

Lecture Note 4: Expectations (CB 2.2–2.3)

The concept of expectation is central to probability theory and statistics. Intuitively, the expectation of a random variable is the long run average obtained by repeatedly and independently performing the experiment. For example, suppose for an entire evening you play roulette. Each time you put five dollars on number 17 and twenty dollars on red. Sometimes you win a lot (if 17 comes up), sometimes you win a little (if a red number comes up), often you lose your twenty-five dollars altogether. If you average your winnings over the entire evening, and if the evening is long enough, these average winnings will be close to the expected value or expectation of the random variable defined as the winnings in a single game. Formally,

Definition 1 The expectation or expected value of a random variable X , denoted $E(X)$, is (i), if X is a discrete random variable with PMF $f_X(x)$, equal to:

$$\sum_x x \cdot f_X(x),$$

provided the sum $\sum_x |x|f_X(x) < \infty$ (the sum “exists”).

(ii), if X is a continuous random variable with PDF $f_X(x)$, equal to

$$\int_{\mathcal{X}} x \cdot f_X(x) dx.$$

provided the integral $\int_x |x|f_X(x) dx < \infty$ (the integral “exists”).

Notice that we are essentially taking a weighted average with respect to the frequency of the different possible values that X can take on.

Example: Suppose you toss a single die. The discrete random variable X , the number on top of the die, has a distribution with PMF:

$$f_X(x) = \begin{cases} 1/6 & x = 1, 2, 3, 4, 5, 6, \\ 0 & \text{otherwise.} \end{cases}$$

The expected value or expectation is

$$E(X) = \sum_{k=1}^6 k \cdot (1/6) = 7/2.$$

Note that $7/2$ is not a typical outcome. In fact, it cannot occur in this experiment. Nevertheless, it is the outcome you get on average, in the sense defined above. For a more typical value one might wish to choose the mode, defined as the most likely value (i.e. the maximizer of the PDF/PMF). In this case that would be any of the values 1, 2, 3, 4, 5 or 6. Alternatively one could report the median, defined as any c such that $F_X(x) \leq 1/2$ for $x < c$ and $F_X(x) \geq 1/2$ for $x > c$. In this case that would be any value in the interval $[3, 4]$. The expectation, median, and mode capture different notions of the “central tendency” of the random variable, and which one is most useful depends on the application at hand.

□.

Example: Suppose X has an exponential distribution with parameter $\beta > 0$ and PDF

$$f_X(x) = \begin{cases} \frac{1}{\beta} \exp\left(-\frac{x}{\beta}\right) & x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Before we have seen the unit exponential distribution with $\beta = 1$. The general form is useful because it is more flexible: different choices for β yield different distributions, although all have a similar shape. This distribution turns out to be a good approximation for many types of durations: waiting time of a customer for a telephone operator, time to failure of a light bulb, length of unemployment spell, etc.

Let us calculate the expected value and see how it depends on the choice of β :

$$E(X) = \int_{\mathcal{X}} x \cdot f_X(x) dx = \int_0^{\infty} x \frac{1}{\beta} \exp\left(-\frac{x}{\beta}\right) dx.$$

Recall the integration by parts formula: $\int u dv = uv - \int v du$. Use

$$\begin{aligned} u &= x, & dv &= \frac{1}{\beta} \exp\left(-\frac{x}{\beta}\right) dx. \\ \Rightarrow du &= dx, & v &= -\exp\left(-\frac{x}{\beta}\right). \end{aligned}$$

Then:

$$\begin{aligned} E(X) &= \int_0^{\infty} x \frac{1}{\beta} \exp\left(-\frac{x}{\beta}\right) dx = -x \exp\left(-\frac{x}{\beta}\right) \Big|_0^{\infty} + \int_0^{\infty} \exp\left(-\frac{x}{\beta}\right) dx \\ &= 0 + \int_0^{\infty} \exp\left(-\frac{x}{\beta}\right) dx = -\beta \exp\left(-\frac{x}{\beta}\right) \Big|_0^{\infty} = \beta. \end{aligned}$$

The mode is zero for this distribution, regardless of the value of β . To calculate the median, solve $1/2 = 1 - F(x) = 1 - \exp\left(-\frac{x}{\beta}\right)$, leading to $x = -\ln(1/2) \cdot \beta \approx 0.69 \cdot \beta$. \square

