Indicators: Now With Pair-wise Flavor!

- Recall I_i is indicator variable for event A_i when:

$$I_i = \begin{cases} 1 & \text{if } A_i \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

  - Let X = # of events that occur: $X = \sum_{i=1}^{n} I_i$

$$E[X] = E\left[\sum_{i=1}^{n} I_i\right] = \sum_{i=1}^{n} E[I_i] = \sum_{i=1}^{n} P(A_i)$$

- Now consider pair of events $A_i\ A_j$ occurring
  - $I_i\ I_j$ = 1 if both events A_i and A_j occur, 0 otherwise
  - Number of pairs of events that occur is $\binom{X}{2} = \sum_{i<j} I_i\ I_j$

From Event Pairs to Variance

- Expected number of pairs of events:

$$E\left[\binom{X}{2}\right] = E\left[\sum_{i<j} I_i I_j\right] = \sum_{i<j} E[I_i I_j] = \sum_{i<j} P(A_i A_j)$$

$$E\left[\frac{X(X-1)}{2}\right] = \frac{1}{2}(E[X^2] - E[X]) = \sum_{i<j} P(A_i A_j)$$

$$E[X^2] - E[X] = 2\sum_{i<j} P(A_i A_j) \;\Rightarrow\; E[X^2] = 2\sum_{i<j} P(A_i A_j) + E[X]$$

- Recall: $\text{Var}(X) = E[X^2] - (E[X])^2$

$$\text{Var}(X) = 2\sum_{i<j} P(A_i A_j) + E[X] - (E[X])^2$$

$$= 2\sum_{i<j} P(A_i A_j) + \sum_{i=1}^{n} P(A_i) - \left(\sum_{i=1}^{n} P(A_i)\right)^2$$

Let's Try It With the Binomial

- X ~ Bin(n, p) $\qquad E[X] = \sum_{i=1}^{n} P(A_i) = np$
  - Each trial: X_i ~ Ber(p) $\qquad E[X_i] = p$
  - Let event A_i = trial *i* is success (i.e., $X_i = 1$)

$$E\left[\binom{X}{2}\right] = \sum_{i<j} E[X_i X_j] = \sum_{i<j} P(A_i A_j) = \sum_{i<j} p^2 = \binom{n}{2} p^2$$

$$E\left[\frac{X(X-1)}{2}\right] = \frac{1}{2}(E[X^2] - E[X]) = \frac{n(n-1)}{2} p^2$$

$$\operatorname{Var}(X) = E[X^2] - (E[X])^2 = (E[X^2] - E[X]) + E[X] - (E[X])^2$$

$$= n(n-1)p^2 + np - (np)^2 = n^2p^2 - np^2 + np - n^2p^2$$

$$= np(1-p)$$

Computer Cluster Utilization

- Computer cluster with *k* servers
  - Requests independently go to server *i* with probability p_i
  - Let event A_i = server *i* receives no requests
  - X = # of events $A_1, A_2, \ldots A_k$ that occur
  - Y = # servers that receive ≥ 1 request = k – X
  - E[Y] after first *n* requests?
  - Since requests independent: $P(A_i) = (1 - p_i)^n$

$$E[X] = \sum_{i=1}^{k} P(A_i) = \sum_{i=1}^{k} (1 - p_i)^n$$

$$E[Y] = k - E[X] = k - \sum_{i=1}^{k} (1 - p_i)^n$$

when $p_i = \frac{1}{k}$ for $1 \le i \le k$, $E[Y] = k - \sum_{i=1}^{k} (1 - \frac{1}{k})^n = k\left(1 - (1 - \frac{1}{k})^n\right)$

Computer Cluster Utilization (cont.)

- Computer cluster with *k* servers
  - Requests independently go to server *i* with probability p_i
  - Let event A_i = server *i* receives no requests
  - X = # of events A_1, A_2, … A_k that occur
  - Y = # servers that receive ≥ 1 request = *k* – X
  - Var(Y) after first *n* requests? (= $(-1)^2$ Var(X) = Var(X))
  - Independent requests: $P(A_i A_j) = (1 - p_i - p_j)^n, \quad i \neq j$

$$E[X(X-1)] = E[X^2] - E[X] = 2\sum_{i<j} P(A_i A_j) = 2\sum_{i<j} (1 - p_i - p_j)^n$$

$$\mathrm{Var}(X) = 2\sum_{i<j} (1 - p_i - p_j)^n + E[X] - (E[X])^2 \qquad E[X] = \sum_{i=1}^{k} (1 - p_i)^n$$

$$= 2\sum_{i<j} (1 - p_i - p_j)^n + \sum_{i=1}^{k} (1 - p_i)^n - \left(\sum_{i=1}^{k} (1 - p_i)^n \right)^2 = \mathrm{Var}(Y)$$

