CS 70 Discrete Mathematics and Probability Theory
Fall 2026 Hug, Sabin

DIS 11

Variance, Covariance, and Correlation

Throughout, X and Y are discrete random variables.

Variance is the average squared distance from the mean:

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2,$$

where $\mathbb{E}[X^2] = \sum_x x^2 \, \mathbb{P}[X = x]$. It is nonnegative, and its square root is the standard deviation. For constants a, b,

$$\mathrm{Var}(aX + b) = a^2 \, \mathrm{Var}(X).$$

If $X_1, \ldots, X_n$ are independent, then $\mathrm{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathrm{Var}(X_i)$. Unlike linearity of expectation, this can fail for dependent variables.

Covariance measures whether X and Y tend to deviate from their means in the same direction (positive) or in opposite directions (negative):

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

where $\mathbb{E}[XY] = \sum_x \sum_y xy \, \mathbb{P}[X = x, Y = y]$. In general,

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{cov}(X, Y).$$

Independent variables have zero covariance, but zero covariance does not imply independence.

Correlation rescales covariance by the standard deviations. When both variances are positive (otherwise it is undefined),

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} \in [-1, 1].$$

Correlation is unitless and measures linear association: it is ± 1 exactly when $Y = aX + b$ with probability 1, with $a > 0$ for $+1$ and $a < 0$ for -1. Variables with zero correlation are called *uncorrelated*; they need not be independent.

1 Binomial Variance

Note 21

Throw n balls into m bins uniformly at random. For a specific ball i, what is the variance of the number of roommates it has (i.e. the number of other balls that it shares its bin with)?

Solution:

Let's say that ball i lands in some bin B_i; we care about how many of the remaining $n-1$ balls land in B_i as well. For each $j \neq i$, call I_j the indicator that ball j lands in B_i. Then, if N_i is the number of roommates, we can write

$$N_i = \sum_{j \neq i} I_j.$$

Since the locations of the balls are all independent, the random variables I_j are as well (knowing that one ball landed in the same place as ball i doesn't tell us anything about whether another one will). Since the variance of a sum of independent random variables is the sum of their individual variances, all we need to know is $\text{var}(I_j)$. But, since there are m bins, $\mathbb{P}[I_j = 1] = 1/m$ and $\mathbb{P}[I_j = 0] = 1 - 1/m$; in other words I_j is a Bernoulli random variable with parameter $1/m$. We discussed in the notes that the variance of such a thing is $1/m(1-1/m)$, so we will have

$$\text{var}(N_i) = \sum_{j \neq i} \text{var}(I_j) = \sum_{j \neq i} \frac{1}{m}\left(1 - \frac{1}{m}\right) = (n-1)\frac{1}{m}\left(1 - \frac{1}{m}\right)$$

2 Diversify Your Hand

Note 20
Note 21

You are dealt 5 cards from a standard 52 card deck. Let X be the number of distinct values in your hand. For instance, the hand (A, A, A, 2, 3) has 3 distinct values.

(a) Calculate $\mathbb{E}[X]$. (Hint: Consider indicator variables X_i representing whether value i appears in the hand.)

(b) Calculate $\text{Var}(X)$. The answer expression will be quite involved; you do not need to simplify anything.

Solution:

(a) Let X_i be the indicator of the ith value appearing in your hand. Then, $X = X_1 + X_2 + ... + X_{13}$. (Here we let 13 correspond to K, 12 correspond to Q, and 11 correspond to J.) By linearity of expectation, $\mathbb{E}[X] = \sum_{i=1}^{13} \mathbb{E}[X_i]$.

We can calculate $\mathbb{P}[X_i = 1]$ by taking the complement, $1 - \mathbb{P}[X_i = 0]$, or 1 minus the probability that value i does not appear in your hand. In that case all 5 cards come from the 48 cards of other values, so this is $1 - \frac{\binom{48}{5}}{\binom{52}{5}}$.

Then, $\mathbb{E}[X] = 13\,\mathbb{P}[X_1 = 1] = 13\left(1 - \frac{\binom{48}{5}}{\binom{52}{5}}\right)$.

(b) To calculate variance, since the indicators are not independent, we have to use the formula $\mathbb{E}[X^2] = \sum_i \mathbb{E}[X_i^2] + \sum_{i \neq j} \mathbb{E}[X_i X_j]$.

