Lisa Yan and Jerry Cain
CS109

Lecture Notes #6
September 25, 2020

Random Variables

Based on a chapter by Chris Piech and Lisa Yan

Pre-recorded lecture: Section 1 and Section 2 (definitions only).
In-lecture: Section 2: examples.
Covered later: Section 3 (as part of pre-lecture Lecture 7)
Other material: Section 4: Proofs.

A **random variable** (RV) is a variable that probabilistically takes on different values. You can think of an RV as being like a variable in a programming language. They take on values, have types and have domains over which they are applicable. We can define events that occur if the random variable takes on values that satisfy a numerical test (e.g., does the variable equal 5? is the variable less than 8?). We often need to know the probabilities of such events.

As an example, let's say we flip three fair coins. We can define a random variable Y to be the total number of "heads" on the three coins. We can ask about the probability of Y taking on different values using the following notation:

- $P(Y = 0) = 1/8$ (T, T, T)
- $P(Y = 1) = 3/8$ (H, T, T), (T, H, T), (T, T, H)
- $P(Y = 2) = 3/8$ (H, H, T), (H, T, H), (T, H, H)
- $P(Y = 3) = 1/8$ (H, H, H)
- $P(Y \geq 4) = 0$

Even though we use the same notation for random variables and for events (both use capital letters), they are distinct concepts. An event is a situation, a random variable is an object. The situation in which a random variable takes on a particular value (or range of values) is an event. When possible, I will try to use letters E, F, G for events and X, Y, Z for random variables.

Using random variables is a convenient notation that assists in decomposing problems. There are many different types of random variables (indicator, binary, choice, Bernoulli, etc). The two main families of random variable types are discrete and continuous. For now we are going to develop intuition around discrete random variables.

1 Probability Mass Function

For a discrete random variable, the most important thing to know is the probability that the random variable will take on each of its possible values. The **probability mass function** (PMF) of a random variable is a function that maps possible outcomes of a random variable to the corresponding probabilities. Because it is a function, we can plot PMF graphs where the x-axis contains the values that the random variable can take on and the y-axis contains the probability of the random variable taking on said value:

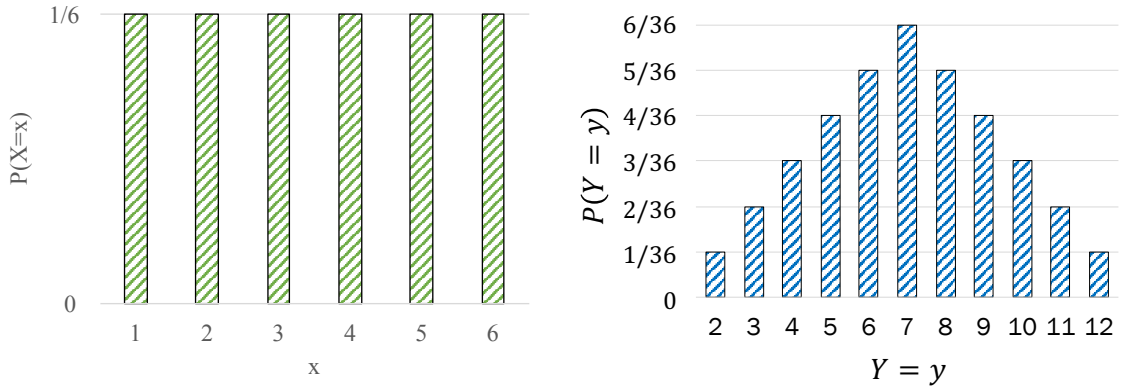


Figure 1: On the left, the PMF of a single 6 sided die roll. On the right, the PMF of the sum of two dice rolls.

There are many ways that probability mass functions can be specified. We can draw a graph. We can build a table (or for you CS folks, a `map`/`HashMap`/`dict`) that lists out all the probabilities for all possible events. Or we could write out a mathematical expression.

