POL 571: Expectation and Functions of Random Variables

Kosuke Imai
Department of Politics, Princeton University

March 10, 2006

1 Expectation and Independence

To gain further insights about the behavior of random variables, we first consider their *expectation*, which is also called *mean value* or *expected value*. The definition of expectation follows our intuition.

Definition 1 *Let X be a random variable and g be any function.*

1. *If X is discrete, then the expectation of $g(X)$ is defined as, then*

$$E[g(X)] = \sum_{x \in \mathcal{X}} g(x) f(x),$$

   where f is the probability mass function of X and $\mathcal{X}$ is the support of X.

2. *If X is continuous, then the expectation of $g(X)$ is defined as,*

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

   where f is the probability density function of X.

If $E(X) = -\infty$ or $E(X) = \infty$ (i.e., $E(|X|) = \infty$), then we say the expectation $E(X)$ does not exist. One sometimes write E_X to emphasize that the expectation is taken with respect to a particular random variable X. For a continuous random variable, the expectation is sometimes written as,

$$E[g(X)] = \int_{-\infty}^{x} g(x)\, d\, F(x).$$

where $F(x)$ is the distribution function of X. The expectation operator has inherits its properties from those of summation and integral. In particular, the following theorem shows that expectation preserves the inequality and is a linear operator.

Theorem 1 (Expectation) *Let X and Y be random variables with finite expectations.*

1. *If $g(x) \geq h(x)$ for all $x \in \mathbf{R}$, then $E[g(X)] \geq E[h(X)]$.*
2. *$E(aX + bY + c) = aE(X) + bE(Y) + c$ for any $a, b, c \in \mathbf{R}$.*

Let's use these definitions and rules to calculate the expectations of the following random variables if they exist.

Example 1

1. *Bernoulli random variable.*
2. *Binomial random variable.*
3. *Poisson random variable.*
4. *Negative binomial random variable.*
5. *Gamma random variable.*
6. *Beta random variable.*
7. *Normal random variable.*
8. **Cauchy distribution.** *A Cauchy random variable takes a value in* $(-\infty, \infty)$ *with the following symmetric and bell-shaped density function.*

$$f(x) = \frac{1}{\pi[1 + (x - \mu)^2]}.$$

The expectation of Bernoulli random variable implies that since an indicator function of a random variable is a Bernoulli random variable, its expectation equals the probability. Formally, given a set A, an indicator function of a random variable X is defined as,

$$1_A(X) = \begin{cases} 1 & \text{if} \quad X \in A \\ 0 & \text{otherwise} \end{cases}.$$

Then, it follows that $E[1_A(X)] = P(X \in A)$. In addition, as we might expect, the expectation serves as a good guess in the following sense.

Example 2 *Show that* $b = E(X)$ *minimizes* $E[(X - b)^2]$.

Finally, we emphasize that the independence of random variables implies the mean independence, but the latter does not necessarily imply the former.

Theorem 2 (Expectation and Independence) *Let* X *and* Y *be independent random variables. Then, the two random variables are mean independent, which is defined as,*

$$E(XY) = E(X)E(Y).$$

More generally, $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$ *holds for any function* g *and* h.

That is, the independence of two random variables implies that both the covariance and correlation are zero. But, the converse is not true. Interestingly, it turns out that this result helps us prove a more general result, which is that the functions of two independent random variables are also independent.

Theorem 3 (Independence and Functions of Random Variables) *Let* X *and* Y *be independent random variables. Then,* $U = g(X)$ *and* $V = h(Y)$ *are also independent for any function* g *and* h.

We will come back to various properties of functions of random variables at the end of this chapter.

2 Moments and Conditional Expectation

Using expectation, we can define the moments and other special functions of a random variable.

Definition 2 *Let X and Y be random variables with their expectations $\mu_X = E(X)$ and $\mu_Y = E(Y)$, and k be a positive integer.*

1. *The kth moment of X is defined as $E(X^k)$. If $k = 1$, it equals the expectation.*
2. *The kth central moment of X is defined as $E[(X - \mu_X)^k]$. If $k = 2$, then it is called the variance of X and is denoted by $\text{var}(X)$. The positive square root of the variance is called the standard deviation.*
3. *The covariance of X and Y is defined as $\text{cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$.*
4. *The correlation (coefficient) of X and Y is defined as $\rho_{XY} = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}$.*

The following properties about the variances are worth memorizing.

