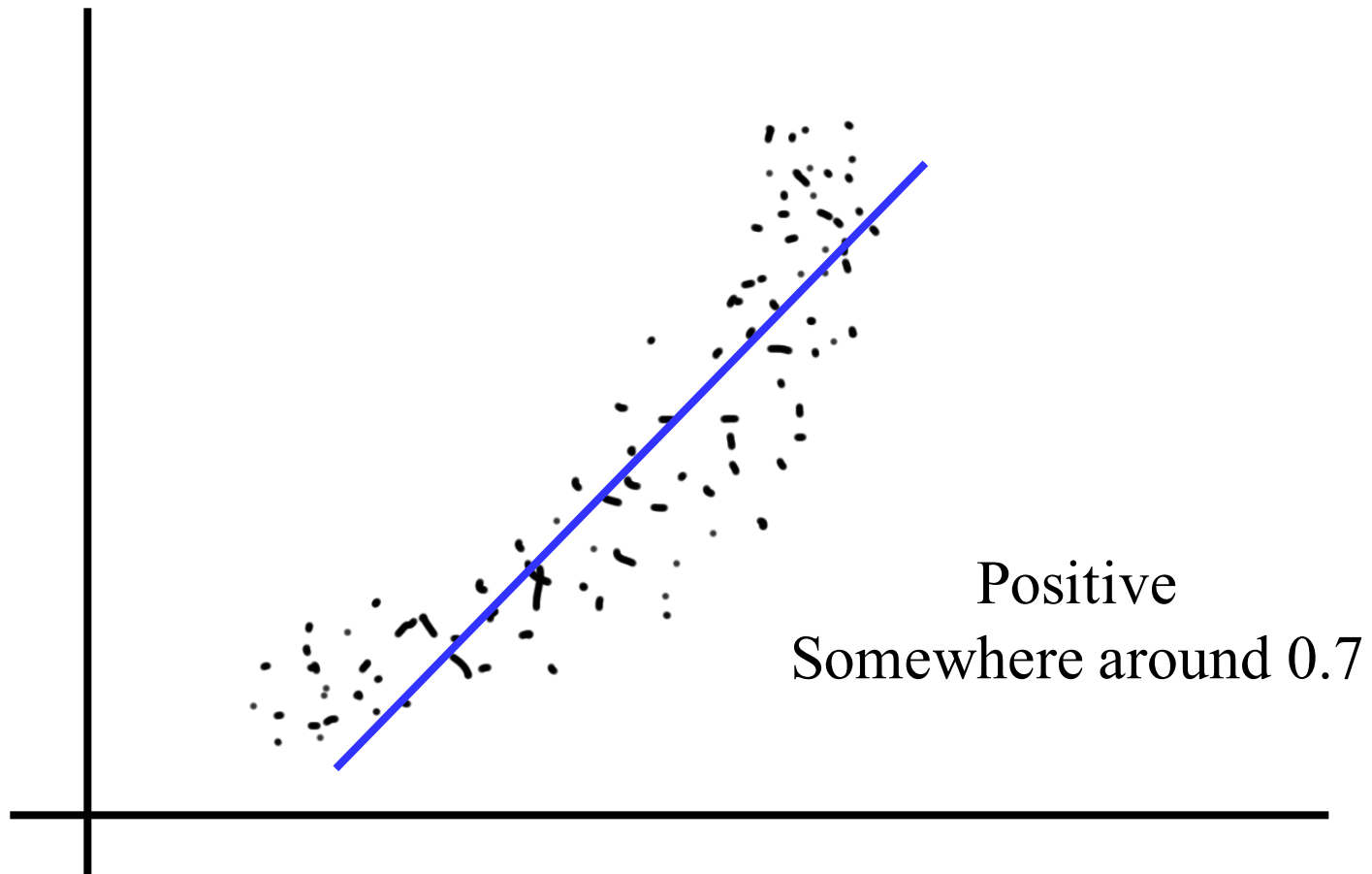
Pearson Correlation

Socrative: (a) positive, (b) negative, (c) zero



Pearson Correlation

Socrative: (a) positive, (b) negative, (c) zero



Viva La Correlación

- Say X and Y are arbitrary random variables

- Correlation of X and Y , denoted $\rho(X, Y)$:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

- Note: $-1 \leq \rho(X, Y) \leq 1$
- Correlation measures linearity between X and Y
- $\rho(X, Y) = 1 \quad \Rightarrow \quad Y = aX + b \quad \text{where } a = \sigma_y/\sigma_x$
- $\rho(X, Y) = -1 \quad \Rightarrow \quad Y = aX + b \quad \text{where } a = -\sigma_y/\sigma_x$
- $\rho(X, Y) = 0 \quad \Rightarrow \quad \text{absence of linear relationship}$
 - But, X and Y can still be related in some other way!
- If $\rho(X, Y) = 0$, we say X and Y are “uncorrelated”
 - Note: Independence implies uncorrelated, but **not** vice versa!