First, we have

$$\sum_i \mathbb{E}[X_i^2] = \sum_i \mathbb{E}[X_i] = 13\left(1 - \frac{\binom{48}{5}}{\binom{52}{5}}\right).$$

Next, we tackle $\sum_{i \neq j} \mathbb{E}[X_i X_j]$. Note that $\mathbb{E}[X_i X_j] = \mathbb{P}[X_i X_j = 1]$, as $X_i X_j$ is either 0 or 1.

To calculate $\mathbb{P}[X_i X_j = 1]$ (the probability that values i and j both appear in our hand), we use inclusion-exclusion on the complement: $\mathbb{P}[X_i X_j = 1] = 1 - \mathbb{P}[X_i = 0] - \mathbb{P}[X_j = 0] + \mathbb{P}[X_i = 0, X_j = 0]$. Here $\mathbb{P}[X_i = 0, X_j = 0] = \frac{\binom{44}{5}}{\binom{52}{5}}$, since all 5 cards must avoid the 8 cards of values i and j. Then

$$\begin{aligned}
\sum_{i \neq j} \mathbb{E}[X_i X_j] &= 13 \cdot 12\,\mathbb{P}[X_i X_j = 1] \\
&= 13 \cdot 12(1 - \mathbb{P}[X_i = 0] - \mathbb{P}[X_j = 0] + \mathbb{P}[X_i = 0, X_j = 0]) \\
&= 156\left(1 - 2\frac{\binom{48}{5}}{\binom{52}{5}} + \frac{\binom{44}{5}}{\binom{52}{5}}\right)
\end{aligned}$$

Putting it all together, we have

$$\begin{aligned}
\text{Var}(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\
&= 13\left(1 - \frac{\binom{48}{5}}{\binom{52}{5}}\right) + 156\left(1 - 2\frac{\binom{48}{5}}{\binom{52}{5}} + \frac{\binom{44}{5}}{\binom{52}{5}}\right) - \left(13\left(1 - \frac{\binom{48}{5}}{\binom{52}{5}}\right)\right)^2.
\end{aligned}$$

3 Covariance and Correlation Warm-Up

Note 22

Let X, Y be random variables where $\mathbb{P}[X=0, Y=0] = \frac{1}{4}$, $\mathbb{P}[X=1, Y=2] = \frac{1}{2}$, and $\mathbb{P}[X=-1, Y=2] = \frac{1}{4}$.

(a) What is $\text{cov}(X, Y)$?

(b) What is $\text{Corr}(X, Y)$?

(c) What is $\text{cov}(X+2Y, 3X-Y)$?

Solution:

(a) The three possible pairs (X, Y) are $(0,0)$, $(1,2)$, and $(-1,2)$. Thus,

$$\mathbb{E}[X] = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + (-1) \cdot \frac{1}{4} = \frac{1}{4},$$
$$\mathbb{E}[Y] = 0 \cdot \frac{1}{4} + 2 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = \frac{3}{2},$$
$$\mathbb{E}[XY] = 0 \cdot \frac{1}{4} + 2 \cdot \frac{1}{2} + (-2) \cdot \frac{1}{4} = \frac{1}{2}.$$

Using the definition of covariance,

$$\text{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = \frac{1}{2} - \frac{1}{4} \cdot \frac{3}{2} = \frac{1}{8}.$$

(b) We have $\mathbb{E}[X^2] = 1 \cdot \frac{1}{2} + 1 \cdot \frac{1}{4} = \frac{3}{4}$ and $\mathbb{E}[Y^2] = 4 \cdot \frac{1}{2} + 4 \cdot \frac{1}{4} = 3$, so

$$\text{Var}(X) = \frac{3}{4} - \left(\frac{1}{4}\right)^2 = \frac{11}{16}, \qquad \text{Var}(Y) = 3 - \left(\frac{3}{2}\right)^2 = \frac{3}{4}.$$

Therefore,

$$\text{Corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}} = \frac{1/8}{\sqrt{33/64}} = \frac{1}{\sqrt{33}}.$$

(c) By bilinearity of covariance, and using $\text{cov}(X, X) = \text{Var}(X)$ and $\text{cov}(Y, Y) = \text{Var}(Y)$,

$$\begin{aligned}\text{cov}(X+2Y, 3X-Y) &= 3\,\text{cov}(X, X) - \text{cov}(X, Y) + 6\,\text{cov}(Y, X) - 2\,\text{cov}(Y, Y) \\ &= 3\,\text{Var}(X) + 5\,\text{cov}(X, Y) - 2\,\text{Var}(Y) \\ &= 3 \cdot \frac{11}{16} + 5 \cdot \frac{1}{8} - 2 \cdot \frac{3}{4} = \frac{19}{16}.\end{aligned}$$

4 Correlation and Independence

Note 20
Note 22

(a) What does it mean for two random variables to be uncorrelated?