Theorem 4 (Variances and Covariances) *Let X and Y be random variables and $a, b \in \mathbf{R}$.*

1. $\text{var}(aX + b) = a^2\text{var}(X)$.
2. $\text{var}(aX + bY) = a^2\text{var}(X) + b^2\text{var}(Y) + 2ab\text{cov}(X, Y)$.
3. $\text{cov}(X, Y) = E(XY) - E(X)E(Y)$
4. $\text{var}(X) = E(X^2) - [E(X)]^2$.

The last property shows that the calculation of variance requires the second moment. How do we find moments of a random variable in general? We can use a function that generates moments of any order so long as they exist.

Definition 3 *Let X be a random variable with a distribution function F. The moment generating function of X is defined by*

$$M(t) = E(e^{tX}),$$

provided that this expectation exists for t in some neighborhood of 0. That is, $E(e^{tX}) < \infty$ for all $t \in (-\epsilon, \epsilon)$ with $\epsilon > 0$.

Here, we do not go into the details of the technical condition about the neighborhood of 0. We note, however, (without a proof) that this condition exists in order to avoid the situation where two random variables with different distributions can have exactly the same moments. If this condition is met, then the distribution of a random variable is uniquely determined. That is, if $M_X(t) = M_Y(t)$ for all t in some neighborhood of 0, then $F_X(u) = F_Y(u)$ for all u.

Before we give the key property of the moment generating function, we need some additional results from intermediate calculus. The following theorem shows that one can interchange the derivative and integral over a finite range.

Theorem 5 (Leibnitz's Rule) *Let $f(x, \theta)$, $a(\theta)$, and $b(\theta)$ be differentiable functions with respect to θ. If $-\infty < a(\theta), b(\theta) < \infty$, then,*

$$\frac{d}{d\theta}\int_{a(\theta)}^{b(\theta)} f(x, \theta)\, d\, x = f(b(\theta), \theta)b'(\theta) - f(a(\theta), \theta)a'(\theta) + \int_{a(\theta)}^{b(\theta)} \frac{\partial}{\partial\theta} f(x, \theta)\, d\, x,$$

where $a'(\theta)$ and $b'(\theta)$ are the derivatives of $a(\theta)$ and $b(\theta)$ with respect to θ.

Note that if $a(\theta)$ and $b(\theta)$ are constants, then only the last term of the RHS remains. What about the situation where the integral is taken over an infinite range? Unfortunately, this question cannot be answered without studying the measure theory, which is beyond the scope of this class. Therefore, we state the result without a proof.

Theorem 6 (Interchange of Integration and Differentiation) *Let $f(x, \theta)$ be a differentiable function with respect to θ. If the following conditions are satisfied,*

1. $\left| \frac{\partial}{\partial \theta} f(x, \theta) \big|_{\theta=\theta^*} \right| \leq g(x, \theta)$ *for all* $\theta^* \in (\theta - \epsilon, \theta + \epsilon)$ *for some* $\epsilon > 0$,
2. $\int_{-\infty}^{\infty} g(x, \theta)\, d\, x < \infty$,

then, the following equality holds,

$$\frac{d}{d\, \theta} \int_{-\infty}^{\infty} f(x, \theta)\, d\, x = \int_{-\infty}^{\infty} \frac{\partial}{\partial\, \theta} f(x, \theta)\, d\, x.$$

The conditions say that the first derivative of the function must be bounded by another function whose integral is finite. Now, we are ready to prove the following theorem.

Theorem 7 (Moment Generating Functions) *If a random variable X has the moment generating function $M(t)$, then*

$$E(X^n) = M^{(n)}(0),$$

where $M^{(n)}(t)$ is the nth derivative of $M(t)$.

The first question in the following example asks you to generalize the result we obtained earlier in this chapter.

