

# Mathematical Statistics I

## Chapter 4: Expected Values

Jesse Wheeler

### Contents

1	Discrete random variables	1
2	Continuous random variables	3
3	Expectation of functions of random variables	4

## 1 Discrete random variables

### Introduction

- This material comes primarily from Rice (2007, Chapter 4).
- We will cover the ideas of expected value, variance, as well as higher-order moments.
- This includes topics such as conditional expectation, which is one of the fundamental ideas behind many branches of statistics and machine learning.
- For instance, most regression / prediction algorithms are built with the idea of minimizing some conditional expectation.

### Expectation: Discrete random variables

#### Definition: Expectation of discrete random variables

Let  $X$  be a discrete random variable with pmf  $p(x)$ , which takes values in the space  $\mathcal{X}$ . The *expected value* of  $X$  is

$$E(X) = \sum_{x \in \mathcal{X}} x p(x),$$

provided that  $\sum_{x \in \mathcal{X}} |x| p(x) < \infty$ ; otherwise, the expectation is not defined.

- This is not the most mathematically precise definition of expectation, but a more complete treatment of the topic is outside the scope of this course (See Resnick, 2019).
- The concept of the expected value parallels the notion of a *weighted average*.
- That is, we weight each possibility  $x \in \mathcal{X}$  by their corresponding probability:  $\sum_x x p(x)$ .
- $E(X)$  is also referred to as the *mean* of  $X$ , and is typically denoted  $\mu$  or  $\mu_X$ .
- If the function  $p$  is thought of as a weight, then  $E(X)$  is the center; that is, if we place the mass  $p(x_i)$  at the points  $x_i$ , then the balancing point is  $E(X)$ .

- Like with the pmf and cdf, we often use subscripts to denote which probability law we are using for the expectation, if it is not clear:  $E_X(X)$ .

### *Roulette*

A roulette wheel has the numbers 1 through 36, as well as 0 and 00. If you bet \$1 that an odd number comes up, you win or lose \$1 according to whether that event occurs. If  $X$  denotes your net gain,  $X = 1$  with probability  $18/38$  and  $X = -1$  with probability  $20/38$ . The expected value of  $X$  is

$$E(X) = 1 \times \frac{18}{38} + (-1) \times \frac{20}{38} = -\frac{1}{19}.$$

- As you might imagine, the expected value coincides in the limit with the actual average loss per game, if you play many games (Chapter 5).
- Most casino games have a negative expected value by design; you may win some money, but if a large number of games are played, the house will come out on top.

### *Geometric Random Variable*

Suppose that items are produced in a plant are independently defective with probability  $p$ . If items are inspected one by one until a defective item is found, then how many items must be inspected on average?

Let  $X$  denote the number of items inspected, up-to and including the first defective item.  $X$  is geometrically distributed, which as pmf

$$p(k) = P(X = k) = p(1-p)^{k-1}.$$

Therefore

$$\begin{aligned} E(X) &= \sum_{k=1}^{\infty} kp(1-p)^{k-1} \\ &= p \sum_{k=1}^{\infty} k(1-p)^{k-1}. \end{aligned}$$

To work out this summation, we will use a trick that is sometimes useful for infinite series. First, let's define  $q = 1 - p$ , and note that  $0 < q < 1$ . Then, the sum becomes

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

You might notice that the summand is a power-rule derivative:

$$\frac{d}{dq} q^k = kq^{k-1}.$$

This fact is going to be useful, because the left-hand side of this derivative equation is a geometric sum, which we know how to calculate:

$$\sum_{k=1}^{\infty} q^k = \sum_{k=1}^{\infty} q q^{k-1} = q \sum_{j=0}^{\infty} q^j = \frac{q}{1-q}.$$

Thus, what we would like to do is write

$$\frac{d}{dq} \left( \frac{q}{1-q} \right) = \frac{d}{dq} \left( \sum_{k=1}^{\infty} q^k \right) \stackrel{?}{=} \sum_{k=1}^{\infty} \frac{d}{dq} q^k = \sum_{k=1}^{\infty} kq^{k-1}.$$

Now we can easily calculate the left-hand side to be  $\frac{1}{(1-q)^2}$ , and therefore we want to make the conclusion

$$\sum_{k=1}^{\infty} k q^{k-1} \stackrel{?}{=} \frac{d}{dq} \left( \frac{q}{1-q} \right) = \frac{1}{(1-q)^2}.$$

The question is: **Can we move the derivative inside of the infinite sum?** For this particular case, the answer is *yes*. In more advanced analysis classes, you learn methods for justifying this step rigorously using uniform convergence. Specifically, what we need is for uniform convergence of the partial sums and their derivatives. Fortunately for this class, all of the sums (and integrals) we will consider will be “well-behaved” and will satisfy these conditions.

