CMPT 210: Probability and Computation

Lecture 16

Sharan Vaswani

July 8, 2022

Recap

Independence: R_1 and R_2 are independent r.v. if for *all* $x_1 \in \text{Range}(R_1)$ and $x_2 \in \text{Range}(R_2)$, events $[R_1 = x_1]$ and $[R_2 = x_2]$ are independent i.e.
$\Pr[(R_1 = x_1) \cap (R_2 = x_2)] = \Pr[(R_1 = x_1)] \Pr[(R_2 = x_2)]$

If R_1 and R_2 are independent r.v., $\mathbb{E}[R_1 R_2] = \mathbb{E}[R_1]\,\mathbb{E}[R_2]$, whereas $\mathbb{E}[R_1 + R_2] = \mathbb{E}[R_1] = \mathbb{E}[R_2]$ holds regardless of independence.

Variance: Standard way to measure the deviation from the mean. For r.v. X,
$\text{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \sum_{x \in \text{Range}(X)} (x - \mu)^2 \Pr[X = x]$, where $\mu := \mathbb{E}[X]$.

Alternate Definition: $\text{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.

If $X \sim \text{Ber}(p)$, $\text{Var}[X] = p(1 - p)$.

If $X \sim \text{Uniform}(\{v_1, v_2, \dots v_n\})$, $\text{Var}[X] = \frac{[v_1^2 + v_2^2 + \dots v_n^2]}{n} - \left(\frac{[v_1 + v_2 + \dots v_n]}{n}\right)^2$.

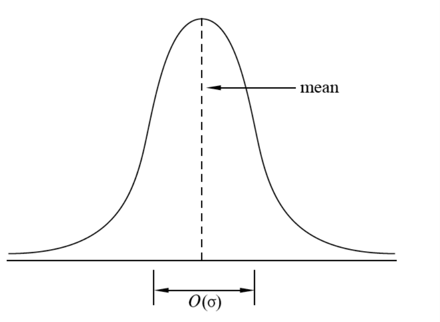
If $X \sim \text{Geo}(p)$, $\text{Var}[X] = \frac{1-p}{p^2}$.

Standard Deviation

Standard Deviation: For r.v. X, the standard deviation in X is defined as:

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

Standard deviation has the same units as expectation.



Standard deviation for a "bell"-shaped distribution indicates how wide the "main part" of the distribution is.

Properties of Variance

For constants a, b and r.v. R, $\text{Var}[aR + b] = a^2\text{Var}[R]$.

$$\begin{aligned}
\text{Var}[aR + b] &= \mathbb{E}[(aR + b)^2] - (\mathbb{E}[aR + b])^2 = \mathbb{E}[a^2R^2 + 2abR + b^2] - (\mathbb{E}[aR] + \mathbb{E}[b])^2 \\
&= (a^2\mathbb{E}[R^2] + 2ab\mathbb{E}[R] + b^2) - (a\mathbb{E}[R] + b)^2 \\
&= (a^2\mathbb{E}[R^2] + 2ab\mathbb{E}[R] + b^2) - (a^2(\mathbb{E}[R])^2 + 2ab\mathbb{E}[R] + b^2) \\
&= a^2 \left[\mathbb{E}[R^2] - (\mathbb{E}[R])^2\right]
\end{aligned}$$

$$\implies \text{Var}[aR + b] = a^2\text{Var}[R]$$

Similarly, for the standard deviation,

$$\sigma_{aR+b} = \sqrt{\text{Var}[aR + b]} = \sqrt{a^2\text{Var}[R]} = |a|\,\sigma_R$$

Note the difference from the property of expectation,

$$\mathbb{E}[aR + b] = a\mathbb{E}[R] + b$$

Q: What is $\mathbb{E}[\mathbb{E}[R]]$ equal to? Ans: $\mathbb{E}[R]$

Properties of Variance

Recall that for r.v's R and S, $\mathbb{E}[R+S] = \mathbb{E}[R] + \mathbb{E}[S]$.

In general, such a property is not true for the variance, i.e. variance of a sum is not necessarily equal to the sum of the variances.

