3.6 Correlation

Def. a) $Cov(X,Y) = E[(X-\mu_1)(Y-\mu_2)]$, $\mu_1 = E(X)$, $\mu_2 = E(Y)$.

b) $\rho = \rho_{X,Y} = \dfrac{Cov(X,Y)}{\sigma_1 \sigma_2}$, $\sigma_1 = \sqrt{Var(X)}$, $\sigma_2 = \sqrt{Var(Y)}$.

Note $Cov(X,Y) = \rho\,\sigma_1 \sigma_2$

Ex 1. Let X, Y be r.v. with $\sigma_2^2 = Var(Y)$, $\sigma_1^2 = Var(X)$, and correlation $\rho = \rho(X,Y)$. Confirm that

(1) $\boxed{\text{a) } Y = \rho \dfrac{\sigma_2}{\sigma_1} X + Z, \text{ and } X, Z \text{ are uncorrelated: } Cov(X,Z) = 0}$

b) $Var(Z) = \sigma_2^2(1-\rho^2)$

Answer. a) Since $Z = Y - \rho \dfrac{\sigma_2}{\sigma_1} X$,

$Cov(X,Z) = Cov\left(X, Y - \rho\dfrac{\sigma_2}{\sigma_1} X\right) = Cov(X,Y) - \rho\dfrac{\sigma_2}{\sigma_1} Var(X) =$

$= Cov(X,Y) - \rho\,\sigma_2 \cdot \sigma_1 = 0$

b) $Var(Z) = Var(Y) + \rho^2 \dfrac{\sigma_2^2}{\sigma_1^2} Var(X) - 2\rho \dfrac{\sigma_2}{\sigma_1} Cov(X,Y) =$

$= \sigma_2^2 + \rho^2 \dfrac{\sigma_2^2}{\cancel{\sigma_1^2}} \cdot \cancel{\sigma_1^2} - 2\rho^2 \dfrac{\sigma_2^2}{\cancel{\sigma_1}} \cancel{\sigma_1} = \sigma_2^2(1-\rho^2)$

Comments on Ex 1. (1) Since $Var(Z) \geq 0$, we have $1-\rho^2 \geq 0$, equivalently, $|\rho| \leq 1$ $(-1 \leq \rho \leq 1)$.

(2) "Extreme" correlation is when $\rho^2 = 1$ or $\rho = \pm 1$: in this case, $Z = c$ (constant) with prob. 1, and $Y = \pm \dfrac{\sigma_2}{\sigma_1} X + c$

(3) The roles of X and Y can be switched:

$X = \rho \dfrac{\sigma_1}{\sigma_2} Y + \tilde{Z}$, where $Y, \tilde{Z}$ are uncorrelated, and

$\mathrm{Var}(\tilde{Z}) = \sigma_1^2 (1-\rho^2)$

Remark 1. Let $E(X)=\mu_1$, $E(Y)=\mu_2$. Eqn (1) can be rewritten as

(1') $Y-\mu_2 = \rho \frac{\sigma_2}{\sigma_1}(X-\mu_1) + V$, where X, V are uncorrelated,

$\underline{E(V)=0}$, $\mathrm{Var}(V) = \sigma_2^2(1-\rho^2)$

on average, positive and negative values of V are "balanced".

$|\rho| \le 1$ by Cauchy-Schwarz inequality:

$$|\rho| = \frac{|\mathrm{Cov}(X,Y)|}{\sigma_1 \sigma_2} \le 1 \text{ means}$$

$$(2) \quad \left| E\left[(X-\mu_1)(Y-\mu_2)\right] \right| \le \sqrt{E\left[(X-\mu_1)^2\right]}\ \sqrt{E\left[(Y-\mu_2)^2\right]};$$

(2) is a consequence of Cauchy-Schwarz inequality

$$(*) \quad \boxed{|E(UV)| \le \sqrt{E(U^2)}\ \sqrt{E(V^2)}}$$

$U = X-\mu_1$
$V = Y-\mu_2$

Proof of $(*)$. For any $t \in \mathbb{R}$,

$$0 \le E\left[(U+tV)^2\right] = E(U^2) + 2E(UV)t + E(V^2)t^2$$

$= c + 2bt + at^2$, where $c = E(U^2)$, $b = E(UV)$, $a = E(V^2)$.

So, discriminant $D = (2b)^2 - 4ac \le 0$, equivalently, $b^2 \le ac$,

$|b| \le \sqrt{a}\sqrt{c}$.

Some inequalities

1. If $X \ge 0$, then $E(X) \ge 0$.

If $X \geq Y$, then $E(X) \geq E(Y)$.

2. $|E(X)| \leq E(|X|)$, because $-|X| \leq X \leq |X|$ implies $-E(|X|) \leq E(X) \leq E(|X|)$.