With this sorted out, we can now use our trick to finish the calculation:

$$\begin{aligned} E(X) &= p \sum_{k=1}^{\infty} k (q)^{k-1} \\ &= p \frac{1}{(1-q)^2} \\ &= \frac{p}{p^2} = \frac{1}{p}. \end{aligned}$$

### Poisson Distribution

The Poisson( $\lambda$ ) distribution has pmf  $p(k) = \frac{\lambda^k}{k!} e^{-\lambda}$ , for all  $k \geq 0$ . Thus, if  $X \sim \text{Pois}(\lambda)$ , then what is  $E[X]$ ?

$$\begin{aligned} E[X] &= \sum_{k=0}^{\infty} \frac{k \lambda^k}{k!} e^{-\lambda} \\ &= e^{-\lambda} \sum_{k=0}^{\infty} \frac{k \lambda^{k-1} \cdot \lambda}{k!} \\ &= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \\ &= \lambda e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} \\ &= \lambda e^{-\lambda} e^{\lambda} = \lambda. \end{aligned}$$

## 2 Continuous random variables

### Expectation: Continuous random variables

#### Definition: Expectation of continuous random variables

Let  $X$  be a continuous random variable with pdf  $f(x)$ , which takes values in the space  $\mathcal{X}$ . The *expected value* of  $X$  is

$$E(X) = \int_{x \in \mathcal{X}} x f(x) dx.$$

provided that  $\int_{x \in \mathcal{X}} x f(x) dx < \infty$ , otherwise the expectation is undefined.

- As before, this is not the most mathematically precise definition of expectation, but a more complete treatment of the topic is outside the scope of this course (See Resnick, 2019).

- We can still think of  $E(X)$  as the center of mass of the density.

### Gamma Density

If  $X$  follows a gamma density with parameters  $\alpha$  and  $\lambda$ , then the pdf of  $X$  is

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \quad x \geq 0.$$

Find  $E(X)$ .

*Solution:* By definition, the expected value of  $X$  is

$$E(X) = \int_0^\infty (x) \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} dx.$$

Combining the factors of  $x$  in the integrand, we obtain

$$E(X) = \int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} x^\alpha e^{-\lambda x} dx.$$

Now we will apply the “integration by density function” trick: we will re-write the integrand so that it corresponds to the density function of some random variable, and use the fact that the density function must integrate to one. Specifically, note that if we let  $\alpha^* = \alpha + 1$ , then  $\alpha = \alpha^* - 1$ , and we can express the integral as:

$$\begin{aligned} E(X) &= \int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} x^\alpha e^{-\lambda x} dx \\ &= \int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha^*-1} e^{-\lambda x} dx \\ &= \left( \frac{\lambda^\alpha}{\Gamma(\alpha)} \right) \left( \frac{\Gamma(\alpha^*)}{\lambda^{\alpha^*}} \right) \int_0^\infty \frac{\lambda^{\alpha^*}}{\Gamma(\alpha^*)} x^{\alpha^*-1} e^{-\lambda x} dx \\ &= \left( \frac{\lambda^\alpha}{\Gamma(\alpha)} \right) \left( \frac{\Gamma(\alpha^*)}{\lambda^{\alpha^*}} \right) \end{aligned}$$

Where the last step is a result of the fact that the integrand (and support of the integral) matches the density of a  $\text{Gamma}(\alpha^*, \lambda)$  distribution. Now using the fact that  $\alpha^* = \alpha + 1$ , and that  $\Gamma(x+1) = x\Gamma(x)$ , we obtain

$$\begin{aligned} E(X) &= \frac{\lambda^\alpha \Gamma(\alpha + 1)}{\Gamma(\alpha) \lambda^{\alpha+1}} \\ &= \frac{\alpha}{\lambda} \end{aligned}$$

## 3 Expectation of functions of random variables

### Functions of random variables

- We are often interested in functions of random variables:  $Y = g(X)$ .
- Ideas that we have already covered enable us to calculate  $E(Y)$ .
- For instance, you could use the change-of-variables theorem to get the density of  $Y$ , then use the definition to calculate  $E[Y]$ .