However, when the r.v's are *independent*, then, $\text{Var}[R+S] = \text{Var}[R] + \text{Var}[S]$.

$$\begin{aligned}
\text{Var}[R+S] &= \mathbb{E}[(R+S)^2] - (\mathbb{E}[R+S])^2 = \mathbb{E}[R^2 + S^2 + 2RS] - (\mathbb{E}[R] + \mathbb{E}[S])^2 \\
&= \mathbb{E}[R^2 + S^2 + 2RS] - [(\mathbb{E}[R])^2 + (\mathbb{E}[S])^2 - 2\mathbb{E}[R]\,\mathbb{E}[S]] \\
&= [\mathbb{E}[R^2] - (\mathbb{E}[R])^2] + [\mathbb{E}[S^2] - (\mathbb{E}[S])^2] + 2(\mathbb{E}[RS] - \mathbb{E}[R]\,\mathbb{E}[S]) \\
&= \text{Var}[R] + \text{Var}[S] + 2(\mathbb{E}[RS] - \mathbb{E}[R]\,\mathbb{E}[S])
\end{aligned}$$

Recall that if r.v. are independent, $\mathbb{E}[RS] = \mathbb{E}[R]\,\mathbb{E}[S]$,

$$\implies \text{Var}[R+S] = \text{Var}[R] + \text{Var}[S]$$

Variance

Random variables $R_1, R_2, R_3, \dots R_n$ are *pairwise* independent if for any pair R_i and R_j, for $x \in \text{Range}(R_i)$ and $y \in \text{Range}(R_j)$, events $\Pr[R_i = x]$ and $\Pr[R_j = y]$ are pairwise independent implying that $\Pr[(R_i = x) \cap (R_j = y)] = \Pr[R_i = x] \Pr[R_j = y]$.

Using a similar derivation as before, we can prove that for any pair of pairwise independent r.v's, R_i and R_j, $\mathbb{E}[R_i R_j] = \mathbb{E}[R_i]\mathbb{E}[R_j]$.

$$\begin{aligned}
\text{Var}[R_1 + R_2 + \dots R_n] &= \mathbb{E}[(R_1 + R_2 + \dots R_n)^2] - (\mathbb{E}[R_1 + R_2 + \dots R_n])^2 \\
&= \sum_{i=1}^{n} [\mathbb{E}[R_i^2] - (\mathbb{E}[R_i])^2] + 2 \sum_{i,j \mid 1 \leq i < j \leq n} [\mathbb{E}[R_i R_j] - \mathbb{E}[R_i]\mathbb{E}[R_j]] \\
\text{Var}[R_1 + R_2 + \dots R_n] &= \sum_{i=1}^{n} \text{Var}[R_i] \qquad \text{(Since the r.v's are pairwise independent)}
\end{aligned}$$

Importantly, we do not require the r.v's to be mutually independent. Similar to events, mutual independence $\implies$ pairwise independence, but pairwise independence $\not\Rightarrow$ mutual independence.

Variance - Examples

Q: If $R \sim \text{Bin}(n, p)$, calculate $\text{Var}[R]$.

Similar to Slide 13 of Lecture 13, we define R_i to be the indicator random variable that we get a heads in toss i of the coin. Recall that R is the random variable equal to the number of heads in n tosses.

Hence,

$$R = R_1 + R_2 + \ldots + R_n \implies \text{Var}[R] = \text{Var}[R_1 + R_2 + \ldots + R_n]$$

Since $R_1, R_2, \ldots, R_n$ are independent indicator random variables, and hence pairwise independent,

$$\text{Var}[R] = \text{Var}[R_1] + \text{Var}[R_2] + \ldots + \text{Var}[R_n]$$

Since the variance of an indicator (Bernoulli) r.v. is $p(1-p)$,

$$\text{Var}[R] = n\,p\,(1-p).$$

Back to throwing dice

Q: We throw a standard dice, and define a random variable R which is equal to 1 if we get an even number and 0 otherwise.What is $\mathrm{Var}[R]$?

Ans: $\frac{1}{4}$

Q: We throw 10 independent dice and define R to be the random variable equal to the number of dice that have an even number. What is $\mathrm{Var}[R]$?

Ans: $10\,\frac{1}{2}\,\frac{1}{2} = 2.5$

Q: We repeatedly and independently throw the dice until we get an even number. We define a random variable R equal to the number of throws we need to get an even number. What is $\mathrm{Var}[R]$?