Can't Get Enough of that Multinomial

- Multinomial distribution

- n independent trials of experiment performed
- Each trials results in one of m outcomes, with respective probabilities: p_1, p_2, \dots, p_m where $\sum_{i=1}^m p_i = 1$
- $X_i =$ number of trials with outcome i

$$P(X_1 = c_1, X_2 = c_2, \dots, X_m = c_m) = \binom{n}{c_1, c_2, \dots, c_m} p_1^{c_1} p_2^{c_2} \dots p_m^{c_m}$$

- E.g., Rolling 6-sided die multiple times and counting how many of each value $\{1, 2, 3, 4, 5, 6\}$ we get
- Would expect that X_i are negatively correlated
- Let's see... when $i \neq j$, what is $\text{Cov}(X_i, X_j)$?

Covariance and the Multinomial

- Computing $\text{Cov}(X_i, X_j)$

- Indicator $I_i(k) = 1$ if trial k has outcome i , 0 otherwise

$$E[I_i(k)] = p_i \qquad X_i = \sum_{k=1}^n I_i(k) \qquad X_j = \sum_{k=1}^n I_j(k)$$

- $\text{Cov}(X_i, X_j) = \sum_{a=1}^n \sum_{b=1}^n \text{Cov}(I_i(b), I_j(a))$

- When $a \neq b$, trial a and b independent: $\text{Cov}(I_i(b), I_j(a)) = 0$
- When $a = b$: $\text{Cov}(I_i(b), I_j(a)) = E[I_i(a)I_j(a)] - E[I_i(a)]E[I_j(a)]$
- Since trial a cannot have outcome i and j : $E[I_i(a)I_j(a)] = 0$

$$\begin{aligned} \text{Cov}(X_i, X_j) &= \sum_{a=b=1}^n \text{Cov}(I_i(b), I_j(a)) = \sum_{a=1}^n (-E[I_i(a)]E[I_j(a)]) \\ &= \sum_{a=1}^n (-p_i p_j) = -n p_i p_j \quad \Rightarrow X_i \text{ and } X_j \text{ negatively correlated} \end{aligned}$$

Multinomials All Around

- Multinomial distributions:
 - Count of strings hashed into buckets in hash table
 - Number of server requests across machines in cluster
 - Distribution of words/tokens in an email
 - Etc.
- When m (# outcomes) is large, p_i is small
 - For equally likely outcomes: $p_i = 1/m$

$$\text{Cov}(X_i, X_j) = -np_i p_j = -\frac{n}{m^2}$$

- Large $m \Rightarrow X_i$ and X_j very mildly negatively correlated
- Poisson paradigm applicable

Break

Conditional Expectation

- X and Y are jointly discrete random variables

- Recall conditional PMF of X given $Y = y$:

$$p_{X|Y}(x | y) = P(X = x | Y = y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}$$

- Define conditional expectation of X given $Y = y$:

$$E[X | Y = y] = \sum_x x P(X = x | Y = y) = \sum_x x p_{X|Y}(x | y)$$

- Analogously, jointly continuous random variables:

$$f_{X|Y}(x | y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} \quad E[X | Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x | y) dx$$

Rolling Dice

- Roll two 6-sided dice D_1 and D_2
 - $X = \text{value of } D_1 + D_2$ $Y = \text{value of } D_2$
 - What is $E[X | Y = 6]$?

$$\begin{aligned} E[X | Y = 6] &= \sum_x xP(X = x | Y = 6) \\ &= \left(\frac{1}{6}\right)(7 + 8 + 9 + 10 + 11 + 12) = \frac{57}{6} = 9.5 \end{aligned}$$

- Intuitively makes sense: $6 + E[\text{value of } D_1] = 6 + 3.5$

Mystery Distribution

- X and Y are independent random variables
 - $X \sim \text{Bin}(n, p)$ $Y \sim \text{Bin}(n, p)$
 - What is $E[X \mid X + Y = m]$, where $m \leq n$?
 - Start by computing $P(X = k \mid X + Y = m)$:

$$\begin{aligned}
 P(X = k \mid X + Y = m) &= \frac{P(X = k, X + Y = m)}{P(X + Y = m)} = \frac{P(X = k, Y = m - k)}{P(X + Y = m)} = \frac{P(X = k)P(Y = m - k)}{P(X + Y = m)} \\
 &= \frac{\binom{n}{k} p^k (1-p)^{n-k} \cdot \binom{n}{m-k} p^{m-k} (1-p)^{n-(m-k)}}{\binom{2n}{m} p^m (1-p)^{2n-m}} = \frac{\binom{n}{k} \cdot \binom{n}{m-k}}{\binom{2n}{m}}
 \end{aligned}$$

- Hypergeometric: $(X \mid X + Y = m) \sim \text{HypG}(m, 2n, n)$
- $E[X \mid X + Y = m] = nm/2n = m/2$ # total draws total balls white balls

White ball: #X heads. Black ball: #Y heads

Paz Fuera A-Pueblo

*That's (literally) Spanish for:
Peace out A-Town*