11

Random Variables II, Variance, LLN, Concentration Inequalities

Computer Science Mentors 70

Nov 16 - Nov 20
Fall 2020

Prepared by: Aishani Sil, Austin Lei, Agnibho Roy, Debayan Bandyopadhyay, Abinav Routhu

I Random Variables II

1. **E[*levator*]**
A building has n floors numbered $1, 2, \ldots, n$, plus a ground floor G. At the ground floor, m people get on the elevator together, and each gets off at a uniformly random one of the n floors (independently of everybody else).

(a) What is the expected value of the last floor someone gets off at? For example, if people leave at floors $1, 2$, and 5, Floor 5 is the last floor.

Solution: Let us define random variable G to be the last floor where someone gets off. By the definition of expected value,

$$E[G] = \sum_{i=1}^{n} i \times P_i,$$

where P_i is the probability that everyone has gotten off before or at floor i, and someone gets off at floor i (this latter constraint is needed, otherwise G would equal $i-1$).

$$P_i = \left(\frac{i}{n}\right)^m - \left(\frac{i-1}{n}\right)^m.$$

We find this by starting with the probability that everyone gets off at or before floor i, and subtracting the probability that no one has gotten off at floor i.
Thus,

$$\begin{aligned}
E[G] &= \sum_{i=1}^{n} i\left(\left(\frac{i}{n}\right)^m - \left(\frac{i-1}{n}\right)^m\right) \\
&= 1\left(\left(\frac{1}{n}\right)^m - \left(\frac{0}{n}\right)^m\right) + 2\left(\left(\frac{2}{n}\right)^m - \left(\frac{1}{n}\right)^m\right) + 3\left(\left(\frac{3}{n}\right)^m - \left(\frac{2}{n}\right)^m\right) + \cdots + +n\left(\left(\frac{n}{n}\right)^m - \left(\frac{n-1}{n}\right)^m\right) \\
&= (1-2)\left(\frac{1}{n}\right)^m + (2-3)\left(\frac{2}{n}\right)^m + (3-4)\left(\frac{3}{n}\right)^m + \cdots + ((n-1)-n)\left(\frac{n-1}{n}\right)^m + n\left(\frac{n}{n}\right)^m \\
&= n - \sum_{i=1}^{n-1}\left(\frac{i}{n}\right)^m.
\end{aligned}$$

(b) What is the expected number of times the elevator stops (not including the ground floor)?

Solution: Let F_i = elevator stopped at floor i.

$$P(F_i = 1) = 1 - P(\text{ no one gets off at floor } i\text{ }) = 1 - \left(\tfrac{n-1}{n}\right)^m.$$

Let F represent the number of floors the elevator is stopping at (the original quantity we are looking for).

$$F = F_1 + F_2 + \ldots F_n, \text{ so } E[F] = E[F_1] + E[F_2] + \ldots E[F_n] = n(1 - (\tfrac{n-1}{n})^m).$$

Notice how we can use linearity of independence even though these individual probabilities are not independent– neat!

2. **Chaotic Santa**
(Fall '17 Disc) A delivery guy is out delivering n packages to n customers, where $n \in \mathbb{N}$, $n > 1$. Not only does he hand a random package to each customer, he opens the package before delivering it with probability $\frac{1}{2}$.

(a) What is the expected number of customers that get their own package unopened?

Solution: Define an indicator variable C_i to be 1 if customer i gets their own package unopened. Since it is an indicator and follows a Bernoulli distribution,

$$E[C_i] = P(C_i) = \tfrac{1}{2} \cdot \tfrac{1}{n}.$$

Let C = number of customers that get their own package unopened;

$$C = C_1 + C_2 + \ldots + C_n = n \cdot \tfrac{1}{2} \cdot \tfrac{1}{n} = \tfrac{1}{2}.$$

(b) What is the variance for the random variable above?

Solution:
Remember that $\text{Var}(C) = E[C^2] - E[C]^2$. We can get the second term by squaring part (a), so we must still find $E[C^2]$.

$$E[C^2] = E[(C_1 + C_2 + \ldots + C_n)^2] = E[\textstyle\sum_{i,j} C_i C_j] = \sum_{i,j} E[C_i C_j].$$

For this kind of setup, we split up the summation into cases where $i = j$ and cases where $i \neq j$, so our expression is

$$E[C^2] = \textstyle\sum_i E[C_i^2] + \sum_{i \neq j} E[C_i C_j]$$
$$E[C_i^2] = E[C_i] = \tfrac{1}{2n}\,;\, E[C_i C_j] = P(C_i C_j = 1)$$