Ans: $\frac{1-1/2}{(1/2)^2} = 2$

Questions?

Matching Birthdays

Q: In a class of n students, what is the probability that two students share the same birthday? Assume that (i) each student is equally likely to be born on any day of the year, (ii) no leap years and (iii) student birthdays are independent of each other.

For $d := 365$,

$$\Pr[\text{two students share the same birthday}] = 1 - \frac{d \times (d-1) \times (d-2) \times \ldots (d-(n-1))}{d^n}$$

Q: On average, how many pairs of students have matching birthdays?

Define M to be the number of pairs of students with matching birthdays. For a fixed ordering of the students, let $X_{i,j}$ be the indicator r.v. corresponding to the event $E_{i,j}$ that the birthdays of students i and j match. Hence,

$$M = \sum_{i,j|1 \leq i < j \leq n} X_{i,j} \implies \mathbb{E}[M] = \mathbb{E}[\sum_{i,j|1 \leq i < j \leq n} X_{i,j}] = \sum_{i,j|1 \leq i < j \leq n} \mathbb{E}[X_{i,j}] = \sum_{i,j|1 \leq i < j \leq n} \Pr[E_{i,j}]$$

(Linearity of expectation)

Matching Birthdays

For a pair of students i, j, let B_i be the r.v. equal to the day of student i's birthday. Range(B_i) = {1,2,. . . , 365} and for all $k \in [365]$, $\Pr[B_i = k] = 1/d$.

$$E_{i,j} = (B_i = 1 \cap B_j = 1) \cup (B_i = 2 \cap B_j = 2) \cup \dots$$

$$\implies \Pr[E_{i,j}] = \sum_{k=1}^{d} \Pr[B_i = k \cap B_j = k] = \sum_{k=1}^{d} \Pr[B_i = k] \Pr[B_j = k] = \sum_{k=1}^{d} \frac{1}{d^2} = \frac{1}{d}$$

$$\implies \mathbb{E}[M] = \sum_{i,j|1 \leq i < j \leq n} \Pr[E_{i,j}] = \frac{1}{d} \sum_{i,j|1 \leq i < j \leq n} (1) = \frac{1}{d}[(n-1) + (n-2) + \dots + 1] = \frac{n(n-1)}{2d}$$

Hence, in our class of 48 students, on average, there are $\frac{(24)(47)}{365} = 3.09$ students with matching birthdays.

Matching Birthdays

Q: Are the $X_{i,j}$ mutually independent?

No, because if $X_{i,j} = 1$ and $X_{j,k} = 1$, then,
$\Pr[X_{i,k} = 1 | X_{j,k} = 1 \cap X_{i,j} = 1] = 1 \neq \frac{1}{d} = \Pr[X_{i,k} = 1]$.

Q: Are the $X_{i,j}$ pairwise independent?

Yes, because for all i, j and i', j' (where $i \neq i'$), $\Pr[X_{i,j} = 1 | X_{i',j'} = 1] = \Pr[X_{i,j} = 1]$ because if students i' and j' have matching birthdays, it does not tell us anything about whether i and j have matching birthdays.

Matching Birthdays

Q: If M is the r.v. equal to the number of pairs of students with matching birthdays, calculate $\text{Var}[M]$.

$$\text{Var}[M] = \text{Var}[\sum_{i,j|1\leq i<j\leq n} X_{i,j}]$$

Since $X_{i,j}$ are pairwise independent, the variance of the sum is equal to the sum of the variance.

$$\implies \text{Var}[M] = \sum_{i,j|1\leq i<j\leq n} \text{Var}[X_{i,j}] = \sum_{i,j|1\leq i<j\leq n} \frac{1}{d}\left(1-\frac{1}{d}\right) = \frac{1}{d}\left(1-\frac{1}{d}\right)\frac{n(n-1)}{2}$$

(Since $X_{i,j}$ is an indicator (Bernoulli) r.v.)

Hence, in our class of 48 students, the standard deviation for the matching birthdays is equal to $\sqrt{\frac{(24)(47)}{365}\frac{364}{365}} \approx 1.75$.

Questions?