Moments or Expectation of Functions of Random Variables

Definition 1 *For a discrete random variable X define the expected value of X as*

$$E(X) = \sum_x xP(X = x)$$

provided this sum is well defined.

The sum is well defined provided it is either finite, $+\infty$ or $-\infty$. The only case it is not well defined is if

$$\sum_{x<0} xP(X = x) = -\infty$$

and

$$\sum_{x>0} xP(X = x) = \infty \ .$$

Suppose we consider a function of a discrete r.v. X, that is $Y = g(X)$. How can we calculate $\mathrm{E}(Y)$? According to Definition 1 we first need to find the distribution of Y. However there is a simpler calculation that can avoid this intermediate step of calculating the distribution of Y. Proposition 1 gives this result. Notice that the formula

$$\mathrm{E}(Y) = \sum_x g(x)P(X = x)$$

is not a new definition for expectation, it is just a consequence of the basic Definition 1.

Another consequence of this definition is that for a continuous r.v. X with pdf f

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} xf(x)dx$$

provided the integral is well defined (recall how such integrals are given in terms of limits of definite integrals). In particular this is not a new definition. It means that amongst other consequences when one needs to find properties of expectation one can usually just work with Definition 1. This derivation is not explored or studied here.

In fact the basic Definition 1 is the only definition we need. All the other properties of expectation are consequences of this one definition. We explore some of these here, but only prove these in the case of discrete random variables.

Proposition 1 *Suppose X is a discrete r.v. and $Y = g(X)$ for some function g. Suppose also that*

$$\sum_x |g(x)|P(X = x) < \infty \ .$$

Then

$$E(Y) = \sum_x g(x)P(X = x) \ .$$

Suppose X is a continuous r.v. and $Y = g(X)$ for some function g. Suppose also that

$$\int_{-\infty}^{\infty} |g(x)| f_X(x)dx < \infty \ .$$

Then

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

Proof : This is only proven for the discrete case.

In terms of Definition 1 we have

$$\begin{aligned}
\mathrm{E}(Y) &= \sum_y yP(Y = y) \\
&= \sum_y y \left\{ \sum_{x:g(x)=y} P(X = x) \right\} \\
&= \sum_x P(X = x) \left\{ \sum_{y:y=g(x)} y \right\} \\
&= \sum_x P(X = x)g(x) \\
&= \sum_x g(x)P(X = x)
\end{aligned}$$

The second line of this calculation computes the distribution of Y in terms of the pmf of X. The third line interchanges the order of summation, which is justified (see first year calculus) if the double sum is absolutely summable, which is implied by the condition in the Proposition. The fourth line calculates the inner part of the iterated sum, which is now a sum over a single value y after fixing x in the outer sum.
End of Proof.

Can we similarly calculate the expected value of functions of several discrete r.v.'s?

Proposition 2 *Suppose $X_1, X_2, \ldots, X_n$ are discrete r.v.s and $Y = g(X_1, X_2, \ldots, X_n)$ for some function g. Suppose also that*

$$\sum_{x_1} \sum_{x_2} \cdots \sum_{x_n} |g(x_1, \ldots, x_n)| P(X_1 = x_1, \ldots, X_n = x_n) < \infty \ .$$

Then

$$E(Y) = \sum_{x_1} \sum_{x_2} \cdots \sum_{x_n} g(x_1, \ldots, x_n) P(X_1 = x_1, \ldots, X_n = x_n) \ .$$

Suppose $X_1, X_2, \ldots, X_n$ are n-variate continuous r.v.s and $Y = g(X_1, X_2, \ldots, X_n)$ for some function g. Suppose also that

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} |g(x_1, \ldots, x_n)| f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) dx_n \ldots x_1 < \infty \ .$$

Then

$$E(Y) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \ldots, x_n) f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) dx_n \ldots x_1 \ .$$

The proof is completely analogous to the proof of Proposition 1.

Proposition 2 lets us avoid having to calculate the distribution of Y, which can be a considerable saving in work, depending on the calculation needed. The result is not a new definition of expectation, but only a consequence of the basic Definition 1.

There is a definition of expected value for a continuous random variable. It is also a consequence of Definition 1, but that is a relation or derivation is not discussed further here.

There are also analogues for expectations of functions of continuous random variables, such as Theorem 4.1.B, page 123 in the text. The proof for this is a technically more complicated than in the discrete case.

A further consequence of Definition 1 is that it is possible to find formula by which one can determine expectations for r.v.s which are neither discrete nor continuous or for bivariate r.v.s for which one component is discrete and one continuous. We do not explore these further in this course.

Suppose X, Y are bivariate discrete. How do we calculate $\mathrm{E}(X)$? Consider the function $g : (x, y) \mapsto x$. Thus $X = g(X, Y)$. By Proposition 2

$$\begin{aligned}
\mathrm{E}(X) &= \mathrm{E}(g(X,Y)) \\
&= \sum_x \sum_y g(x,y) P(X = x, Y = y) \\
&= \sum_x x \left\{ \sum_y P(X = x, Y = y) \right\} \\
&= \sum_x x P(X = x)
\end{aligned}$$

by the property of calculating the marginal distribution of X from the joint distribution of Y.