Example 3

1. *Show that if X and Y are independent random variables with the moment generating functions $M_X(t)$ and $M_Y(t)$, then $Z = X + Y$ has the moment generating function, $M_Z(t) = M_X(t) M_Y(t)$.*
2. *Find a variance of the random variables in Example 1.*

Finally, we can also define the *conditional expectation*, $E(X \mid Y)$, and *conditional variance*, $E[(X - \mu_X)^2 \mid Y)]$, of a random variable X given another random variable Y. The expectation is over the conditional distribution, $f(X \mid Y)$. The *conditional covariance* of X and Y given X is similarly defined as $E[(X - \mu_X)(Y - \mu_Y) \mid Z]$ where the expectation is over $f(X, Y \mid Z)$. Theorem 2 implies that the conditional independence implies the conditional mean independence, but the latter does not imply the former. The conditional mean and variance have the following useful properties.

Theorem 8 (Conditional Expectation and Conditional Variance) *Let X and Y be random variables.*

1. *(Law of Iterated Expectation)* $E(X) = E[E(X \mid Y)]$.
2. $\text{var}(X) = E[\text{var}(X \mid Y)] + \text{var}[E(X \mid Y)]$.

These properties are useful when deriving the mean and variance of a random variable that arises in a hierarchical structure.

Example 4 *Derive the mean and variance of the following random variable X,*

$$\begin{aligned} X \mid n, Y &\sim \text{Binomial}(n, Y) \\ Y &\sim \text{Beta}(\alpha, \beta) \end{aligned}$$

3 Expectation and Inequalities

In this section, we learn some key equalities and inequalities about the expectation of random variables. Our goals are to become comfortable with the expectation operator and learn about some useful properties. The first theorem can be useful when deriving a lower bound of the expectation and when deriving an upper bound of a probability.

Theorem 9 (Chebychev's Inequality) *Let X be a random variable and let g be a nonnegative function. Then, for any positive real number a,*

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

When $g(X) = |X|$, it is called **Markov's inequality**. Let's use this result to answer the following question.

Example 5 *Let X be any random variable with mean μ and variance σ^2. Show that $P(|X - \mu| \geq 2\sigma) \leq 0.25$.*

That is, the probability that *any* random variable whose mean and variance are finite takes a value more than 2 standard deviation away from its mean is at most 0.25. Although this is a very general result, this bound is often very conservative. For example, if X is a normal random variable, this probability is approximately 0.05.

Next, we consider the inequality, which will be used again in POL 572.

Theorem 10 (Jensen's Inequality) *Let X be a random variable with $E(|X|) < \infty$. If g is a convex function, then*

$$E[g(X)] \geq g(E(X)),$$

provided $E(|g(X)|) < \infty$.

Note that if g is a concave function, then the inequality will be reversed, i.e., $E[g(X)] \leq g(E(X))$. This result is readily applicable to many commonly used functions.

Example 6 *Use Jensen's inequality to answer the following questions.*

1. *Establish the inequalities between the following pairs: (a) $E(X^2)$ and $[E(X)]^2$, (b) $E(1/X)$ and $1/E(X)$, and (c) $E[\log(X)]$ and $\log[E(X)]$.*

2. *Suppose $a_1, a_2, \ldots, a_n$ are positive numbers. Establish the inequalities among the following quantities,*

$$a_A = \frac{1}{n}\sum_{i=1}^{n} a_i, \quad a_G = \left(\prod_{i=1}^{n} a_i\right)^{1/n}, \quad a_H = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{a_i}},$$

   where a_A is the arithmetic mean, a_G is the geometric mean, and a_H is the harmonic mean.

The following theorem is a generalization of many important results.

Theorem 11 (Hölder's Inequality) *Let X and Y be random variables and $p, q \in (1, \infty)$ satisfying $1/p + 1/q = 1$. Then,*

$$E(|XY|) \leq [E(|X|^p)]^{1/p}[E(|Y|^q)]^{1/q}.$$

When $p = q = 2$, the inequality is called **Cauchy-Schwartz inequality**,

$$E(|XY|) \leq \sqrt{E(X^2)E(Y^2)},$$

a variant of which we studied in POL 502. We give a couple of applications of Hölder's inequality.

