Sec 6.4 Stat 134 lec 37

Covariance and variance of a sum

$X, Y, \quad S = X + Y$

mean $\mu_X, \mu_Y, \mu_S = \mu_X + \mu_Y$

$$Var(S) = E\left((S-\mu_S)^2\right) = E(D_S^2)$$

D_S deviation from mean

$$D_S = S - \mu_S$$
$$= X + Y - (\mu_X + \mu_Y)$$
$$= D_X + D_Y$$

$Var(X) = E(X^2) - E(X)^2$
$= E\left((X - E(X))^2\right)$

$$Var(S) = E\left((D_X + D_Y)^2\right)$$
$$= E\left(D_X^2 + D_Y^2 + 2D_X D_Y\right)$$
$$= \underbrace{E(D_X^2)}_{=Var(X)} + \underbrace{E(D_Y^2)}_{=Var(Y)} + 2\underbrace{E(D_X D_Y)}_{=Cov(X,Y)}$$

defn

$$Cov(X,Y) = E(D_X D_Y) = E\left((X-\mu_X)(Y-\mu_Y)\right)$$
$$= E\left(XY - \mu_X Y - X\mu_Y + \mu_X \mu_Y\right)$$
$$= E(XY) - \mu_X\mu_Y - \mu_X\mu_Y + \mu_X\mu_Y)$$

$$= \boxed{E[XY] - E[X]E[Y]}$$

Easy Facts

$$\boxed{\begin{aligned} &Cov(X,X) = Var(X) \\ &Cov(X,Y) = Cov(Y,X) \\ &Cov(X,c) = 0 \end{aligned}}$$

Bilinearity property:

$$\boxed{\begin{aligned} &Cov(X+Y, Z) = Cov(X,Z) + Cov(Y,Z) \\ &Cov(aX, bY) = ab\, Cov(X,Y) \end{aligned}}$$

$E((X+Y)Z)$
$- E(X+Y)E(Z)$
etc.

More generally

$$\boxed{Cov\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j Cov(X_i, Y_j)}$$

ex $Cov(X-5Y,\ 3X+Y-Z+10)$

$= 3Var(X) + Cov(X,Y) - Cov(X,Z) + 0$

$\quad -15 Cov(Y,X) - 5Var(Y) + 5Cov(Y,Z) + 0$

If X, Y are indep

$$\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y) = 0$$

Ex I_A, I_B indicators

$$\mathrm{Cov}(I_A, I_B) = E(\underbrace{I_A I_B}_{= I_{AB}}) - E(I_A)E(I_B)$$

$$= P(AB) - P(A)P(B)$$

if $\mathrm{Cov}(I_A, I_B) > 0$ then $P(AB) > P(A)\cancel{P(B)}$

$$\parallel$$

$$P(A|B)\cancel{P(B)}$$

we say A and B are positively dependent

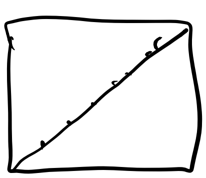
$\Rightarrow P(A|B) > P(A)$

says that given B A is more likely to occur.

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

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

X_i are all identically distributed,

n^2 

$$= n\mathrm{Var}(X_1) + n(n-1)\,\mathrm{Cov}(X_1, X_2)$$

not diagonal

ex You have n letters labelled $1, \dots, n$

You randomly put them into n envelopes labelled $1, \dots, n$.

Let M = # matches.

Find $E(M)$, $Var(M)$

$$M = I_1 + \dots + I_n \qquad I_1 = \begin{cases} 1 & \text{if } 1^{st} \text{ letter match} \\ 0 \end{cases} \quad \frac{1}{n}$$

$$E(M) = n \cdot \frac{1}{n} = 1$$

$$I_{1,2} = \begin{cases} 1 & \text{if } 1^{st} \text{ and } 2^{nd} \text{ letter match} \\ 0 \end{cases} \quad \frac{1}{n} \cdot \frac{1}{n-1}$$

$$Var(M) = Var(I_1 + \dots + I_n)$$

$$= \sum_{i=1}^{n} Var(I_i) + \sum_{i \neq j} Cov(I_i, I_j)$$

$$= n \underbrace{Var(I_1)}_{= \frac{1}{n} \frac{n-1}{n}} + n(n-1) \underbrace{Cov(I_1, I_2)}_{= E(I_{12}) - E(I_1)E(I_2)}$$

$$E(I_{12}) = \frac{1}{n} \cdot \frac{1}{n-1} \qquad E(I_1) = \frac{1}{n} \qquad E(I_2) = \frac{1}{n}$$

$$= 1$$

ex Population of N numbers, population avg μ, pop $SD = \sigma$.

Take a simple random sample (SRS) size n.

Sample sum $= S_n = X_1 + \dots + X_n$

$$E(S_n) = nE(X_1) = n\mu.$$

$$Var(S_n) = n\underbrace{Var(X_1)}_{\sigma^2} + n(n-1)\,Cov(X_1, X_2)$$

trick,

take a census, $n = N$

$$var(S_N) = 0$$

$$N\sigma^2 + N(N-1)\,Cov(X_1, X_2)$$

$$\Rightarrow \boxed{Cov(X_1, X_2) = \frac{-\sigma^2}{N-1}}$$

$$\text{so } Var(S_n) = n\sigma^2 + n(n-1)\left(\frac{-\sigma^2}{N-1}\right)$$

$$= \boxed{n\sigma^2\left(\frac{N-n}{N-1}\right)}$$

correction factor < 1

$Var(S_n)$ if draw with replacement (i.e $X_1, \dots, X_n$ indep) $<$ $Var(S_n)$ if draw without replacement.

Recall, we have seen this correction factor before (p241)

$$X \sim Bin(n, p) \Rightarrow Var(X) = (npq)$$

$$X \sim \text{hyper-geom}(n, N, 6) \Rightarrow Var(X) = n \underbrace{\left(\frac{6}{N}\right)}_{p} \underbrace{\left(\frac{N-6}{N}\right)}_{q} \underbrace{\left(\frac{N-n}{N-1}\right)}_{\text{correction factor}}.$$