We must now find the probability that the two selected people both get their own package unopened. Note that we cannot use linearity of expectation here, so we need to count this as one event.

$$P(C_i C_j = 1) =$$
$\frac{1}{2n}$ (first person gets their package unopened)

$\cdot$

$\frac{1}{2(n-1)}$ (second person gets their package unopened after the first person's stuff arrives)

So, we add up n instances of the first part of the summation, and $n(n-1)$ instances of the second part for each combination where $i \neq j$...

$$E[C^2] = n(\tfrac{1}{2n}) + n(n-1)\tfrac{1}{2n \cdot 2(n-1)} = \tfrac{3}{4}$$
$$\text{Var}(C) = E[C^2] - E[C]^2 = \tfrac{3}{4} - \tfrac{1}{4} = \tfrac{1}{2}.$$

II Variance

3. **Introduction to Variance**
Let us say that we are dealing with a biased die and we want to know how often my roll varies from turn to turn. Consider the following 6-length tuple assigning probabilities to each of the 6 rolls on the die, whose distribution we name X:

$$(p_1, p_2, p_3, p_4, p_5, p_6)$$

(a) Find the mean and variance for the die with distribution $(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$

Solution: The expected value is:

$$E[X] = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = \frac{7}{2} = 3.5$$

The variance is:

$$\text{Var}(X) = E[X^2] - E[X]^2 = \frac{1}{6}(1+4+9+16+25+36) - 3.5^2 = \frac{91}{6} - \frac{49}{4} = \frac{35}{12}$$

(b) Find the mean and variance for the die with distribution $(\frac{1}{3}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{3})$

Solution: The expected value is:

$$E[X] = \frac{1}{12}(2+3+4+5) + \frac{1}{3}(1+6) = \frac{7}{2} = 3.5$$

The variance is:

$$\text{Var}(X) = E[X^2] - E[X]^2 = \frac{1}{12}(4+9+16+25) + \frac{1}{3}(1+36) - 3.5^2 = \frac{101}{6} - \frac{49}{4} = \frac{55}{12}$$

(c) Give an intuitive explanation for the difference between the results in part(a) and part (b)

Solution: Although the means of the two die are the same, we expect there to be a higher variance because the probabilities of a 1 and 6 are higher in part (b) than part (a).

4. **Money Balls**
We are drawing dollar bills out of a bag consisting of R red balls and B blue balls. We get 1 dollar for every blue ball, but whenever we draw a red ball, the game stops and we retrieve the amount of money we have made so far.

(a) Find the expected value of the amount of money made from this game.

Solution: We can create an indicator, X_i, which represents the indicator that the i^{th} blue ball appears before the first red ball. The total number of balls that we draw before we draw this blue ball represents the total amount of money made in the game, which can be represented as:

$$X = \sum_{i=1}^{B} X_i$$

The probability that the event represented by the indicator occurs is $\frac{1}{R+1}$. Consider drawing all balls out of the bag. Out of the $R+1$ choices (in between all red balls) that the i^{th} blue ball can be placed, it must be placed before all of them. Using the linearity of expectation:

$$X = \sum_{i=1}^{B} E[X_i] = B * E[X_1] = \frac{B}{R+1}$$

(b) Find the variance of the amount of money made in this game.

Solution: To compute the variance, we have: $\text{Var}(X) = E[X^2] - E[X]^2$. We first find:

$$\begin{aligned}
E[X^2] &= E[(X_1 + X_2 + \cdots + X_B)^2] \\
&= \sum_{i=1}^{B} E[X_i^2] + \sum_{i \neq j} E[X_i X_j] \\
&= BE[X_i^2] + B(B-1)E[X_1 X_2] \\
&= B(\frac{1}{R+1}) + B(B-1)\frac{2}{(R+1)(R+2)} \\
&= \frac{B(R+2) + 2B(B-1)}{(R+1)(R+2)}
\end{aligned}$$

Keep in mind that the square of an indicator is the same as the indicator itself, which is the simplification in line 3 of the above simplification. Furthermore, $E[X_i X_j] = \frac{2}{(R+1)(R+2)}$ because after the i^{th} blue ball, we have a total of $(R+1)+1$ places to place the jth blue ball, and it can either go at the beginning or after the i^{th} blue ball, leading to 2 spots. We simplify the summation using the linearity of expectation. Combining everything, we have:

$$\text{Var}(X) = \frac{B(R+2) + 2B(B-1)}{(R+1)(R+2)} - \left(\frac{B}{R+1}\right)^2$$

III Concentration Inequalities

5. **Introduction to Markov and Chebyshev Bounds**

(a) Let X be a nonegative random variable with $E[X] = 5$. Using Markov's Inequality, determine an upper bound for the probability that $X > 10$.