3. $|E(X)| \leq E(|X|) \leq \sqrt{E(X^2)}$ because by Cauchy-Schwarz

$$E(|X|) = E(|X| \cdot 1) \leq \sqrt{E(X^2)} \cdot \sqrt{E(1)} = \sqrt{E(X^2)}.$$

3.4 Indicator method

Recall

(a) $1_A = \begin{cases} 1 & \text{if } A \\ 0 & \text{if not } A \end{cases}$ is Bernoulli(p) with $p = P(A)$

$$1_{A^c} = 1 - 1_A, \quad 1_A 1_B = 1_{A \cap B}$$

(b) $E(1_A) = P(A) = p$, $Var(1_A) = P(A) - P(A)^2$

$$Cov(1_A, 1_B) = E(1_A 1_B) - E(1_A)E(1_B)$$
$$= P(A \cap B) - P(A)P(B)$$

A, B independent iff $Cov(1_A, 1_B) = 0$.

Ex 1. Hats of n people are mixed up. Everyone randomly one by one picks a hat. Let $X = \#$ of matches.

Find $E(X) = 1$, and $Var(X) = 1$.

Answer. A_i = "ith person gets right hat", $i = 1, \dots, n$.

Then $X = \sum_{i=1}^{n} 1_{A_i}$, $E(X) = \sum_{i=1}^{n} P(A_i) = n \cdot \frac{1}{n} = 1$.

$P(A_i) = \frac{(n-1)!}{n!} = \frac{1}{n}$

$$Var(X) = E(X^2) - E(X)^2 = 2 - 1 = 1.$$

$$E(X^2) = E\left(\sum_{i,j=1}^{n} 1_{A_i} 1_{A_j}\right) = E\left(\sum_{i=1}^{n} 1_{A_i} + 2\sum_{i<j} 1_{A_iA_j}\right) = \sum_{i=1}^{n} P(A_i) + 2\sum_{i<j} P(A_iA_j)$$

$$= n \cdot \frac{1}{n} + \cancel{2} \cdot \frac{n(n-1)}{\cancel{2}} \cdot \frac{1}{n(n-1)} = 2, \text{ because}$$

$$P(A_iA_j) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)}.$$

$$\boxed{\text{If } X = \sum_{i=1}^{n} 1_{A_i}, \text{ then } X^2 = \left(\sum_{i=1}^{n} 1_{A_i}\right)^2 = \sum_{i,j=1}^{n} 1_{A_i} 1_{A_j} = \sum_{i=1}^{n} 1_{A_i} + 2\sum_{i<j} 1_{A_iA_j}, \text{ and } E(X^2) = \sum_{i=1}^{n} P(A_i) + 2\sum_{i<j} P(A_iA_j)}$$

Ex 2. There are m red and w white balls in the box, $K = m + w$. n balls are removed one by one. Assume, $n < m < K$. Let X = # of red balls among n. X is hypergeometric.

a) Find $E(X)$.

Answer. A_i = "ith removed ball is red". Then

$$X = 1_{A_1} + \dots + 1_{A_n}, \quad E(X) = \sum_{i=1}^{n} P(A_i) = np = n \cdot \frac{m}{K}.$$

$$P(A_i) = \frac{m \cdot \cancel{(K-1) \cdots (K-1-(n-1)+1)}}{K \cancel{(K-1) \cdots (K-n+1)}} = \frac{m}{K} =: p$$

b) Find $Var(X) = E(X^2) - E(X)^2 = E(X^2) - n^2p^2$.

Answer. $$X^2 = \sum_{i,j} 1_{A_i} 1_{A_j} = \sum_{i=1}^{n} 1_{A_i} + 2\sum_{i<j} 1_{A_iA_j}$$

$$E(X^2) = \sum_{i=1}^{n} P(A_i) + 2\sum_{i<j} P(A_iA_j) =$$

$$\boxed{P(A_i) = p, \quad P(A_iA_j) = \frac{m(m-1)\cancel{(K-2) \cdots (K-2-(n-2)+1)}}{K(K-1)\cancel{\cdots (K-n+1)}} = \frac{m}{K} \frac{(m-1)}{K-1} \left(\approx p^2\right)}$$

$$= np + \cancel{2} \frac{n(n-1)}{\cancel{2}} \cdot \frac{m}{K} \cdot \frac{m-1}{K-1} = np + n(n-1) \frac{m}{K} \cdot \frac{m-1}{K-1}$$

$$= np + \cancel{2}\,\frac{n(n-1)}{\cancel{2}} \cdot \frac{m}{K} \cdot \frac{m-1}{K-1} = np + n(n-1)\,\frac{m}{K} \cdot \frac{m-1}{K-1}$$

$$Var(X) = np + n(n-1)\,\frac{m}{K}\,\frac{m-1}{K-1} - n^2 p^2 .$$