Often we are interested in expectations of transformations of X . Given that we have already defined transformations, the expected values for transformations are defined implicitly:

$$E[r(X)] = E[Y], \quad \text{where } Y = r(X).$$

We therefore do not need to define expectations of transformations. However, we do not have to calculate these expectations by first calculating the CDF or PDF/PMF of these transformations. A much more direct route is provided by the following result:

Result 1 *The expectation of a function $r(\cdot)$ of a random variable X is*

(i), if X is a discrete random variable with PMF $f_X(x)$,

$$\sum_x r(x) \cdot f_X(x).$$

(ii), if X is a continuous random variable with PDF $f_X(x)$,

$$\int_{\mathcal{X}} r(x) \cdot f_X(x) dx.$$

For the discrete case with a monotone transformation:

$$E(Y) = \sum_y y \cdot f_X(r^{-1}(y)) = \sum_x r(x) f_X(x).$$

Here is also a sketch of a proof for the continuous case with $r(\cdot)$ a monotone (increasing) transformation (which is not required for the result). Let $Y = r(X)$. Then

$$E[r(X)] = E(Y) = \int_{\mathcal{Y}} y \cdot f_Y(y) dy = \int_{\mathcal{Y}} y f_X(r^{-1}(y)) \cdot \frac{\partial r^{-1}}{\partial y}(y) dy.$$

Transform the integrand from y to $z = r^{-1}(y)$ with inverse transformation $y = r(z)$ to get

$$\begin{aligned} \int_{\mathcal{Y}} y f_X(r^{-1}(y)) \cdot \frac{\partial r^{-1}}{\partial y}(y) dy &= \int_{\mathcal{Z}} r(z) f_X(z) \cdot \frac{\partial r^{-1}}{\partial r(z)}(r(z)) dr(z) \\ &= \int_{\mathcal{Z}} r(z) f_X(z) \cdot \frac{\partial r^{-1}}{\partial y}(r(z)) \frac{\partial r(z)}{\partial z}(z) dz \\ &= \int_{\mathcal{Z}} r(z) f_X(z) dz. \end{aligned}$$

For the last step, note that $z = r^{-1}(r(z))$ and so by the chain rule for differentiation and by differentiating both sides with respect to z , $1 = \frac{\partial r^{-1}}{\partial r(z)}(r(z)) \frac{\partial r(z)}{\partial z}(z)$ and hence $\frac{\partial r^{-1}}{\partial r(z)}(r(z)) = 1/\frac{\partial r}{\partial z}r(z)$.

Next, let us consider some special expectations. In all cases X is a random variable.

1. The mean is another name for the expectation or expected value of X , and we often use the Greek letter $\mu = E(X)$.
2. The k th moment of X is $\mu_k = E(X^k)$ (so $\mu_1 = \mu$).
3. The variance of X is $V(X) = \sigma^2 = E((X - \mu)^2)$. With a bit of algebra you can show that $\sigma^2 = \mu_2 - \mu_1^2$.
4. The expectation of a linear function of a random variable is the linear function of the expectation:

$$E(a + b \cdot X) = a + b \cdot E(X).$$

The variance of a linear function of a random variable is the variance of the random variable multiplied by the square of the slope coefficient:

$$V(a + b \cdot X) = b^2 \cdot V(X).$$

5. More generally,

$$E[a \cdot g_1(X) + b \cdot g_2(X) + c] = aE[g_1(X)] + bE[g_2(X)] + c.$$

6. If $g_1(x) \geq g_2(x)$ for all x , then $E[g_1(X)] \geq E[g_2(X)]$.

See CB, Theorem 2.2.5 for some other handy results using expectations.

Example: Suppose we perform an experiment that can have one of two outcomes, success or failure. Let p be the probability of success. Let Y be the indicator for success, equal to one if the experiment is a success and zero otherwise. This is referred to as a Bernoulli trial. Now repeat this experiment n times, and assume that the repetitions are independent. Let Y_i , for $i = 1, \dots, n$ denote the success in the i th Bernoulli trial. Define $X = \sum_{i=1}^n Y_i$ as the total number of successes. Then X has a binomial distribution. To figure out its PMF, consider the probability of particular sequence of x successes and $n - x$ failures, for example first x one's and then $n - x$ zero's. The probability of *any* one such a sequence is $p^x(1 - p)^{n-x}$. To get the probability of x successes and $n - x$ failures, we need to count the number of such sequences. This is equal to the number of ways you can select x objects out of a set of n , which is $\binom{n}{x}$. Hence the PMF of X is

$$f_X(x) = \binom{n}{x} \cdot p^x \cdot (1 - p)^{(n-x)},$$

for $x = 0, 1, 2, \dots, n$, and zero otherwise.

