

Discrete Random Variables: Expected Value

Class 4, TF, AID-M

Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Know how to compute expected value (mean) of a discrete random variable.
2. Know the expected value of Bernoulli, binomial and geometric random variables.

2 Expected Value

In the R reading questions for this lecture, you simulated the average value of rolling a die many times. You should have gotten a value close to the exact answer of 3.5. To motivate the formal definition of the average, or [expected value](#), we first consider some examples.

Example 1. Suppose we have a six-sided die marked with five 3's and one 6. (This was the red one from our non-transitive dice.) What would you expect the average of 6000 rolls to be?

Solution: If we knew the value of each roll, we could compute the average by summing the 6000 values and dividing by 6000. Without knowing the values, we can compute the [expected average](#) as follows.

Since there are five 3's and one six we expect roughly 5/6 of the rolls will give 3 and 1/6 will give 6. Assuming this to be exactly true, we have the following table of values and counts:

value:	3	6
expected counts:	5000	1000

The average of these 6000 values is then

$$\frac{5000 \cdot 3 + 1000 \cdot 6}{6000} = \frac{5}{6} \cdot 3 + \frac{1}{6} \cdot 6 = 3.5$$

We consider this the expected average in the sense that we 'expect' each of the possible values to occur with the given frequencies.

Example 2. We roll two standard 6-sided dice. You win \$1000 if the sum is 2 and lose \$100 otherwise. How much do you expect to win on average per trial?

Solution: The probability of a 2 is 1/36. If you play N times, you can 'expect' $\frac{1}{36} \cdot N$ of the trials to give a 2 and $\frac{35}{36} \cdot N$ of the trials to give something else. Thus your total expected winnings are

$$1000 \cdot \frac{N}{36} - 100 \cdot \frac{35N}{36}.$$

To get the expected average per trial we divide the total by N :

$$\text{expected average} = 1000 \cdot \frac{1}{36} - 100 \cdot \frac{35}{36} = -69.44.$$

Think: Would you be willing to play this game one time? Multiple times?

Notice that in both examples the sum for the expected average consists of terms which are a value of the random variable times its probability. This leads to the following definition.

Definition: Suppose X is a discrete random variable that takes values x_1, x_2, \dots, x_n with probabilities $p(x_1), p(x_2), \dots, p(x_n)$. The **expected value** of X is denoted $E[X]$ and defined by

$$E[X] = \sum_{j=1}^n p(x_j) x_j = p(x_1)x_1 + p(x_2)x_2 + \dots + p(x_n)x_n.$$

Notes:

1. The expected value is also called the **mean** or **average** of X and often denoted by μ (“mu”).
2. As seen in the above examples, the expected value need not be a possible value of the random variable. Rather it is a weighted average of the possible values.
3. Expected value is a **summary statistic**, providing a measure of the **location** or **central tendency** of a random variable.
4. If all the values are equally probable then the expected value is just the usual average of the values.

Example 3. Find $E[X]$ for the random variable X with table:

values of X :	1	3	5
pmf:	1/6	1/6	2/3

Solution: $E[X] = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 3 + \frac{2}{3} \cdot 5 = \frac{24}{6} = 4$

Example 4. Let X be a Bernoulli(p) random variable. Find $E[X]$.

Solution: X takes values 1 and 0 with probabilities p and $1 - p$, so

$$E[X] = p \cdot 1 + (1 - p) \cdot 0 = p.$$

Important: This is an important example. Be sure to remember that the expected value of a Bernoulli(p) random variable is p .

Think: What is the expected value of the sum of two dice?

2.1 Mean and center of mass

You may have wondered why we use the name ‘probability mass function’. Here’s one reason: if we place an object of mass $p(x_j)$ at position x_j for each j , then $E[X]$ is the position of the center of mass. Let’s recall the latter notion via an example.

Example 5. Suppose we have two masses along the x -axis, mass $m_1 = 500$ at position $x_1 = 3$ and mass $m_2 = 100$ at position $x_2 = 6$. Where is the center of mass?

Solution: Intuitively we know that the center of mass is closer to the larger mass.



From physics we know the center of mass is

$$\bar{x} = \frac{m_1 x_1 + m_2 x_2}{m_1 + m_2} = \frac{500 \cdot 3 + 100 \cdot 6}{600} = 3.5.$$

We call this formula a ‘weighted’ average of the x_1 and x_2 . Here x_1 is weighted more heavily because it has more mass.