(b) What does it mean for two random variables to be independent?

(c) Are all uncorrelated variables independent? Are all independent variables uncorrelated? If your answer is yes, justify your answer; if your answer is no, give a counterexample.

Solution:

(a) Recall that for two random variables X and Y,

$$\text{cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y].$$

Two random variables are uncorrelated iff their covariance is equal to zero. If X and Y are uncorrelated, then there is no linear relationship between them.

(b) Recall that two random variables X and Y are independent if and only if the following criteria are met (the three criteria are equivalent and connected by Bayes rule):

$$\begin{aligned}\mathbb{P}[X = x \mid Y = y] &= \mathbb{P}[X = x] \\ \mathbb{P}[Y = y \mid X = x] &= \mathbb{P}[Y = y] \\ \mathbb{P}[X = x, Y = y] &= \mathbb{P}[X = x]\,\mathbb{P}[Y = y]\end{aligned}$$

for all x, y such that $\mathbb{P}[X = x], \mathbb{P}[Y = y] > 0$.

If X and Y are independent, any information about one variable offers no information whatsoever about the other variable.

(c) All independent variables are uncorrelated: if X and Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$, so $\text{cov}(X,Y) = 0$. The converse, however, is not true: two uncorrelated variables may not be independent. Consider two variables X and Y that follow a uniform joint distribution over the points $(1,0), (0,1), (-1,0), (0,-1)$. See Figure 1.

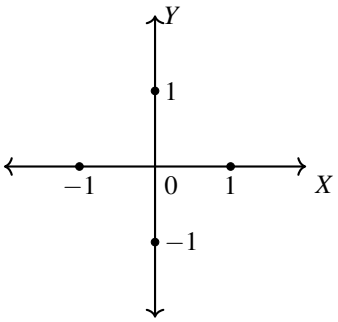


Figure 1: Choose one of the four points shown uniformly at random.

Then

$$\text{cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = 0.$$

To see why, observe that $XY = 0$ always because at least one of X and Y is always 0, and furthermore $\mathbb{E}[X] = \mathbb{E}[Y] = 0$ because both X and Y are symmetric around 0. So, there is no linear relationship, but X and Y are not independent (for example, $\mathbb{P}[Y = 0] = 1/2$ but $\mathbb{P}[Y = 0 \mid X = 1] = 1$).

5 Coupon Collector Variance

Note 21

It's that time of the year again (again)—Safeway is offering its Monopoly Card promotion. Each time you visit Safeway, you are given one of n different Monopoly Cards with equal probability. You need to collect them all to redeem the grand prize.

Let X be the number of visits you have to make before you can redeem the grand prize. Show that

$$\text{Var}(X) = n^2\left(\sum_{i=1}^{n}\frac{1}{i^2}\right) - \mathbb{E}[X].$$

Solution:

Note that this is the coupon collector's problem, but now we have to find the variance. Let X_i be the number of additional visits needed to collect the ith unique Monopoly card, given that we have already collected $i-1$ unique Monopoly cards. Then $X = \sum_{i=1}^{n} X_i$ and each X_i is geometrically distributed with $p = (n-i+1)/n$. Moreover, the random variables themselves are independent, since each time you collect a new card, you are starting from a clean slate.

$$\begin{aligned}
\text{Var}(X) &= \sum_{i=1}^{n}\text{Var}(X_i) && \text{(as the } X_i \text{ are independent)}\\
&= \sum_{i=1}^{n}\frac{1-(n-i+1)/n}{[(n-i+1)/n]^2} && \text{(variance of a geometric r.v. is } (1-p)/p^2)\\
&= \sum_{i=1}^{n}\frac{1-i/n}{(i/n)^2} && \text{(substituting } i \to n-i+1, \text{ which still ranges from 1 to } n)\\
&= \sum_{i=1}^{n}\frac{n(n-i)}{i^2}\\
&= \sum_{i=1}^{n}\frac{n^2}{i^2} - \sum_{i=1}^{n}\frac{n}{i}\\
&= n^2\left(\sum_{i=1}^{n}\frac{1}{i^2}\right) - \mathbb{E}[X] && \text{(using the coupon collector problem expected value).}
\end{aligned}$$