1 Coupon Collector Variance

Note 19

It's that time of the year 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 $\text{Var}(X) = n^2\left(\sum_{i=1}^n i^{-2}\right) - \mathbb{E}[X]$.

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 visits we need to make before we have collected the ith unique Monopoly card actually obtained, 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 indicators themselves are independent, since each time you collect a new card, you are starting from a clean slate.

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

2 Diversify Your Hand

Note 15
Note 16

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 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 the card does not appear in your hand. 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 $\mathbb{E}[X^2] = \sum_{i=j} \mathbb{E}[X_i^2] + \sum_{i \neq j} \mathbb{E}[X_i X_j]$.

First, we have

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

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 we have both cards in our hand), we note that $\mathbb{P}[X_i X_j = 1] = 1 - \mathbb{P}[X_i = 0] - \mathbb{P}[X_j = 0] + \mathbb{P}[X_i = 0, X_j = 0]$. 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 Double-Check Your Intuition Again

Note 16

(a) You roll a fair six-sided die and record the result X. You roll the die again and record the result Y.

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

(ii) Prove that $X+Y$ and $X-Y$ are not independent.

For each of the problems below, if you think the answer is "yes" then provide a proof. If you think the answer is "no", then provide a counterexample.

(b) If X is a random variable and $\text{Var}(X) = 0$, then must X be a constant?

(c) If X is a random variable and c is a constant, then is $\text{Var}(cX) = c\,\text{Var}(X)$?

(d) If A and B are random variables with nonzero standard deviations and $\text{Corr}(A, B) = 0$, then are A and B independent?

(e) If X and Y are not necessarily independent random variables, but $\text{Corr}(X,Y) = 0$, and X and Y have nonzero standard deviations, then is $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$?

(f) If X and Y are random variables then is $\mathbb{E}[\max(X,Y)\min(X,Y)] = \mathbb{E}[XY]$?

(g) If X and Y are independent random variables with nonzero standard deviations, then is

$$\text{Corr}(\max(X,Y), \min(X,Y)) = \text{Corr}(X,Y)?$$

Solution:

(a) (i) Using bilinearity of covariance, we have

$$\begin{aligned}\text{cov}(X+Y, X-Y) &= \text{cov}(X,X) + \text{cov}(X,Y) - \text{cov}(Y,X) - \text{cov}(Y,Y) \\ &= \text{cov}(X,X) - \text{cov}(Y,Y), \\ &= 0\end{aligned}$$

where we use that $\text{cov}(X,Y) = \text{cov}(Y,X)$ to get the second equality.

(ii) Observe that $\mathbb{P}[X+Y=7, X-Y=0] = 0$ because if $X-Y=0$, then the sum of our two dice rolls must be even. However, both $\mathbb{P}[X+Y=7]$ and $\mathbb{P}[X-Y=0]$ are nonzero, so $\mathbb{P}[X+Y=7, X-Y=0] \neq \mathbb{P}[X+Y=7] \cdot \mathbb{P}[X-Y=0]$.

(b) Yes. If we write $\mu = \mathbb{E}[X]$, then $0 = \text{Var}(X) = \mathbb{E}[(X-\mu)^2]$ so $(X-\mu)^2$ must be identically 0 since perfect squares are non-negative. Thus $X = \mu$.

(c) No. We have $\text{Var}(cX) = \mathbb{E}[(cX - \mathbb{E}[cX])^2] = c^2\,\mathbb{E}[(X - \mathbb{E}[X])^2] = c^2\,\text{Var}(X)$ so if $\text{Var}(X) \neq 0$ and $c \neq 0$ or $c \neq 1$ then $\text{Var}(cX) \neq c\,\text{Var}(X)$. This does prove that $\sigma(cX) = c\sigma(X)$ though.

(d) No. Let $A = X+Y$ and $B = X-Y$ from part (a). Since A and B are not constants then part (b) says they must have nonzero variances which means they also have nonzero standard deviations. Part (a) says that their covariance is 0 which means they are uncorrelated, and that they are not independent.