Now look at the definition of expected value $E[X]$. It is a weighted average of the values of X with the weights being probabilities $p(x_i)$ rather than masses! We might say that “The expected value is the point at which the distribution would balance”. Note the similarity between the physics example and Example 1.

2.2 Algebraic properties of $E[X]$

When we add, scale or shift random variables the expected values do the same. The shorthand mathematical way of saying this is that $E[X]$ is [linear](#).

1. If X and Y are random variables on a sample space Ω then

$$E[X + Y] = E[X] + E[Y]$$

2. If a and b are constants then

$$E[aX + b] = aE[X] + b.$$

We will think of $aX + b$ as [scaling](#) X by a and [shifting](#) it by b .

Before proving these properties, let’s see them in action with a few examples.

Example 6. Roll two dice and let X be the sum. Find $E[X]$.

Solution: Let X_1 be the value on the first die and let X_2 be the value on the second die. Since $X = X_1 + X_2$ we have $E[X] = E[X_1] + E[X_2]$. Earlier we computed that $E[X_1] = E[X_2] = 3.5$, therefore $E[X] = 7$.

Example 7. Let $X \sim \text{binomial}(n, p)$. Find $E[X]$.

Solution: Recall that X models the number of successes in n Bernoulli(p) random variables, which we’ll call X_1, \dots, X_n . The key fact, which we highlighted in the previous reading for this class, is that

$$X = X_1 + X_2 + \dots + X_n = \sum_{j=1}^n X_j.$$

Now we can use the Algebraic Property (1) to make the calculation simple.

$$X = \sum_{j=1}^n X_j \Rightarrow E[X] = \sum_j E[X_j] = \sum_j p = \boxed{np}.$$

We could have computed $E[X]$ directly as

$$E[X] = \sum_{k=0}^n kp(k) = \sum_{k=0}^n k \binom{n}{k} p^k (1-p)^{n-k}.$$

It is possible to show that the sum of this series is indeed np . We think you'll agree that the method using Property (1) is much easier.

Example 8. (For infinite random variables the mean does not always exist.) Suppose X has an infinite number of values according to the following table.

values x :	2	2^2	2^3	\dots	2^k	\dots	Try to compute the mean.
pmf $p(x)$:	$1/2$	$1/2^2$	$1/2^3$	\dots	$1/2^k$	\dots	

Solution: The mean is

$$E[X] = \sum_{k=1}^{\infty} 2^k \frac{1}{2^k} = \sum_{k=1}^{\infty} 1 = \infty.$$

The mean does not exist! This can happen with infinite series.

Example 9. Mean of a geometric distribution

Let $X \sim \text{geo}(p)$. Recall this means X takes values $k = 0, 1, 2, \dots$ with probabilities $p(k) = (1-p)^k p$. (X models the number of tails before the first heads in a sequence of Bernoulli trials.) The mean is given by

$$E[X] = \frac{1-p}{p}.$$

To see this requires a clever trick. Mathematicians love this sort of thing and we hope you are able to follow the logic and enjoy it. In this class we will not ask you to come up with something like this on an exam.

Here's the trick.: to compute $E[X]$ we have to sum the infinite series

$$E[X] = \sum_{k=0}^{\infty} k(1-p)^k p.$$

Now, we know the sum of the geometric series: $\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}$.

Differentiate both sides: $\sum_{k=0}^{\infty} kx^{k-1} = \frac{1}{(1-x)^2}$.

Multiply by x : $\sum_{k=0}^{\infty} kx^k = \frac{x}{(1-x)^2}$.

Replace x by $1-p$: $\sum_{k=0}^{\infty} k(1-p)^k = \frac{1-p}{p^2}$.

Multiply by p : $\sum_{k=0}^{\infty} k(1-p)^k p = \frac{1-p}{p}$.

This last expression is the mean.

$$E[X] = \frac{1-p}{p}.$$

Example 10. Flip a fair coin until you get heads for the first time. What is the expected number of times you flipped tails?

Solution: The number of tails before the first head is modeled by $X \sim \text{geo}(1/2)$. From the previous example $E[X] = \frac{1/2}{1/2} = 1$. This is a surprisingly small number.

Example 11. Michael Jordan, perhaps the greatest basketball player ever, made 80% of his free throws. In a game what is the expected number he would make before his first miss?