Now let us calculate the mean of the binomial distribution. One approach exploits the result that the PMF adds up to one:

$$\sum_{x=0}^n \binom{n}{x} \cdot p^x \cdot (1 - p)^{(n-x)} = \sum_{x=0}^n f_X(x) = 1.$$

Now,

$$\begin{aligned} E(X) &= \sum_{x=0}^n x \cdot \binom{n}{x} \cdot p^x \cdot (1 - p)^{(n-x)} = \sum_{x=1}^n x \cdot \binom{n}{x} \cdot p^x \cdot (1 - p)^{(n-x)} \\ &= \sum_{x=1}^n x \cdot \frac{n!}{x!(n-x)!} \cdot p^x \cdot (1 - p)^{(n-x)} = \sum_{x=1}^n \frac{n!}{(x-1)!(n-x)!} \cdot p^x \cdot (1 - p)^{(n-x)} \\ &= \sum_{x=1}^n n \cdot p \cdot \frac{(n-1)!}{(x-1)!((n-1)-(x-1))!} \cdot p^{(x-1)} \cdot (1 - p)^{((n-1)-(x-1))}. \end{aligned}$$

Define $m = n - 1$ and $y = x - 1$. Then instead of summing from $x = 1$ to $x = n - 1$ we sum from $y = 0$ to $y = m$ and get:

$$\begin{aligned} E(X) &= n \cdot p \cdot \sum_{y=0}^m \frac{m!}{y!(m-y)!} \cdot p^y \cdot (1 - p)^{(m-y)} \\ &= n \cdot p \cdot \sum_{y=0}^m \binom{m}{y} \cdot p^y \cdot (1 - p)^{(m-y)} = np. \end{aligned}$$

A much simpler approach in this case is to write X as the sum of independent and identically distributed random variables, $X = \sum_{i=1}^n Y_i$, with

$$f_Y(y) = p^y \cdot (1 - p)^{(1-y)},$$

and mean p . Expectations are additive: if Y_1 and Y_2 are two random variables, then $E[Y_1 + Y_2] = E[Y_1] + E[Y_2]$. Hence the mean of X is

$$E(X) = E\left(\sum_{i=1}^n Y_i\right) = \sum_{i=1}^n E(Y_i) = np. \square$$

Expectation as Best Predictor

Earlier we discussed the intuition of the expected value as a long-run average. Another useful way to view the expected value, is from the context of a simple prediction problem.

Suppose we want to choose a single value b as a prediction for the random outcome X . We might measure the quality of the prediction by $(X - b)^2$. This is the squared prediction error, and presumably we want this to be as small as possible. We might try to choose b to minimize the expectation

$$E[(X - b)^2].$$

We can solve this minimization problem by calculus, but there is an alternative way to solve the problem. Write $X - b$ as $X - \mu + \mu - b$, where $\mu = E[X]$. Then

$$\begin{aligned} E[(X - b)^2] &= E[((X - \mu) + (\mu - b))^2] \\ &= E[(X - \mu)^2 + (\mu - b)^2 + 2(X - \mu)(\mu - b)] \\ &= E[(X - \mu)^2] + (\mu - b)^2 + 2(\mu - b)E[(X - \mu)] \end{aligned}$$

Note however that $E[X - \mu] = 0$ by definition. So

$$E[(X - b)^2] = E[(X - \mu)^2] + (\mu - b)^2.$$

The first term is equal to the variance, and is the same regardless of the choice of b . The second term is clearly minimized by setting $b = \mu$. Thus, the best predictor in this problem is $b = \mu$.

You might wonder what would happen if we used a different definition of prediction error. For example, what if we try to choose b to minimize the expected absolute prediction error:

$$\min_b E(|X - b|).$$

Then, it can be shown that the solution is to set b equal to the median of X .