Recall from lecture that the converse is true though.

(e) Yes. If $\text{Corr}(X,Y) = 0$, then $\text{cov}(X,Y) = 0$. We have $\text{Var}(X+Y) = \text{cov}(X+Y, X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\,\text{cov}(X,Y) = \text{Var}(X) + \text{Var}(Y)$.

(f) Yes. For any values x, y we have $\max(x,y)\min(x,y) = xy$. Thus, $\mathbb{E}[\max(X,Y)\min(X,Y)] = \mathbb{E}[XY]$.

(g) No. You may be tempted to think that because $(\max(x,y), \min(x,y))$ is either (x,y) or (y,x), then $\text{Corr}(\max(X,Y), \min(X,Y)) = \text{Corr}(X,Y)$ because $\text{Corr}(X,Y) = \text{Corr}(Y,X)$. That reasoning is flawed because $(\max(X,Y), \min(X,Y))$ is not always equal to (X,Y) or always equal to (Y,X) and the inconsistency affects the correlation. It is possible for X and Y to be independent while $\max(X,Y)$ and $\min(X,Y)$ are not.

For a concrete example, suppose X is either 0 or 1 with probability $1/2$ each and Y is independently drawn from the same distribution. Then $\text{Corr}(X,Y) = 0$ because X and Y are independent. Even though X never gives information about Y, if you know $\max(X,Y) = 0$ then you know for sure $\min(X,Y) = 0$.

More formally, $\max(X,Y) = 1$ with probability $3/4$ and 0 with probability $1/4$, and $\min(X,Y) = 1$ with probability $1/4$ and 0 with probability $3/4$. This means

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

and

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

Thus,

$$\begin{aligned}\text{cov}(\max(X,Y), \min(X,Y)) &= \mathbb{E}[\max(X,Y)\min(X,Y)] - \frac{3}{16} \\ &= \frac{1}{4} - \frac{3}{16} = \frac{1}{16} \neq 0\end{aligned}$$

We conclude that $\text{Corr}(\max(X,Y), \min(X,Y)) \neq 0 = \text{Corr}(X,Y)$.

4 Dice Games

Note 20

(a) Alice rolls a die until she gets a 1. Let X be the number of total rolls she makes (including the last one), and let Y be the number of rolls on which she gets an even number. Compute $\mathbb{E}[Y \mid X = x]$, and use it to calculate $\mathbb{E}[Y]$.

(b) Bob plays a game in which he starts off with one die. At each time step, he rolls all the dice he has. Then, for each die, if it comes up as an odd number, he puts that die back, and adds a number of dice equal to the number displayed to his collection. (For example, if he rolls a one on the first time step, he puts that die back along with an extra die.) However, if it comes up as an even number, he removes that die from his collection.

Compute the expected number of dice Bob will have after n time steps. (Hint: compute the value of $\mathbb{E}[X_k \mid X_{k-1} = m]$ to derive a recursive expression for X_k, where X_i is the random variable representing the number of dice after i time steps.)

Solution:

(a) Let's compute $\mathbb{E}[Y \mid X = x]$. If Alice makes x total rolls, then before rolling a 1, she makes $x-1$ rolls that are not a 1. Since these rolls are independent, Y follows a binomial distribution with $n = x-1$ and $p = 3/5$, and $\mathbb{E}[Y \mid X = x] = \frac{3}{5}(x-1)$.

Now, we'd like to compute $\mathbb{E}[Y]$. With total expectation, we have

$$\begin{aligned}
\mathbb{E}[Y] &= \sum_x \mathbb{E}[Y \mid X = x]\,\mathbb{P}[X = x] \\
&= \sum_x \frac{3}{5}(x-1)\,\mathbb{P}[X = x] \\
&= \frac{3}{5}\sum_x x \cdot \mathbb{P}[X = x] - \frac{3}{5}\sum_x \mathbb{P}[X = x] \\
&= \frac{3}{5}\mathbb{E}[X] - \frac{3}{5}
\end{aligned}$$