Solution: Here is an example where we want the number of successes before the first failure. Using the neutral language of heads and tails: success is tails (probability $1-p$) and failure is heads (probability p). Therefore $p = 0.2$ and the number of tails (made free throws) before the first heads (missed free throw) is modeled by a $X \sim \text{geo}(0.2)$. We saw in Example 9 that this is

$$E[X] = \frac{1-p}{p} = \frac{0.8}{0.2} = 4.$$

2.3 Expected values of functions of a random variable

([The change of variables formula.](#))

If X is a discrete random variable taking values x_1, x_2, \dots and h is a function then $h(X)$ is a new random variable. Its expected value is

$$E[h(X)] = \sum_j h(x_j)p(x_j).$$

We illustrate this with several examples.

Example 12. Let X be the value of a roll of one die and let $Y = X^2$. Find $E[Y]$.

Solution: Since there are a small number of values we can make a table.

X	1	2	3	4	5	6
Y	1	4	9	16	25	36
prob	1/6	1/6	1/6	1/6	1/6	1/6

Notice the probability for each Y value is the same as that of the corresponding X value. So,

$$E[Y] = E[X^2] = 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + \dots + 6^2 \cdot \frac{1}{6} = 15.167.$$

Example 13. Roll two dice and let X be the sum. Suppose the payoff function is given by $Y = X^2 - 6X + 1$. Is this a good bet?

Solution: We have $E[Y] = \sum_{j=2}^{12} (j^2 - 6j + 1)p(j)$, where $p(j) = P(X = j)$.

We show the table, but really we'll use R to do the calculation.

X	2	3	4	5	6	7	8	9	10	11	12
Y	-7	-8	-7	-4	1	8	17	28	41	56	73
prob	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Here's the R code I used to compute $E[Y] = 13.833$.

```
x = 2:12
y = x^2 - 6*x + 1
p = c(1 2 3 4 5 6 5 4 3 2 1)/36
ave = sum(p*y)
```

It gave $E[Y] = 13.833$.

To answer the question above: since the expected payoff is positive it looks like a bet worth taking.

Quiz: If $Y = h(X)$ does $E[Y] = h(E[X])$? **Solution:** NO!!! This is not true in general!

Think: Is it true in the previous example?

Quiz: If $Y = 3X + 77$ does $E[Y] = 3E[X] + 77$?

Solution: Yes. By property (2), scaling and shifting does behave like this.

2.4 Proofs of the algebraic properties of $E[X]$

We finish by proving the algebraic properties of $E[X]$.

1. For random variables X and Y on a sample space Ω : $E[X + Y] = E[X] + E[Y]$
2. For constants a, b and random variable X : $E[aX + b] = aE[X] + b$.

The proof of Property (1) is simple, but there is some subtlety in even understanding what it means to add two random variables. Recall that the value of random variable is a number determined by the outcome of an experiment. To add X and Y means to add the values of X and Y for the same outcome. In table form this looks like:

outcome ω :	ω_1	ω_2	ω_3	\dots	ω_n
value of X :	x_1	x_2	x_3	\dots	x_n
value of Y :	y_1	y_2	y_3	\dots	y_n
value of $X + Y$:	$x_1 + y_1$	$x_2 + y_2$	$x_3 + y_3$	\dots	$x_n + y_n$
prob. $P(\omega)$:	$P(\omega_1)$	$P(\omega_2)$	$P(\omega_3)$	\dots	$P(\omega_n)$

The proof of (1) follows immediately:

$$E[X + Y] = \sum (x_i + y_i)P(\omega_i) = \sum x_i P(\omega_i) + \sum y_i P(\omega_i) = E[X] + E[Y].$$

The proof of Property (2) only takes one line.

$$E[aX + b] = \sum p(x_i)(ax_i + b) = a \sum p(x_i)x_i + b \sum p(x_i) = aE[X] + b.$$

The b term in the last expression follows because $\sum p(x_i) = 1$.

Note by Prof. Mayer: This script is the work of the original authors (Jeremy Orloff and Jonathan Bloom), which appear in the title and all praise for this work must be to them. The only modifications done to fit it to "Theoretical fundamentals of AI and Data Science" (DIT, Faculty AI, Study course AIN-B) are leaving out certain topics and changing the class numbering.

The complete original script (and more: slides, exams, question sheets, and solutions) can be downloaded at [MIT OpenCourseWare: Introduction To Probability And Statistics, 18.05, Spring 2014, Undergraduate](#)

The OpenCourseWare material is distributed under the following Terms of use and license:
<https://ocw.mit.edu/pages/privacy-and-terms-of-use/>

Prof. Mayer thanks Jeremy Orloff and Jonathan Bloom not only for making the course material open to the public, but also for giving him access to the original LaTeX sources to tailor the material better to the DITs needs. Without their help, this course would by far not have the quality of its current state.