For example, consider the random variable Y which is the sum of two dice rolls. The probability mass function can be defined by the graph on the right of Figure 1. It can also be defined using the equation:

$$p_Y(y) = \begin{cases} \frac{y-1}{36} & \text{if } y \in \mathbb{Z}\,,\, 2 \leq y \leq 7 \\ \frac{13-y}{36} & \text{if } y \in \mathbb{Z}\,,\, 8 \leq y \leq 12 \\ 0 & \text{else} \end{cases}$$

The probability mass function, $p_Y(y)$, defines the probability of Y taking on the value y. The new notation $p_Y(y)$ is simply different notation for writing $P(Y = y)$. Using this new notation makes it more apparent that we are specifying a function. Try a few values of y, and compare the value of $p_Y(y)$ to the graph in Figure 1. They should be the same.

Cumulative Distribution Function

The **cumulative distribution function** (CDF) of a random variable X is a function F specified as $F(a) = P(X \leq a)$, the probability that X takes on a value less than or equal to some value a. For discrete random variable X with PMF $p(x)$:

$$F(x) = P(X \leq a) = \sum_{\text{all} x \leq a} p(x)$$

The CDF's range is 0 to 1: For example, let X be the outcome of a 6-sided die roll. $F(-2) = 0$ (it is impossible to roll less than or equal to -2), $F(6) = 1$ (it will always be the case that you will roll less than or equal to 6), and $F(100) = 1$ (it will always be the case that you roll something less than or equal to 100).The usefulness of cumulative distribution functions will come when we talk about continuous random variables later this week. For now, just note that both CDFs and PMFs represent probabilities!

2 Expectation

A useful piece of information about a random variable is the average value of the random variable over many repetitions of the experiment it represents. This average is called the **expectation**.

The **expectation** of a discrete random variable X is defined as:

$$E[X] = \sum_{x:P(x)>0} xP(x)$$

It goes by many other names: *mean*, *expected value*, *weighted average*, *center of mass*, *1st moment*.

Example: Dice expectation

The random variable X represents the outcome of one roll of a six-sided die. What is the $E[X]$? This is the same as asking for the average value of a die roll.

$$E[X] = 1(1/6) + 2(1/6) + 3(1/6) + 4(1/6) + 5(1/6) + 6(1/6) = 7/2$$

Example: Expected class size

A school has 3 classes with 5, 10, and 150 students. Each student is only in one of the three classes. If we randomly choose a class with equal probability and let X = the size of the chosen class:

$$\begin{aligned} E[X] &= 5(1/3) + 10(1/3) + 150(1/3) \\ &= 165/3 = 55 \end{aligned}$$

However, if instead we randomly choose a student with equal probability and let Y = the size of the class the student is in:

$$\begin{aligned} E[Y] &= 5(5/165) + 10(10/165) + 150(150/165) \\ &= 22635/165 \approx 137 \end{aligned}$$

2.1 Properties of Expectation

Expectations preserve linearity. Mathematically, this means that

$$E[aX + bY + c] = aE[X] + bE[Y] + c$$

There are two parts to this: First, linearly scaling a random variable will linearly scale expectation. Second, expectation of a sum is the sum of expectations. So if you have an expectation of a sum of quantities, this is equal to the sum of the expectations of those quantities. We will return to the implications of this very useful fact later in the course.

One can also calculate the expected value of a function $g(X)$ of a random variable X when one knows the probability distribution of X but one does not explicitly know the distribution of $g(X)$:

$$E[g(X)] = \sum_x g(x) \cdot p_X(x)$$

This identity has the humorous name of "the Law of the Unconscious Statistician" (LOTUS), for the fact that even statisticians are known—perhaps unfairly—to ignore the difference between this identity and the basic definition of expectation (the basic definition doesn't have a function g).

We can use this to compute, for example, the expectation of the square of a random variable (called the *second moment*):

$$\begin{aligned} E[X^2] &= E[g(X)] && \text{where } g(X) = X^2 \\ &= \sum_x g(x) \cdot p_X(x) && \text{by LOTUS} \\ &= \sum_x x^2 \cdot p_X(x) && \text{definition of } g \end{aligned}$$

Proofs of these properties are provided in the last section.

Example: Absolute Value

Let X be a discrete random variable where $P(X = x) = \frac{1}{3}$ for $x \in \{-1, 0, 1\}$. Suppose we define random variable $Y = |X|$. What is $E[Y]$?

Solution 1: We compute $E[Y]$ by definition, which requires us to first determine the PMF of Y. $P(Y = 0) = \frac{1}{3}$ and $P(Y = 1) = \frac{2}{3}$, and thus $E[Y] = \sum_y yP(Y = y) = \frac{1}{3} \cdot 0 + \frac{2}{3} \cdot 1 = \frac{2}{3}$.