Since X follows a geometric distribution with $p = 1/6$, $\mathbb{E}[X] = 6$, and

$$\mathbb{E}[Y] = \frac{3}{5}\mathbb{E}[X] - \frac{3}{5} = \frac{3}{5}\cdot 6 - \frac{3}{5} = 3.$$

(b) Let X_k be a random variable representing the number of dice after k time steps. In particular, this means that $X_0 = 1$. To compute the number of dice at step k, we first condition on $X_{k-1} = m$. Each one of the m dice is expected to leave behind 2 in its place, since there's a $\frac{1}{2}$ probability that it leaves behind 0 dice, a $\frac{1}{6}$ probability for each of 2, 4, and 6 dice, corresponding to rolling a 1, 3, and 5 respectively.

Therefore, we have $\mathbb{E}[X_k \mid X_{k-1} = m] = 2m$, so with total expectation, we have

$$\begin{aligned}
\mathbb{E}[X_k] &= \sum_m \mathbb{E}[X_k \mid X_{k-1} = m]\,\mathbb{P}[X_{k-1} = m] \\
&= \sum_m 2m \cdot \mathbb{P}[X_{k-1} = m] \\
&= 2\sum_m m \cdot \mathbb{P}[X_{k-1} = m] \\
&= 2\,\mathbb{E}[X_{k-1}]
\end{aligned}$$

This means that we expect to have $\mathbb{E}[X_n] = 2\,\mathbb{E}[X_{n-1}] = 2^2\,\mathbb{E}[X_{n-2}] = \cdots = 2^n\,\mathbb{E}[X_0] = 2^n$ dice.

5 Iterated Expectation

Note 20

In this question, we will try to achieve more familiarity with the law of iterated expectation.

(a) You lost your phone charger! It will take D days for the new phone charger you ordered to arrive at your house (here, D is a random variable). Suppose that on day i, the amount of battery you lose is B_i, where $\mathbb{E}[B_i] = \beta$. Let $B = \sum_{i=1}^{D} B_i$ be the total amount of battery drained between now and when your new phone charger arrives. Apply the law of iterated expectation to show that $\mathbb{E}[B] = \beta\,\mathbb{E}[D]$.

(Here, the law of iterated expectation has a very clear interpretation: the amount of battery you expect to drain is the average number of days it takes for your phone charger to arrive, multiplied by the average amount of battery drained per day.)

(b) Consider now the setting of independent Bernoulli trials, each with probability of success p. Let S_i be the number of successes in the first i trials. Compute $\mathbb{E}[S_m \mid S_n = k]$.

(You will need to consider three cases based on whether $m > n$, $m = n$, or $m < n$. Try using your intuition rather than proceeding by calculations.)

Solution:

(a) This is simply Wald's Identity from lecture. Condition on $D = d$; then $B = \sum_{i=1}^{d} B_i$ and

$$\mathbb{E}[B \mid D = d] = \sum_{i=1}^{d} \mathbb{E}[B_i] = \beta d.$$

Therefore, $\mathbb{E}[B \mid D] = \beta D$ and by the law of iterated expectation, $\mathbb{E}[B] = \mathbb{E}[\mathbb{E}[B \mid D]] = \beta\,\mathbb{E}[D]$.

(b) Suppose $m > n$. Then we already know that the first n trials resulted in k successes, and there are $m - n$ trials for which we do not know the outcome. Each of these $m - n$ trials has probability of success p, so we expect $(m-n)p$ additional successes. Hence, $\mathbb{E}[S_m \mid S_n = k] = k + (m-n)p$.

Next, consider when $m = n$. Here, $\mathbb{E}[S_m \mid S_n = k] = k$.

Finally, suppose that $m < n$. In n trials, we have k successes, and due to symmetry, we expect the k successes to be distributed uniformly among the n trials. In particular, if we look at the first m trials only, then we expect a proportion m/n of the total successes to be distributed among the first m successes. Therefore, $\mathbb{E}[S_m \mid S_n = k] = mk/n$.