- Fortunately, we don't have to do this. We can instead calculate  $E[Y]$  by integrating (or summing) with respect to  $X$ :

$$E[g(X)] = \int_{x \in \mathcal{X}} g(x) f(x) dx.$$

- We will justify this for the discrete analog.

**Theorem 4.1: Expectation of transformed random variables**

Suppose that  $X$  is a random variable and that  $Y = g(X)$  for some function  $g$ . Then,

- If  $X$  is discrete with pmf  $p(x)$ :

$$E(Y) = \sum_x g(x) p(x),$$

provided that  $\sum_x |g(x)| p(x) < \infty$ .

- If  $X$  is continuous with pdf  $f(x)$ :

$$E(Y) = \int_{-\infty}^{\infty} g(x) f(x) dx,$$

provided that  $\int |g(x)| f(x) dx < \infty$ .

**Functions of random variables: proof**

*Proof:* By definition of expectation,

$$E(Y) = \sum_i y_i p_Y(y_i).$$

Now let  $A_i$  denote the set of  $x$ 's that are mapped to  $y_i$  by  $g$ . That is,  $A_i$  is the pre-image of  $y_i$ , meaning that  $x \in A_i$  if  $g(x) = y_i$ . Then,

$$p_Y(y_i) = \sum_{x \in A_i} p(x),$$

and we can express the expectation as

$$\begin{aligned} E(Y) &= \sum_i y_i p_Y(y_i) \\ &= \sum_i y_i \sum_{x \in A_i} p(x) \\ &= \sum_i \sum_{x \in A_i} y_i p(x) \\ &= \sum_i \sum_{x \in A_i} g(x) p(x) \\ &= \sum_x g(x) p(x) \end{aligned}$$

Here, the second to last step is because for all  $x \in A_i$ ,  $g(x) = y_i$  by definition. The final step is a result of the fact that the  $A_i$  are disjoint, and every  $x$  belongs to some  $A_i$ , and thus the sum over  $i$  and  $x \in A_i$  is the sum of all  $x$ .

- The proof for the continuous case is similar, but does require a measure-theoretic approach to integration.
- One important thing to note is that  $g(E(X))$  is not usually equal to  $E(g(x))$ .

- For example, let  $Z$  be a standard normal. We know that  $E[Z] = 0$ , because it's symmetric. However,  $P(|Z| > 0) = 1$ , thus we can readily deduce that  $E[|Z|] \geq 0 = |E[Z]|$ .
- An immediate consequence is that if for all non-negative random variables  $X$  that have finite expectation, if  $g(x) \leq x$  for some function  $g$ , then  $E[g(X)] \leq E[X]$ .

### Expected value of indicator functions

- An interesting example is *indicator* functions.
- For example, suppose that  $X$  is a random variable. Then  $Y = 1[X \in A]$  for some  $A \subset \mathcal{X}$  is a random variable.
- Example: Let  $X$  follow a standard normal distribution, and  $A = [-1, 1]$ . Then  $Y = 1[X \in A]$  is defined as the random variables such that  $Y(\omega) = 1$  if  $X(\omega) \in A$ , and  $Y(\omega) = 0$  otherwise.
- Expectations of indicator variables are *probabilities*:

$$\begin{aligned} E(Y) &= E(1[X \in A]) \\ &= \int_{x \in \mathcal{X}} 1[X \in A] f(x) dx \\ &= \int_{x \in A} f(x) dx = P(X \in A). \end{aligned}$$

- This fact is useful for deriving some important inequalities.
- Let  $X$  be a continuous random variable with expectation  $E(X)$ . From our definition, this implies that  $\int |x| f(x) dx < \infty$ .
- Now suppose that for some random variable  $Y = g(X)$  such that  $|Y| \leq |X|$ . Then, if  $Y$  has a pdf, we can deduce that  $\int |y| f(x) dx < \infty$ , and therefore  $E[Y]$  exists.
- Now suppose that  $\varphi$  is a non-decreasing, non-negative function, and that for some  $a \in \mathbb{R}$ ,  $\varphi(a) > 0$ . Then, for all  $x \geq a$ ,  $\varphi(x)/\varphi(a) \geq 1$ .
- Define  $Y = 1[X \geq a]$ . Note that for all possible outcomes  $\omega \in \Omega$ ,

$$Y = 1[X \geq a] \leq \varphi(X)/\varphi(a) 1[X \geq a] \leq \varphi(X)/\varphi(a).$$