Solution 2: We compute $E[Y]$ using LOTUS, by treating Y as a function of X and therefore $E[Y] = \sum_x |x|P(X = x) = \frac{1}{3} \cdot |-1| + \frac{1}{3} \cdot |0| + \frac{1}{3} \cdot |1| = \frac{2}{3}$.

Note that by using LOTUS, we do not have to explicitly compute a distribution (i.e., a PMF) on Y; we can simply use the known distribution of X. This property of expectation is especially useful when we must compute the expectation of a complex function of X. See the St. Petersburg Paradox example.

Example: St. Petersburg Paradox

Consider a game played with a fair coin which comes up heads with p = 0.5. Let N = the number of coin flips before the first "tails". In this game you win $\$2^N$. How many dollars do you expect to win? Let W be a random variable which represents your winnings.

Note that $W = 2^N$ and $P(N = n) = \left(\frac{1}{2}\right)^{n+1}$. We do not explicitly compute the PMF of W. Instead, using LOTUS,

$$E[W] = E[2^N] = \left(\frac{1}{2}\right)^1 2^0 + \left(\frac{1}{2}\right)^2 2^1 + \left(\frac{1}{2}\right)^3 2^2 + \left(\frac{1}{2}\right)^4 2^3 + \cdots = \sum_{i=0}^{\infty} \left(\frac{1}{2}\right)^{i+1} 2^i = \sum_{i=0}^{\infty} \frac{1}{2} = \infty$$

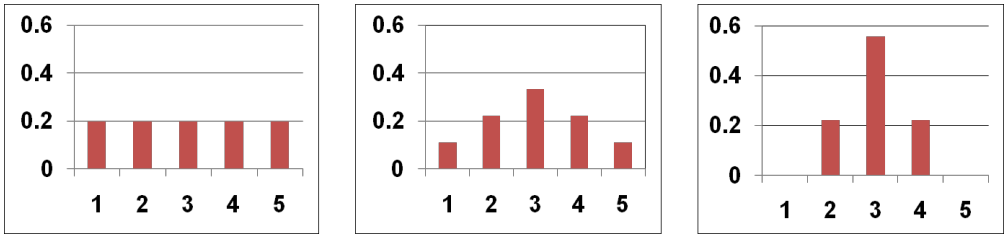
This example is nicknamed a paradox because consider the more realistic scenario, where the game dealer has a maximum amount of money (say, \$65,536), and you cannot win more than this amount. If you are projected to win more, then the dealer goes home, you get kicked out of the hall, and you win nothing. In this case, when $N = 16$, you win $W = 2^{16} = 65,536$, and if $N \geq 17$ you win nothing ($W = 0$). Using LOTUS, where $k = \log_2(65,536) = 16$:

$$E[W] = \left(\frac{1}{2}\right)^1 2^0 + \left(\frac{1}{2}\right)^2 2^1 + \left(\frac{1}{2}\right)^3 2^2 + \cdots + \left(\frac{1}{2}\right)^k 2^i + 0 = \sum_{i=0}^{k} \left(\frac{1}{2}\right)^{i+1} 2^i = \sum_{i=0}^{16} \frac{1}{2} = 8.5$$

That's much less than infinite winnings!

3 Variance

Expectation is a useful statistic, but it does not give a detailed view of the probability mass function. Consider the following 3 distributions (PMFs)



All three have the same expected value, $E[X] = 3$, but the "spread" in the distributions is quite different. Variance is a formal quantification of "spread". There is more than one way to quantify spread; variance uses the average square distance from the mean.

The variance of a discrete random variable X with expected value μ is:

$$\text{Var}(X) = E[(X - \mu)^2]$$

Standard deviation is the square root of variance: $\text{SD}(X) = \sqrt{\text{Var}(X)}$. Intuitively, standard deviation is a kind of average distance of a sample to the mean. (Specifically, it is a *root-mean-square* [RMS] average.) Variance is the square of this average distance.

Properties of Variance

When computing the variance, we often use a different form of the same equation:

$$\text{Var}(X) = E[X^2] - E[X]^2$$

Another useful identity for variance is that $\text{Var}(aX + b) = a^2 \text{Var}(X)$. **Note that this implies that variance is nonlinear**: Adding a constant doesn't change the "spread". Multiplying by a constant (negative or positive) changes the "spread" quadratically.