Computer Cluster = Coupon Collecting

- Computer cluster with k servers
  - Requests independently go to server i with probability p_i
  - Let event A_i = server i receives no requests
  - X = # of events $A_1, A_2, \ldots A_k$ that occur
  - Y = # servers that receive ≥ 1 request = $k - X$
- This is really another "Coupon Collector" problem
  - Each server is a "coupon type"
  - Request to server = collecting a coupon of that type
- Hash table version
  - Each server is a bucket in table
  - Request to server = string gets hashed to that bucket

Product of Expectations

- Say X and Y are <u>independent</u> random variables, and $g(\bullet)$ and $h(\bullet)$ are real-valued functions

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)]$$

  - Proof:

$$E[g(X)h(Y)] = \int_{y=-\infty}^{\infty} \int_{x=-\infty}^{\infty} g(x)h(y)\, f_{X,Y}(x,y)\, dx\, dy$$

$$= \int_{y=-\infty}^{\infty} \int_{x=-\infty}^{\infty} g(x)h(y)\, f_X(x)\, f_Y(y)\, dx\, dy$$

$$= \int_{x=-\infty}^{\infty} g(x)\, f_X(x)\, dx \cdot \int_{y=-\infty}^{\infty} h(y)\, f_Y(y)\, dy$$

$$= E[g(X)]E[h(Y)]$$

The Dance of the Covariance

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

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

- Equivalently:

$$\begin{aligned}\mathrm{Cov}(X,Y) &= E[XY - E[X]Y - XE[Y] + E[Y]E[X]] \\ &= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y]\end{aligned}$$

  - X and Y independent, E[XY] = E[X]E[Y] → Cov(X,Y) = 0
  - But Cov(X,Y) = 0 does **not** imply X and Y independent!

Dependence and Covariance

- X and Y are random variables with PMF:

Y \ X	-1	0	1	$p_Y(y)$
0	1/3	0	1/3	2/3
1	0	1/3	0	1/3
$p_X(x)$	1/3	1/3	1/3	1

$$Y = \begin{cases} 0 & \text{if } X \neq 0 \\ 1 & \text{otherwise} \end{cases}$$

  - E[X] = -1(1/3) + 0(1/3) + 1(1/3) = 0
  - E[Y] = 0(2/3) + 1(1/3) = 1/3
  - Since XY = 0, E[XY] = 0
  - Cov(X, Y) = E[XY] – E[X]E[Y] = 0 – 0 = 0
- But, X and Y are clearly dependent!

Example of Covariance

- Consider rolling a 6-sided die
  - Let indicator variable X = 1 if roll is 1, 2, 3, or 4
  - Let indicator variable Y = 1 if roll is 3, 4, 5, or 6
- What is Cov(X, Y)?
  - E[X] = 2/3 and E[Y] = 2/3
  - E[XY] $= \sum_x \sum_y xy\ p(x, y)$

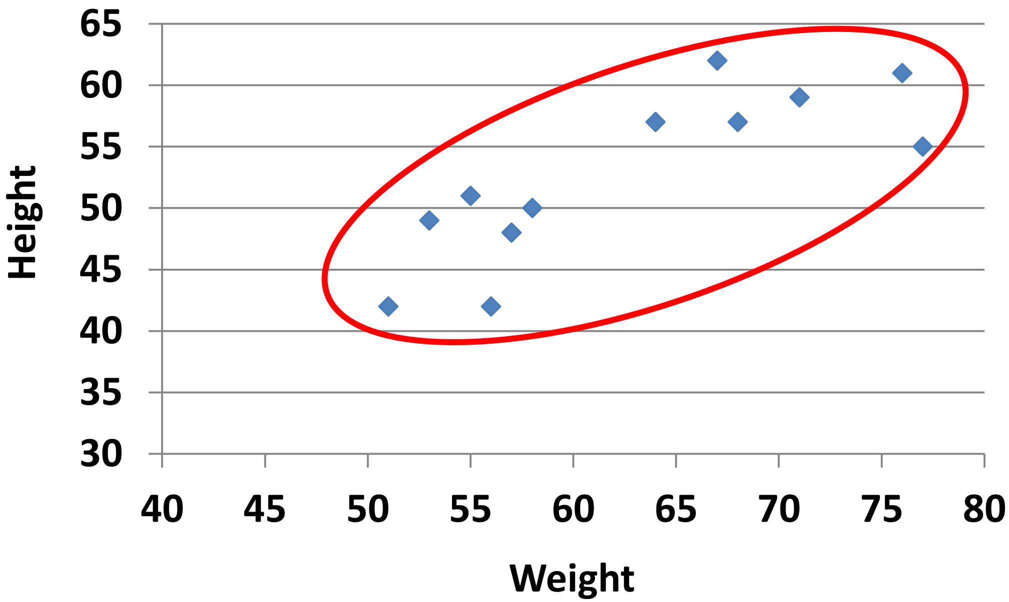
    = (0 * 0) + (0 * 1/3) + (0 * 1/3) + (1 * 1/3) = 1/3
  - Cov(X, Y) = E[XY] – E[X]E[Y] = 1/3 – 4/9 = -1/9
  - Consider: P(X = 1) = 2/3 and P(X = 1 | Y = 1) = 1/2
    - Observing Y = 1 makes X = 1 *less* likely