- Taking expectations of both sides,

$$E(1[X \geq a]) = P(X \geq a) \leq \frac{E[\varphi(X)]}{\varphi(a)} = E[\varphi(X)/\varphi(a)].$$

- This inequality is known as *Markov's (general) inequality*, and is very useful for bounding the probability of particular events.
- Specifically, if  $\varphi(x) = |x|^p$ , with  $p > 0$ , then because  $|X|$  is always positive,  $\varphi$  is non-negative, non-decreasing, and therefore

$$P(|X| \geq a) \leq \frac{E[|X|^p]}{a^p},$$

- If we restrict ourselves to the case where  $X$  is non-negative, we get the most standard version of the inequality:

$$P(X \geq a) \leq E(X)/a.$$

### Markov's Inequality in Action

Suppose that an individual is taken randomly from a population that has an average salary of \$50,000. If we assume that salary from the population is approximately independently and identically distributed, we can provide an upper-bound for the probability that the individual is wealthy.

Let  $X_i$  be the salary of individual  $i$ , randomly drawn from said population. Even though all we know is the average salary, Markov's inequality tells use that:

$$P(X \geq 200,000) \leq \frac{50,000}{200,000} = \frac{1}{4}.$$

- Returning to expectations of functions of random variables, we can extend to the multi-variate case

### Theorem 4.2: functions of multiple variables

Suppose that  $X_1, \dots, X_n$  are jointly distributed RVs and  $Y = g(X_1, \dots, X_n)$ . Then

- IF  $X_i$  are discrete with pmf  $p(x_1, \dots, x_n)$ , then

$$E(Y) = \sum_{x_1, \dots, x_n} g(x_1, \dots, x_n) p(x_1, \dots, x_n).$$

- If  $X_i$  are continuous with pdf  $f(x_1, \dots, x_n)$ , then

$$E(Y) = \int_{\mathcal{X}_1, \dots, \mathcal{X}_n} g(x_1, \dots, x_n) f(x_1, \dots, x_n) dx_1 \dots dx_n.$$

In both cases, we need the sum (or integral) of  $|g|$  to converge.

- The proof for the discrete case of Theorem 4.2 follows directly that of Theorem 4.1
- An immediate consequence of Theorem 4.2 is the following

### Corollary 4.2.1

If  $X$  and  $Y$  are independent random variables, and  $g$  and  $h$  are fixed functions, then

$$E[g(X)h(Y)] = \left( E[g(X)] E[h(Y)] \right),$$

provided that the expectations on the right-hand side exist.

### Example: Breaking sticks

A stick of unit-length is broken randomly (uniformly) in two places. What is the average length of the middle piece?

We will interpret this problem to mean that the locations of the two break-points are independent uniform random variables,  $U_1$  and  $U_2$ , and we need to computing  $E|U_1 - U_2|$ .

*Solution:* Theorem 4.2 implies that we do not need to find the density function of  $U_1 - U_2$ . Instead, we just need to integrate  $|u_1 - u_2|$  against the joint density:  $f(u_1, u_2) = 1$ , with  $0 \leq u_1, u_2 \leq 1$ . Thus

$$E|U_1 - U_2| = \int_0^1 \int_0^1 |u_1 - u_2| du_1 du_2$$

Splitting this integral into two regions, one where  $u_1 \geq u_2$  and one where  $u_2 > u_1$ , we get

$$\begin{aligned} E|U_1 - U_2| &= \int_0^1 \int_0^{u_1} (u_1 - u_2) du_2 du_1 + \int_0^1 \int_{u_1}^1 (u_2 - u_1) du_2 du_1 \\ &= \frac{1}{3} \end{aligned}$$

Logically, this result makes sense as it suggests that if we have two break points (three pieces) and the break points are uniform and random, the middle piece, on average, will be 1/3 of the length of the original stick.

## Linear Combinations of Random Variables

- A useful property of expectation is that it is a *linear operator*.

