## Variance Intro

**Variance**: denoted by $\mathrm{Var}(X)$; measure of how much $X$ deviates from its mean, i.e. its spread.

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

Properties: for random variables $X$, $Y$ and constant $a$,

- $\mathrm{Var}(aX) = a^2 \mathrm{Var}(X)$
- $\mathrm{Var}(X + a) = \mathrm{Var}(X)$
- If $X$, $Y$ independent, then $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

**Variance of sum of (not necessarily independent) indicator variables**: Let $X_1, \ldots, X_n$ be indicator variables for events $A_1, \ldots, A_n$, respectively. The variance of the sum $X = X_1 + \cdots + X_n$ can be calculated as:

$$\mathrm{Var}(X) = \mathbb{E}[(X_1 + \cdots + X_n)^2] - \mathbb{E}[X_1 + \cdots + X_n]^2 = \sum_{i=1}^{n} \mathbb{E}[X_i^2] + \sum_{i \neq j} \mathbb{E}[X_i X_j] - \left(\sum_{i=1}^{n} \mathbb{E}[X_i]\right)^2$$

$\mathbb{E}[X_i^2] = \mathbb{E}[X_i] = \mathbb{P}[A_i]$ since $X_i^2 = X_i$ for indicator variables, and $\mathbb{E}[X_i X_j] = \mathbb{P}[A_i \cap A_j]$.

**Covariance**: measure of the relationship between two RVs

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

The sign of $\mathrm{cov}(X,Y)$ illustrates how $X$ and $Y$ are related; a positive value means that $X$ and $Y$ increase and decrease together, while a negative value means that $X$ increases as $Y$ decreases (and vice versa). A covariance of zero means that the two random variables are uncorrelated—there is no relationship between them.

Properties: for random variables $X$, $Y$, $Z$ and constant $a$,

- $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{cov}(X,Y)$
- $\mathrm{cov}(X,X) = \mathrm{Var}(X)$
- $\mathrm{cov}(X,Y) = \mathrm{cov}(Y,X)$
- Bilinearity: $\mathrm{cov}(X+Y,Z) = \mathrm{cov}(X,Z) + \mathrm{cov}(Y,Z)$ and $\mathrm{cov}(aX,Y) = a\,\mathrm{cov}(X,Y)$

**Correlation**: standardized form of covariance, always between $-1$ and $+1$.

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y},$$

where $\sigma_X = \sqrt{\mathrm{Var}(X)}$ is the standard deviation of $X$.

# 1 Dice Variance

Note 16

(a) Let $X$ be a random variable representing the outcome of the roll of one fair 6-sided die. What is $\text{Var}(X)$?

(b) Let $Z$ be a random variable representing the average of $n$ rolls of a fair 6-sided die. What is $\text{Var}(Z)$?

**Solution:**

(a) Recall that $\text{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$. We can compute each of the individual terms using the definition of expectation:

$$\begin{aligned}
\mathbb{E}[X] &= \frac{1}{6}(1+2+3+4+5+6) = \frac{7}{2} \\
\mathbb{E}\left[X^2\right] &= \frac{1}{6}(1^2+2^2+3^2+4^2+5^2+6^2) \\
&= \frac{1}{6}(1+4+9+16+25+36) = \frac{91}{6}
\end{aligned}$$

Now, we plug back into the variance expression:

$$\begin{aligned}
\text{Var}(X) &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \\
&= \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}
\end{aligned}$$

(b) Because each die roll is independent of the others, we can utilize the fact that for independent random variables $X$ and $Y$, $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$. Let $X_i$ be a random variable representing the outcome of the $i$th dice roll. We now have:

$$\begin{aligned}
\text{Var}(Z) &= \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \left(\frac{1}{n}\right)^2 \text{Var}(\sum_{i=1}^{n} X_i) \\
&= \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \text{Var}(X_i) && (X_i\text{'s are independent}) \\
&= \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \frac{35}{12} && \text{(from (a))} \\
&= \left(\frac{1}{n}\right)^2 \cdot n \cdot \frac{35}{12} = \frac{35}{12n}
\end{aligned}$$

# 2 Student Life

Note 19

In an attempt to avoid having to do laundry often, Marcus comes up with a system. Every night, he designates one of his shirts as his dirtiest shirt. In the morning, he randomly picks one of his shirts to wear. If he picked the dirtiest one, he puts it in a dirty pile at the end of the day (a shirt in the dirty pile is not used again until it is cleaned).