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

E[W] = 62.75 **E[H] = 52.75** **E[W*H] = 3355.83**



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

Properties of Covariance

- Say X and Y are arbitrary random variables
  - $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$
  - $\mathrm{Cov}(X,X) = E[X^2] - E[X]E[X] = \mathrm{Var}(X)$
  - $\mathrm{Cov}(aX+b,Y) = a\mathrm{Cov}(X,Y)$
- Covariance of sums of random variables
  - $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ are random variables
  - $\mathrm{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{Cov}(X_i, Y_j)$

Variance of Sum of Variables

- $\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2\sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathrm{Cov}(X_i, X_j)$
  - Proof:

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \mathrm{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{Cov}(X_i, X_j)$$

Note: $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$

$$= \sum_{i=1}^{n} \mathrm{Var}(X_i) + \sum_{i=1}^{n} \sum_{j=1,\, j\neq i}^{n} \mathrm{Cov}(X_i, X_j)$$

By symmetry: $\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i)$

$$= \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2\sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathrm{Cov}(X_i, X_j)$$

  - If all X_i and X_j independent ($i \neq j$): $\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i)$

Hola Compadre: La Distribución Binomial

- Let $Y \sim \mathrm{Bin}(n, p)$
  - *n* independent trials
  - Let $X_i = 1$ if *i*-th trial is "success", 0 otherwise
  - $X_i \sim \mathrm{Ber}(p)$ $\quad E[X_i] = p$
  - $\mathrm{Var}(Y) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \ldots + \mathrm{Var}(X_n)$
  - $\mathrm{Var}(X_i) = E[X_i^2] - (E[X_i])^2$

    $= E[X_i] - (E[X_i])^2 \qquad$ since $X_i^2 = X_i$

    $= p - p^2 = p(1 - p)$
  - $\mathrm{Var}(Y) = n\mathrm{Var}(X_i) = np(1 - p)$

Variance of Sample Mean

- Consider *n* I.I.D. random variables $X_1, X_2, \dots X_n$
  - X_i have distribution *F* with $E[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$
  - We call sequence of X_i a **sample** from distribution *F*
  - Recall sample mean: $\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$ where $E[\bar{X}] = \mu$
  - What is $\mathrm{Var}(\bar{X})$?

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\sum_{i=1}^{n} \frac{X_i}{n}\right) = \left(\frac{1}{n}\right)^2 \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right)$$

$$= \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \mathrm{Var}(X_i) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \sigma^2 = \left(\frac{1}{n}\right)^2 n\sigma^2$$

$$= \frac{\sigma^2}{n}$$

Sample Variance

- Consider *n* I.I.D. random variables $X_1, X_2, \ldots X_n$
  - X_i have distribution *F* with $E[X_i] = \mu$ and $Var(X_i) = \sigma^2$
  - We call sequence of X_i a **sample** from distribution *F*
  - Recall sample mean: $\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$ where $E[\bar{X}] = \mu$
  - Sample deviation: $\bar{X} - X_i$ for i = 1, 2, ..., n
  - Sample variance: $S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$

  - What is $E[S^2]$?
  - $E[S^2] = \sigma^2$
  - We say S^2 is “unbiased estimate” of σ^2

Proof that E[S^2] = σ^2 (just for reference)

$$E[S^2] = E\left[\sum_{i=1}^{n}\frac{(X_i - \overline{X})^2}{n-1}\right] \;\Rightarrow\; (n-1)E[S^2] = E\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right]$$

$$\begin{aligned}
(n-1)E[S^2] &= E\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right] = E\left[\sum_{i=1}^{n}((X_i - \mu) + (\mu - \overline{X}))^2\right] \\
&= E\left[\sum_{i=1}^{n}(X_i - \mu)^2 + \sum_{i=1}^{n}(\mu - \overline{X})^2 + 2\sum_{i=1}^{n}(X_i - \mu)(\mu - \overline{X})\right] \\
&= E\left[\sum_{i=1}^{n}(X_i - \mu)^2 + n(\mu - \overline{X})^2 + 2(\mu - \overline{X})\sum_{i=1}^{n}(X_i - \mu)\right] \\
&= E\left[\sum_{i=1}^{n}(X_i - \mu)^2 + n(\mu - \overline{X})^2 + 2(\mu - \overline{X})n(\overline{X} - \mu)\right] \\
&= E\left[\sum_{i=1}^{n}(X_i - \mu)^2 - n(\mu - \overline{X})^2\right] = \sum_{i=1}^{n}E[(X_i - \mu)^2] - nE[(\mu - \overline{X})^2] \\
&= n\sigma^2 - n\mathrm{Var}(\overline{X}) = n\sigma^2 - n\frac{\sigma^2}{n} = n\sigma^2 - \sigma^2 = (n-1)\sigma^2
\end{aligned}$$

- So, E[S^2] = σ^2