Proofs of these two properties are provided in the last section.

Example: Dice Variance

Let X = the value on one roll of a 6 sided die. Recall that $E[X] = 7/2$. What is $\text{Var}(X)$?

Answer: First, we can calculate $E[X^2]$:

$$E[X^2] = (1^2)\frac{1}{6} + (2^2)\frac{1}{6} + (3^2)\frac{1}{6} + (4^2)\frac{1}{6} + (5^2)\frac{1}{6} + (6^2)\frac{1}{6} = \frac{91}{6}$$

We can then use the expectation formula for variance:

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

4 Proofs

4.1 Linearity of Expectation

For a single random variable X, $E[aX + b] = aE[X] + b$.

Proof: Let X be a random variable with PMF $p(x) = P(X = x)$.

$$\begin{aligned}E[aX + b] &= \sum_x (ax + b)p(x) = \sum_x (axp(x) + bp(x)) && \text{(LOTUS:}g(X) = aX + b\text{)} \\ &= a\sum_x xp(x) + b\sum_x p(x) && \text{(linearity of summation)} \\ &= aE[X] + b\sum_x p(x) && \text{(definition of expectation)} \\ &= aE[X] + b \cdot 1 && \text{(definition of PMF)}\end{aligned}$$

Proving the general case, where for two random variables X and Y, $E[aX + bY + c] = aE[X] + bE[Y] + c$ (Expectation of sum = sum of expectation): At this point in the course, we have yet to learn how to describe how two (or more) random variables interact. Once we learn about *joint* distributions of multiple random variables next week, we will revisit this proof!

4.2 Law of The Unconscious Statistician

$$E[g(X)] = \sum_x g(x) \cdot p_X(x)$$

Proof: Let X be a random variable with PMF $p(x) = P(X = x)$, and let $Y = g(X)$, where g is a real-valued function.

$$\begin{aligned} E[g(X)] = E[Y] &= \sum_j y_j P(Y = y_j) = \sum_j y_j \left(\sum_{i:g(x_i)=y_j} p(x_i) \right) \\ &= \sum_j \sum_{i:g(x_i)=y_j} y_j p(x_i) = \sum_j \sum_{i:g(x_i)=y_j} g(x_i) p(x_i) = \sum_i g(x_i) p(x_i) \end{aligned}$$

4.3 Variance identity

$$\text{Var}(X) = E[X^2] - E[X]^2$$

Proof: Let X be a random variable with PMF $p(x) = P(X = x)$.

$$\begin{aligned} \text{Var}(X) &= E[(X - E[X])^2] = E[(X - \mu)^2] && \text{Let } E[X] = \mu \\ &= \sum_x (x - \mu)^2 p(x) = \sum_x (x^2 - 2\mu x + \mu^2) p(x) && \text{(def. expectation, distributive property)} \\ &= \sum_x^2 p(x) - 2\mu \sum_x x p(x) + \mu^2 \sum_x p(x) && \text{(linearity of summation)} \\ &= E[X^2] - 2\mu E[X] + \mu^2 \cdot 1 && \text{(LOTUS, def. expectation, def. PMF)} \\ &= E[X^2] - 2\mu^2 + \mu^2 = E[X^2] - \mu^2 && \text{(rearrange terms)} \\ &= E[X^2] - (E[X])^2 \end{aligned}$$

4.4 Non-linearity of Variance

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

Proof: Let X be a random variable with PMF $p(x) = P(X = x)$, and let a, b be real-valued scalars.

$$\begin{aligned} \text{Var}(aX + b) &= E[(aX + b)^2] - (E[aX + b])^2 && \text{(variance identity)} \\ &= E[a^2 X^2 + 2abX + b^2] - (aE[X] + b)^2 && \text{(linearity, factoring)} \\ &= a^2 E[X^2] + 2abE[X] + b^2 - (a^2 (E[X])^2 + 2abE[X] + b^2) && \text{(linearity, factoring)} \\ &= a^2 E[X^2] - a^2 (E[X])^2 = a^2 (E[X^2] - (E[X])^2) \\ &= a^2 \, \text{Var}(X) && \text{(variance identity)} \end{aligned}$$