Some Useful Inequalities

In some cases it is difficult to exactly calculate certain probabilities, but they can be bounded if one knows certain moments of the distribution:¹

Result 2 Markov's Inequality: Suppose X is a nonnegative random variable. Then for any $a > 0$,

$$Pr(X \geq a) \leq \frac{E[X]}{a}.$$

Proof: Consider the “indicator function”

$$1(x \geq a) = \begin{cases} 1 & x \geq a \\ 0 & \text{otherwise} \end{cases}$$

This is a step function equal to 0 for values less than a , and equal to one for values greater than or equal to a . Notice that if X is continuous,

$$E[1(x \geq a)] = \int_0^\infty 1(x \geq a)f_X(x)dx = \int_a^\infty f_X(x)dx = Pr(X \geq a).$$

¹Note my labelling of these theorems differs slightly from CB's in sections 3.6 and 3.8. Mine is better.

Similarly, if X is discrete then we also have $E[1(x \geq a)] = Pr(X \geq a)$. Notice that we will always have

$$1(x \geq a) \leq \frac{x}{a},$$

since for $x < a$, the indicator function is zero, while for $x > a$, x/a is greater than one. So

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

□

Result 3 Chebyshev's Inequality: For any random variable Y with mean μ and variance σ^2 , and for any $k > 0$

$$Pr\left(|Y - \mu| \geq k \cdot \sigma\right) \leq 1/k^2.$$

Proof: Apply Markov's inequality with $X = (Y - \mu)^2$ and $c = k^2\sigma^2$ to rewrite the inequality

$$Pr(X \geq c) \leq E[X]/c,$$

as

$$Pr((Y - \mu)^2 \geq k^2\sigma^2) \leq 1/k^2,$$

which is equivalent to the statement in the result. □

Generating Functions

The CDF completely characterizes the distribution of a random variable. Sometimes, it is useful to consider other ways of characterizing distributions.

1. The moment generating function (mgf), denoted by $M_X(t)$, is the expected value of $\exp(t \cdot X)$. We are interested in this function for values of t around zero. It has a number of interesting properties:
 - (a) It uniquely defines the distribution of a random variable: if two random variables have the same mgf for all t in an interval around zero, they have the same CDF and PMF/PDF (up to sets of measure zero).
 - (b) The k th moment is equal to the k th derivative of the mgf evaluated at zero:

$$\mu_k = \frac{\partial^k M_X}{\partial t^k}(0).$$

In particular, using $M_X^k(t)$ as shorthand for $\frac{\partial^k M_X}{\partial t^k}(t)$, we have $M_X(0) = 1$, $\mu = M_X^1(0)$, $\sigma^2 = M_X^2(0) - M_X^1(0)^2$.

2. The cumulant generating function, is the logarithm of the moment generating function, $K_X(t) = \ln M_X(t)$. Here $\mu = K_X^1(0)$ and $\sigma^2 = K_X^2(0)$.
3. The characteristic function defined as

$$\psi_X(t) = E[\exp(Xit)],$$

where $i = \sqrt{-1}$. Working with the characteristic function rather than the moment generating function has the advantage that the former always exists, while the latter does not always exist.

Example: Consider the exponential distribution with PDF $\frac{1}{\beta} \exp(-\frac{x}{\beta})$ for $x > 0$ and zero otherwise. We have already calculated the expected value. Here we shall see that using the moment generating function is a much easier way of calculating this expectation, avoiding the integration by parts. First, it's handy to reparametrize the distribution with $\theta = 1/\beta$, so that the PDF is now $\theta \exp(-\theta x)$. Then

$$\begin{aligned} M_X(t) &= \int_0^\infty \theta \exp(-\theta x) \exp(tx) dx = \int_0^\infty \theta \exp(-x(\theta - t)) dx \\ &= -\frac{\theta}{\theta - t} \exp(-x(\theta - t)) \Big|_0^\infty = \frac{\theta}{\theta - t}. \end{aligned}$$

Note that this only works for $t < \theta$ but we only care about values of t around zero so that is no problem. The derivative of the mgf is

$$M_X^k(t) = k! \frac{\theta}{(\theta - t)^{k+1}},$$

so that

$$E[X^k] = M_X^k(0) = \frac{k!}{\theta^k},$$

and thus the mean is $1/\theta = \beta$, and the second moment $2/\theta^2 = 2\beta^2$. Using the formula $\sigma^2 = \mu_2 - \mu_1^2$, the variance is $1/\theta^2 = \beta^2$. \square