### Theorem 4.3: Linear combinations

If  $X_1, \dots, X_n$  are jointly distributed random variables with expectations  $E(X_i)$ , respectively, and  $Y = a + \sum_{i=1}^n b_i X_i$ , then,

$$E(Y) = a + \sum_{i=1}^n b_i E(X_i).$$

*Proof.* We will show this for the continuous case with  $n = 2$ . The proof for the discrete case is similar, and this argument is readily extended to the case that  $n > 2$ . First, we will argue that the expectation is well-defined. By definition,

$$E|Y| = \int |a + b_1 x_1 + b_2 x_2| f(x_1, x_2) dx_1 dx_2$$

and by the triangle inequality  $|a + b_1 x_1 + b_2 x_2| \leq |a| + |b_1||x_1| + |b_2||x_2|$ , and the fact that  $E[X_i]$  exists (which implies that  $E|X_i| < \infty$ ), we see that  $E|Y| < \infty$ . Now we can calculate the expected value. By Theorem 4.2,

$$\begin{aligned} E(Y) &= \int (a + b_1 x_1 + b_2 x_2) f(x_1, x_2) dx_1 dx_2 \\ &= a \int f(x_1, x_2) dx_1 dx_2 + b_1 \int x_1 f(x_1, x_2) dx_1 dx_2 \\ &\quad + b_2 \int x_2 f(x_1, x_2) dx_1 dx_2 \end{aligned}$$

Note that the first integral is equal to 1, because it's the integral of a joint pdf over the support. We'll focus now on the second integral, which is evaluated in a similar way as the third integral due to symmetry.

$$\begin{aligned} b_1 \int x_1 f(x_1, x_2) dx_1 dx_2 &= b_1 \int x_1 \left( \int f(x_1, x_2) dx_2 \right) dx_1 \\ &= b_1 \int x_1 f(x_1) dx_1 \\ &= b_1 E[X_1]. \end{aligned}$$

Thus, applying the same idea to the third integral, we get

$$E(Y) = a + b_1 E(X_1) + b_2 E(X_2).$$

- The previous theorem is extremely useful for calculating expected values.
- An obvious example is *sums* of random variables, such as the arithmetic average.
- It's also useful because some distributions can be expressed as the sum of other distributions.
- For instance, we saw in a previous example that the sum of two exponential random variables has a Gamma distribution. Thus, if we know the mean of an exponential, we can readily calculate the mean of a Gamma distribution.



*Expectation of a binomial distribution*

Let  $Y$  follow a Binomial( $p, q$ ) distribution. Find the expected value of  $Y$ .

*Solution.* The pmf of a binomial distribution is given by

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Therefore, to find the expected value directly, we need to calculate the sum

$$E(Y) = \sum_{k=0}^n k \binom{n}{k} p^k (1-p)^{n-k}.$$

It's not immediately clear how one can calculate this sum directly. Instead, we can use the fact that a binomial random variable is defined by the sum of independent Bernoulli distributed random variables. That is, let  $X_1, X_2, \dots, X_n$  be Bernoulli random variables with parameter  $p$ . Then

$$Y \stackrel{d}{=} \sum_{i=1}^n X_i$$

where the symbol  $\stackrel{d}{=}$  is used to indicate that  $Y$  and  $\sum_i X_i$  have the same distribution<sup>1</sup>. Then it is very easy to calculate the expected value of  $Y$ , as it is the sum of expected values of  $X_i$ :

$$E[X_i] = p(1) + (1-p)(0) = p.$$

Thus,

$$E[Y] = \sum_i E[X_i] = \sum_i p = np.$$


*Example: Coupon Collection*

Suppose that you collect coupons, that there are  $n$  distinct coupons, and that on each trial you are equally likely to get a coupon of any of the types. TODO: finish example.

---

<sup>1</sup>they are not necessarily equal to each-other, as they are not necessarily defined using the same probability space. Instead, we only require that they have the same distribution.

## Acknowledgments

- Compiled on August 26, 2025 using R version 4.5.1.
- Licensed under the [Creative Commons Attribution-NonCommercial license](#).  Please share and remix non-commercially, mentioning its origin.
- We acknowledge [students and instructors for previous versions of this course / slides](#).

## References

Resnick S (2019). *A probability path*. Springer. [2](#), [3](#)

Rice JA (2007). *Mathematical statistics and data analysis*, volume 371. 3 edition. Thomson/Brooks/Cole Belmont, CA. [1](#)