Solution: By Markov's Inequality,

$$P(X > 10) \leq \frac{E[X]}{10} = \frac{1}{2}$$

(b) Let X be a random variable with $E[X] = 3$ and $\text{Var}(X) = 4$. Using Chebyshev's Inequality, determine an upper bound for the probability that $X > 13$.

Solution: By Chebyshev's Inequality, we have that

$$\begin{aligned}
P(X > 13) &\leq P(X - 3 > 10) \\
&\leq P(|X - 3| > 10) \\
&\leq \frac{\text{Var}(X)}{10^2} \\
&= \frac{1}{25}.
\end{aligned}$$

6. **Strengthening Inequalities**

Imagine that we have a biased coin with a probability of getting heads as 0.3. This coin is flipped 10 times.

(a) Find a bound for the probability that the coin lands on heads at least 8 times with Markov's inequality. How good is this bound compared to the true probability?

Solution: By Markov's inequality, we have:

$$P(X \geq 8) \leq \frac{E[X]}{8} = \frac{(10)(0.3)}{8} = \frac{3}{8} = 0.375$$

However, we know that by the binomial distribution, the true value is:

$$P(X \geq 8) = \sum_{k=8}^{10} \binom{10}{k} (0.3)^k (0.7)^{10-k} = 0.0016$$

This value is way lower than the upper bound estimate using Markov's inequality, making it not that good of a bound.

(b) Now, find the same bound (coin lands on heads at least 8 times) using Chebyshev's inequality.

Solution: First, $\text{Var}(X) = np(1-p) = 10(0.3)(0.7) = 2.1$. By Chebyshev's inequality:

$$P(X \geq 8) \leq P(X \geq 8 \cap X \leq -2) = P(|X-3| \geq 5) \leq \frac{\text{Var}(X)}{5^2} = \frac{2.1}{25} = 0.084$$

(c) Is this bound better? Can we divide this bound by 2 to get the bound on the tail we are interested in?

Solution: Chebyshev's bound seems to be much better since $0.084 < 0.375$. We cannot divide this bound by 2 because the tails are not symmetric i.e. the probability of 2 heads and 2 tails are not equal for a biased coin.

IV LLN

7. **Introduction to LLN**

Leanne has a weighted coin that shows up heads with probability $\frac{4}{5}$ and tails with probability $\frac{1}{5}$. Leanne flips the coin 100 times, and computes X, the average number of coins that show up heads.

(a) What is $E[X]$?

Solution: Let I_i be the indicator variable for the event that the ith coin flip is heads. We can express $X = \frac{I_1 + \cdots + I_{100}}{100}$. Note that $E[I_i] = \frac{4}{5}$ for all i, so by Linearity of Expectation we have that

$$E[X] = E\left[\frac{I_1 + \cdots + I_{100}}{100}\right] = \frac{1}{100}E[I_1] + \cdots + \frac{1}{100}E[I_{100}] = \frac{1}{100} \cdot 100 \cdot \frac{4}{5} = \frac{4}{5}.$$

(b) What is $\text{Var}(X)$?

Solution: Note that for any i, I_i is a Bernoulli random variable with parameter $p = \frac{4}{5}$, so $\text{Var} I_i = \frac{4}{5}\left(1 - \frac{4}{5}\right) = \frac{4}{25}$.

Since we can express $X = \frac{I_1+\cdots+I_{100}}{100}$, and the I_i are independent, we have that

$$\begin{aligned}
\text{Var}(X) &= \text{Var}\left(\frac{I_1 + \cdots + I_{100}}{100}\right) \\
&= \left(\frac{1}{100}\right)^2 \text{Var}\left(I_1 + \cdots + I_{100}\right) \\
&= \left(\frac{1}{100}\right)^2 \text{Var}\left(I_1 + \cdots + I_{100}\right) \\
&= \left(\frac{1}{100}\right)^2 \left(\text{Var}(I_1) + \cdots + \text{Var}(I_{100})\right) \\
&= \frac{100 \cdot \frac{4}{25}}{10000} \\
&= \frac{1}{625}.
\end{aligned}$$

(c) Suppose Leanne flips n coins instead of 100. What does the LLN tell us about X?

Solution: The LLN tells us that for any $\varepsilon > 0$, the probability that X is within ε of 0.8 goes to 1 as $n \to \infty$. This can be seen for $n = 100$; the variance is small, so it is unlikely for X to be very far from 0.8.