Probability and Statistics

Fall 2000

Lecture 4: Expectations (CB 2.2–2.4)

An important role in statistics is played by expectations. 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 of a random variable 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 $\sum_x |x| f_X(x)$ exists,
(ii), if X is a continuous random variable with pdf $f_X(x)$, equal to

$$\int_x x \cdot f_X(x) dx.$$

provided the integral $\int_x |x| f_X(x) dx$ exists.

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. 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 $[3, 4]$. For this distribution the mean is a more useful measure of central tendency than the mode or the median.
□.

Example: Suppose X has an exponential distribution with parameter λ and pdf

$$f_X(x) = \begin{cases} \lambda \exp(-\lambda x) & x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Before we have seen the unit exponential distribution with $\lambda = 1$. The general form is useful because it is more flexible. It can be used to fit distributions of durations in a variety of settings: waiting times, light bulbs, unemployment spells, and many others. The expected value is

$$E(X) = \int_x x \cdot f_X(x) dx = \int_0^\infty x\lambda \exp(-\lambda x) dx.$$

Integration by parts gives

$$E(X) = \int_0^\infty x\lambda \exp(-\lambda x) dx = -x \exp(-\lambda x)\Big|_0^\infty + \int_0^\infty \exp(-\lambda x) dx$$

$$= 0 + \int_0^\infty \exp(-\lambda x) dx = -\frac{1}{\lambda} \exp(-\lambda x)\Big|_0^\infty = 1/\lambda.$$

The mode is zero for this distribution, irrespective of the value of λ. To calculate the median, solve $1/2 = 1 - \exp(-\lambda x)$, leading to $-\ln(1/2)/lambda \approx 0.69/\lambda$. □

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:

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

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 $f_X(x)$,

$$\int_x r(x) \cdot f_X(x) dx.$$

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

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

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

$$\int_y y f_X(r^{-1}(y)) \cdot \frac{\partial r^{-1}}{\partial y}(y) dy = \int_z r(z) f_X(z) \cdot \frac{\partial r^{-1}}{\partial r(z)}(r(z)) dr(z)$$

$$= \int_z r(z) f_X(z) \cdot \frac{\partial r^{-1}}{\partial y}(r(z)) \frac{\partial r(z)}{\partial z}(z) dz$$

$$= \int_z r(z) f_X(z) dz.$$

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 with pdf/pmf equal to $f_X(x)$:

1. The mean is another name for the expectation or expected value of X, $\mu = E(X)$.
2. The kth moment of X is $\mu_k = E(X^k)$ (so $\mu_1 = \mu$).
3. The variance of X is $V(X) = \sigma^2 = E((X-\mu)^2) = \mu_2 - \mu_1^2$.
4. 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 kth moment is equal to the kth 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 $M_X(0) = 1$, $\mu = M_X^1(0)$, $\sigma^2 = M_X^2(0) - M_X^1(0)^2$.
5. The cumulative generating function, is the logaritm of the moment generating function, $K_X(t) = \ln M_X(t)$. Here $\mu = K_X^1(0)$ and $\sigma^2 = K_X^2(0)$.
6. 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.
7. 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:

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

Example: Consider the exponential distribution with pdf $\lambda \exp(-\lambda x)$ 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.

$$M_X(t) = \int_0^\infty \lambda \exp(-\lambda x) \exp(tx) dx = \int_0^\infty \lambda \exp(-x(\lambda - t)) dx$$

$$= \frac{\lambda}{\lambda - t} \exp(-x(\lambda - t))\Big|_0^\infty = \frac{\lambda}{\lambda - t}.$$

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

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

so that

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

and thus the mean is $1/\lambda$, the second moment $2/\lambda^2$, and the variance $1/\lambda^2$. □

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 $i = 1, \ldots, n$ denote the success in the ith 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 $\left(\begin{array}{c} n \\ x \end{array} \right)$. Hence the pmf of X is

$$f_X(x) = \left(\begin{array}{c} n \\ x \end{array} \right) \cdot p^x \cdot (1-p)^{(n-x)},$$

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

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

$$\sum_{x=0}^{n} \left(\begin{array}{c} n \\ x \end{array} \right) \cdot p^x \cdot (1-p)^{(n-x)} = \sum_x f_X(x) = 1.$$

Now,

$$E(X) = \sum_{x=0}^{n} x \cdot \left(\begin{array}{c} n \\ x \end{array} \right) \cdot p^x \cdot (1-p)^{(n-x)} = \sum_{x=1}^{n} x \cdot \left(\begin{array}{c} n \\ x \end{array} \right) \cdot p^x \cdot (1-p)^{(n-x)}$$

$$= \sum_{x=1}^{n} x \cdot \frac{n!}{x!(n-x)!} \cdot p^x \cdot (1-p)^{(n-x)} = \sum_{x=1}^{n} \frac{n!}{(x-1)!(n-x)!} \cdot p^x \cdot (1-p)^{(n-x)}$$

$$= \sum_{x=1}^{n} n \cdot p \cdot \frac{(n-1)!}{(x-1)!((n-1)-(x-1))!} \cdot p^{(x-1)} \cdot (1-p)^{((n-1)-(x-1))}.$$

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:

$$E(X) = n \cdot p \cdot \sum_{y=0}^{m} \frac{m!}{y!(m-y)!} \cdot p^y \cdot (1-p)^{(m-y)}$$

$$= n \cdot p \cdot \sum_{y=0}^{m} \left(\begin{array}{c} m \\ y \end{array} \right) \cdot p^y \cdot (1-p)^{(m-y)} = np.$$

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. Hence the mean of X is

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

□

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

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

First we prove a simpler result. Let X be a nonnegative random variable. Then, for any positive c,

$$Pr(X \geq c) \leq \frac{E(X)}{c}.$$

This is true because (for the continuous case)

$$\begin{aligned} E(X) &= \int_0^\infty x f_X(x) dx \\ &= \int_C^\infty x f_X(x) dx + \int_0^c x f_X(x) dx \\ &\geq \int_C^\infty x f_X(x) dx + \\ &\geq \int_C^\infty c f_X(x) dx = c \cdot Pr(X \geq c). \end{aligned}$$

Now apply this to $X = (Y - \mu)^2$ 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. □