When Marcus puts his last shirt into the dirty pile, he finally does his laundry, and again designates one of his shirts as his dirtiest shirt (laundry isn't perfect) before going to bed. This process then repeats.

(a) If Marcus has $n$ shirts, what is the expected number of days that transpire between laundry events? Your answer should be a function of $n$ involving no summations.

(b) Say he gets even lazier, and instead of organizing his shirts in his dresser every night, he throws his shirts randomly onto one of $n$ different locations in his room (one shirt per location), designates one of his shirts as his dirtiest shirt, and one location as the dirtiest location.

In the morning, if he happens to pick the dirtiest shirt, *and* the dirtiest shirt was in the dirtiest location, then he puts the shirt into the dirty pile at the end of the day and does not throw any future shirts into that location and also does not consider it as a candidate for future dirtiest locations (it is too dirty).

What is the expected number of days that transpire between laundry events now? Again, your answer should be a function of $n$ involving no summations.

**Solution:**

(a) The number of days that it takes for him to throw a shirt into the dirty pile can be represented as a geometric RV. For the first shirt, this is the geometric RV with $p = 1/n$. We can see this by noticing that every day up to the day he picks the dirtiest shirt, the probability of getting the dirtiest shirt remains $1/n$.

We'll call $X_i$ the number of days that go until he throws the $i$th shirt into the dirty pile. Since on the $i$th shirt, there are $n-i+1$ shirts left, we get that $X_i \sim \text{Geometric}(1/(n-i+1))$. The number of days until he does his laundry is a sum of these variables. Therefore, we can get the following result:

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

(b) For this part we can use a similar approach but the probability for $X_i$ becomes $1/(n-i+1)^2$. This is because the dirtiest shirt falls into the dirtiest spot with probability $1/(n-i+1)$ and we pick it after that with probability $1/(n-i+1)$, so the probability of picking the dirtiest shirt from the dirtiest spot for the $i$th shirt is $1/(n-i+1)^2$. Using the same approach, we get the following sum:

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n}(n-i+1)^2 = \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$$

## 3 Elevator Variance

Note 16

A building has $n$ upper floors numbered $1, 2, \ldots, n$, plus a ground floor $G$. At the ground floor, $m$ people get on the elevator together, and each person gets off at one of the $n$ upper floors uniformly at random and independently of everyone else. What is the *variance* of the number of floors the elevator *does not* stop at?

**Solution:** Let $N$ be the number of floors the elevator does not stop at. We can represent $N$ as the sum of the indicator variables $I_1, \ldots, I_n$, where $I_i = 1$ if no one gets off on floor $i$. Thus, we have

$$\mathbb{E}[I_i] = \mathbb{P}[I_i = 1] = \left(\frac{n-1}{n}\right)^m,$$

and from linearity of expectation,

$$\mathbb{E}[N] = \sum_{i=1}^{n} \mathbb{E}[I_i] = n\left(\frac{n-1}{n}\right)^m.$$

To find the variance, we cannot simply sum the variance of our indicator variables. However, since $\mathrm{Var}(N) = \mathbb{E}[N^2] - \mathbb{E}[N]^2$ the only piece we don't already know is $\mathbb{E}[N^2]$. We can calculate this by again expanding $N$ as a sum:

$$\mathbb{E}[N^2] = \mathbb{E}[(I_1 + \cdots + I_n)^2] = \mathbb{E}\left[\sum_{i,j} I_i I_j\right] = \sum_{i,j} \mathbb{E}[I_i I_j] = \sum_{i}^{n} \mathbb{E}[I_i^2] + \sum_{i \neq j} \mathbb{E}[I_i I_j].$$