Example 7 *Prove the following statements where X and Y are random variables.*

1. **Covariance inequality.** $\text{cov}(X, Y) \leq \sqrt{\text{var}(X)\text{var}(Y)}$.
2. **Liapounov's inequality.** $[E(|X|^r)]^{1/r} \leq [E(|X|^s)]^{1/s}$ *where* $1 < r < s < \infty$.

Using Hölder's inequality, we can also prove the final theorem of this section, which we already proved once in POL 502.

Theorem 12 (Minkowski's Inequality) *Let X and Y be random variables. Then, for $1 \leq p < \infty$,*

$$[E(|X + Y|)^p]^{1/p} \leq [E(|X|^p)]^{1/p} + [E(|Y|^p)]^{1/p}.$$

When $p = 1$, we get something you are very familiar with; i.e., a version of **triangle inequality**!!

4 Functions of Random Variables

So far, we have been dealing with random variables themselves. Here, we study the functions of random variables and their distributions. Since a random variable is a function mapping the sample space to a real line, a function of random variables is also a random variable. We are interested in the distribution of such functions. First, we investigate the sums of random variables.

Theorem 13 (Convolution) *Let X and Y be random variables with the joint probability mass (density) function f.*

1. *If X and Y are discrete, then*

$$P(X+Y=z) = f_{X+Y}(z) = \sum_{x \in \mathcal{X}} f(x, z-x),$$

   where $\mathcal{X}$ is the support of X. Furthermore, if X and Y are independent, then $f_{X+Y}(z) = \sum_{x \in \mathcal{X}} f_X(x) f_Y(z-x) = \sum_{y \in \mathcal{Y}} f_X(z-y) f_Y(y)$, where $\mathcal{Y}$ is the support of Y.

2. *If X and Y are continuous, then*

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f(x, z-x)\, dx.$$

   Furthermore, if X and Y are independent, then $f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\, dx = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y)\, dy$.

Note that we have seen the moment generating functions of a covolution random variale. If X and Y are independent random variables with the moment generating functions, $M_X(t)$ and $M_Y(t)$, then the moment generating function of $Z = X+Y$ is given by $M_Z(t) = M_X(t) M_Y(t)$. Let's apply these results to the following examples.

Example 8

1. **Normal convolution.** *Let $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$. What is the distribution of $X+Y$?*

2. **Geometric convolution.** *Let X_i for $i = 1, 2, \ldots, n$ be independent and identical Geometric random variables with the probability mass function $f(x) = p(1-p)^{x-1}$. What is the distribution of $\sum_{i=1}^n X_i$?*

3. **Gamma convolution.** *Let X_i be independently distributed Gamma random variables with parameters α_i (can be different across i) and β (common across i). What is the distribution of $\sum_{i=1}^n X_i$?*

Now, consider a more general case. That is, we want to figure out the distribution of $Y = g(X)$ given a random variable X whose distribution is known. The following theorem generalizes the inverse CDF method we studied earlier,

Theorem 14 (Transformation of a Univariate Random Variable I) *Let X be a random variable with the distribution function $F_X(x)$. Define $Y = g(X)$ where g is a monotone function. Let also $\mathcal{X}$ and $\mathcal{Y}$ denote the support of distributions for X and Y, respectively.*

1. *If g is an increasing function, the distribution function of Y is given by*

$$F_Y(y) = F_X(g^{-1}(y))$$

for all $y \in \mathcal{Y}$.

2. *If g is a decreasing function and X is a continuous random variable, the distribution function of Y is given by*

$$F_Y(y) = 1 - F_X(g^{-1}(y))$$

for all $y \in \mathcal{Y}$.

3. *If X is a continuous random variable with the probability density function $f_X(x)$, then the probability density function of Y is given by*

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|,$$

for all $y \in \mathcal{Y}$.

Example 9 *Derive the probability function for the following transformations of a random variable.*

1. **Exponential distribution.** *$Y = -\log(X)$ where $X \sim Unif(0, 1)$.*
2. **Inverse Gamma distribution.** *$Y = 1/X$ where $X \sim Gamma(\alpha, \beta)$.*

Of course, we may be interested in non-monotone functions. In those cases, we need to partition the sample space so that within each partition the function is monotone. We state the following theorem without a proof.

Theorem 15 (Transformation of a Univariate Random Variable II) *Let X be a continuous random variable with the probability density function, f_X, and the sample space $\mathcal{X}$. Suppose there exists a partition, A_0, A_1, ..., A_n, of $\mathcal{X}$ such that $P(X \in A_0) = 0$ and f_X is continuous on each A_i for $i = 1, 2, \ldots, n$. Let $g_1(x), g_2(x), \ldots, g_n(x)$, be functions defined on $A_1, A_2, \ldots, A_n$, respectively. Suppose that these functions satisfy the following conditions:*

1. *$g(x) = g_i(x)$ for all $x \in A_i$ and each i,*
2. *$\mathcal{Y} = \{y : y = g_i(x) \text{ for some } x \in A_i\}$ is the same for each i,*
3. *$g_i(x)$ is monotone on A_i and $g_i^{-1}(y)$ has a continuous derivative on $\mathcal{Y}$ for each i.*

Then, for all $y \in \mathcal{Y}$,

$$f_Y(y) = \sum_{i=1}^{n} f_X(g_i^{-1}(y)) \left| \frac{d}{dy} g_i^{-1}(y) \right|.$$

The condition $P(X \in A_0) = 0$ exists so that the boundary points can be dealt with. Note that $g_i(x)$ is a one-to-one function mapping from A_i onto $\mathcal{Y}$, while $g_i^{-1}(y)$ is a one-to-one function mapping from $\mathcal{Y}$ onto A_i.

Example 10 χ^2 distribution. *Let X be the standard normal random variable. Derive the probability density function of $Y = X^2$.*

These rules can be easily generalized to the transformation of multivariate random variables. We state the result for bivariate random variables, but the similar theorems can be derived for the case of multivariate random variables.

Theorem 16 (Transformation of Bivariate Random Variables) *Let (X, Y) be two continuous random variables with the joint probability density function $f_{X,Y}(x, y)$. Consider a bijective transformation $U = g_1(X, Y)$ and $V = g_2(X, Y)$, and define its inverse as $X = h_1(U, V)$ and $Y = h_2(U, V)$. Then, the joint probability density function for (U, V) is given by*

$$f_{U,V}(u, v) = f_{X,Y}(h_1(u, v), h_2(u, v)) \, |J| \quad \text{where} \quad J = \begin{vmatrix} \frac{\partial}{\partial u} h_1(u, v) & \frac{\partial}{\partial_v} h_1(u, v) \\ \frac{\partial}{\partial u} h_2(u, v) & \frac{\partial}{\partial v} h_2(u, v) \end{vmatrix}$$

where J is called the Jacobian of the transformation.

You should see that the convolution theorem can be derived easily using this theorem by considering the joint distribution of $X = X$ and $Z = X + Y$ and then integrating out X. Similar to the univariate case, if the functions are not bijective, then one must partition the support of the joint distribution, $\{(x, y) : f(x, y) > 0\}$, such that within each partition, the functions are bijective, and the following rules can be used,

$$f_{U,V}(u, v) = \sum_{i=1}^{n} f_{X,Y}(h_{1i}(u, v), h_{2i}(u, v)) \, |J_i| \,,$$

Let's try some examples.

Example 11 *Derive the joint distribution of (U, V) in the following scenarios.*

1. *$U = X + Y$ and $V = X - Y$ where both X and Y are independent standard normal random variables.*
2. *$U = X + Y$ and $V = X/Y$ where both X and Y are independent exponential random variables with parameters α and β.*

We end this section with a key property of the multivariate normal distribution, which is a very important distribution in statistics.

Theorem 17 (Multivariate Normal Distribution) *Let $X = (X_1, X_2, \ldots, X_n)$ be an n-dimensional normal random vector with mean μ and variance Σ. Suppose $Y = DX$ for some $m \times n$ matrix D of rank $m \leq n$, then the distribution of Y is $N(D\mu, \, D\Sigma D^\top)$.*

We give a proof in the special case when $m = n$. This is a good time for you to review the notes on linear algebra from POL 502. Although the general proof is beyond the scope of this course, the theorem is extremely useful in many applied settings. Using this theorem, you can easily answer question 1 of Example 9 and question 1 of Example 11.