The first term is simple to calculate: since $I_i$ is an indicator, $I_i^2 = I_i$, so we have

$$\mathbb{E}[I_i^2] = \mathbb{E}[I_i] = \mathbb{P}[I_i = 1] = \left(\frac{n-1}{n}\right)^m,$$

meaning that

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

From the definition of the variables $I_i$, we see that $I_i I_j = 1$ when both $I_i$ and $I_j$ are 1, which means no one gets off the elevator on floor $i$ and floor $j$. This happens with probability

$$\mathbb{P}[I_i = I_j = 1] = \mathbb{P}[I_i = 1 \cap I_j = 1] = \left(\frac{n-2}{n}\right)^m.$$

Thus we now know

$$\sum_{i \neq j} \mathbb{E}[I_i I_j] = n(n-1)\left(\frac{n-2}{n}\right)^m,$$

and we can assemble everything we've done so far to see that

$$\mathrm{Var}(N) = \mathbb{E}[N^2] - \mathbb{E}[N]^2 = n\left(\frac{n-1}{n}\right)^m + n(n-1)\left(\frac{n-2}{n}\right)^m - n^2\left(\frac{n-1}{n}\right)^{2m}.$$

## 4 Covariance

Note 16

(a) We have a bag of 5 red and 5 blue balls. We take two balls uniformly at random from the bag without replacement. Let $X_1$ and $X_2$ be indicator random variables for the events of the first and second ball being red, respectively. What is $\text{cov}(X_1, X_2)$? Recall that $\text{cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$.

(b) Now, we have two bags A and B, with 5 red and 5 blue balls each. Draw a ball uniformly at random from A, record its color, and then place it in B. Then draw a ball uniformly at random from B and record its color. Let $X_1$ and $X_2$ be indicator random variables for the events of the first and second draws being red, respectively. What is $\text{cov}(X_1, X_2)$?

**Solution:**

(a) We can use the formula $\text{cov}(X_1, X_2) = \mathbb{E}[X_1X_2] - \mathbb{E}[X_1]\,\mathbb{E}[X_2]$.

$$\begin{aligned}
\mathbb{E}[X_1] &= \frac{5}{10} \times 1 + \frac{5}{10} \times 0 = \frac{1}{2}, \\
\mathbb{E}[X_2] &= \frac{5}{10} \times 1 + \frac{5}{10} \times 0 = \frac{1}{2}, \\
\mathbb{E}[X_1X_2] &= \frac{5}{10} \cdot \frac{4}{9} \times 1 + \left(1 - \frac{5}{10} \cdot \frac{4}{9}\right) \times 0 = \frac{2}{9}.
\end{aligned}$$

Therefore,

$$\text{cov}(X_1, X_2) = \mathbb{E}[X_1X_2] - \mathbb{E}[X_1]\,\mathbb{E}[X_2] = \frac{2}{9} - \frac{1}{2} \times \frac{1}{2} = -\frac{1}{36}.$$

(b) Again, we use the formula $\text{cov}(X_1, X_2) = \mathbb{E}[X_1X_2] - \mathbb{E}[X_1]\mathbb{E}[X_2]$.

$$\begin{aligned}
\mathbb{E}[X_1] &= \frac{5}{10} \times 1 + \frac{5}{10} \times 0 = \frac{1}{2} \\
\mathbb{E}[X_2] &= \left(\frac{5}{10} \times \frac{6}{11} + \frac{5}{10} \times \frac{5}{11}\right) \times 1 + \left(\frac{5}{10} \times \frac{5}{11} + \frac{5}{10} \times \frac{6}{11}\right) \times 0 = \frac{1}{2} \\
\mathbb{E}[X_1X_2] &= \frac{5}{10} \times \frac{6}{11} \times 1 = \frac{30}{110}.
\end{aligned}$$

Therefore,

$$\mathbb{E}[X_1X_2] - \mathbb{E}[X_1]\mathbb{E}[X_2] = \frac{30}{110} - \frac{1}{4} = \frac{1}{44}.$$

Note that in part (a), if one event happened, the other would be less likely to happen, and thus the covariance was negative. Similarly, in part (b), if one event happened, the other would be more likely to happen, and thus the covariance was positive.