Continuing with this same example how do we calculate the expected value of XY? Consider the function $g : (x, y) \mapsto xy$. Thus $XY = g(X, Y)$. By Proposition 2

$$\begin{aligned}
\mathrm{E}(XY) &= \mathrm{E}(g(X,Y)) \\
&= \sum_x \sum_y g(x,y) P(X = x, Y = y) \\
&= \sum_x \sum_y xy P(X = x, Y = y)
\end{aligned}$$

How can we calculate the expected value of $aX + bY$ for some constants a and b? Consider the function $g : (x, y) \mapsto ax + by$. Thus $aX + bY = g(X, Y)$. By Proposition 2, and using an additional requirement that $\mathrm{E}(|X|) < \infty$ and $\mathrm{E}(|Y|) < \infty$, then

$$\begin{aligned}
\mathrm{E}(aX + bY) &= \mathrm{E}(g(X,Y)) \\
&= \sum_x \sum_y g(x,y) P(X = x, Y = y) \\
&= \sum_x \sum_y (ax + by) P(X = x, Y = y) \\
&= a \sum_x \sum_y x P(X = x, Y = y) + b \sum_x \sum_y y P(X = x, Y = y) \\
&= a \sum_x x P(X = x) + b \sum_y y P(Y = y)
\end{aligned}$$

Where do we use the fact that $\mathrm{E}(|X|)$ and $\mathrm{E}(|Y|)$ are finite?

Suppose that X and Y are independent. Suppose also $\mathrm{E}(X)$ and $\mathrm{E}(Y)$ are finite. Then

$$\begin{aligned}
\mathrm{E}(XY) &= \sum_x \sum_y xyP(X=x, Y=y) \\
&= \sum_x \sum_y xyP(X=x)P(Y=y) \\
&= \sum_x \left\{ xP(X=x) \left(\sum_y yP(Y=y) \right) \right\} \\
&= \sum_x \{xP(X=x)\mathrm{E}(Y)\} \\
&= \left\{ \sum_x xP(X=x) \right\} \mathrm{E}(Y) \\
&= \mathrm{E}(X)\mathrm{E}(Y)
\end{aligned}$$

Where is it required that $\mathrm{E}(X)$ and $\mathrm{E}(Y)$ are finite numbers?

The student should now try to find a formula to obtain $\mathrm{E}(h(X)g(Y))$ in the same setting. That is suppose that X and Y are independent, that $h(X)$ has finite expectation, and that $g(Y)$ has finite expectation, then prove that

$$\mathrm{E}(h(X)g(Y)) = \mathrm{E}(h(X))\mathrm{E}(g(Y)) \ .$$

Notation and terminology

- $\mathrm{E}(X^k)$ is called the k-th moment of the r.v. X, or more precisely of the distribution of X.
- Suppose X has mean $\mu = \mathrm{E}(X)$. The k-th central moment of X (or more precisely of the distribution of X) is

$$\mathrm{E}\left((X-\mu)^k\right)$$

Of course by definition the first central moment is 0. The second central moment is the variance. If a r.v. has mean 0, then the moments and central moments are the same.

Consider a standardized r.v. (with enough moments as needed for these terms below)

$$Y = \frac{X-\mu}{\sigma} \ .$$

Y has mean 0 and variance 1.

- Skewness

$$\mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{1}{\sigma^3}\mathrm{E}\left((X-\mu)^3\right)$$

- Kurtosis

$$\mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{1}{\sigma^4}\mathrm{E}\left((X-\mu)^4\right)$$

For a symmetric distribution with 3 moments, skewness = 0. Kurtotis is a measure often used as a test of heavy tails, that is heavy with respect to the normal distribution. Kurtosis for a normal distribution = 3; the student should evaluate this to verify the result.

Example 1

X = the number or trials (tosses) till the first success.

$X \sim$ Geometric(p) with pmf

$$f(k) = P(X = k) = (1-p)^{k-1}p,\ k = 1, 2, \ldots$$

where $0 < p < 1$.

Calculate

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

and

$$\mathrm{E}(X^2) = \sum_{k=1}^{\infty} k^2(1-p)^{k-1}p$$

To do this we use some properties of power series, in particular the power series

$$S(x) = \sum_{k=0}^{\infty} x^k = \frac{1}{1-x}$$

provided that $|x| < 1$. The radius of convergence for this power series is $\rho = 1$. If $|x| < \rho$, that is less than the radius of convergence, then one can integrate from 0 to x or differentiate at x term by term, and the new power series also has the same radius of convergence. Thus we obtain for example

$$S'(x) = \sum_{k=1}^{\infty} kx^{k-1} = \frac{dS(x)}{dx} = (1-x)^{-2}$$

and

$$S''(x) = \sum_{k=2}^{\infty} k(k-1)x^{k-2} = \frac{d^2S(x)}{dx^2} = 2(1-x)^{-3}$$

Note for example in the first derivative the term corresponding to $k = 0$ has derivative 0 and so the sum goes from 1 to ∞. Similarly for the second derivative the derivatives in the sum for $k = 0$ and $k = 1$ are both 0, so the sum goes from 2 to ∞.

Since $0 < p < 1$, thus

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

$$\begin{aligned}
\mathrm{E}(X^2) &= \left\{\sum_{k=1}^{\infty} k^2(1-p)^{k-1}\right\} p \\
&= \left\{\sum_{k=1}^{\infty} k(k-1+1)(1-p)^{k-1}\right\} p \\
&= \left\{\sum_{k=1}^{\infty} k(k-1)(1-p)^{k-2}\right\} (1-p)p + \left\{\sum_{k=1}^{\infty} (1-p)^{k-1}\right\} p
\end{aligned}$$

$$\begin{aligned} &= \frac{2}{(1-(1-p))^3}(1-p)p + \frac{1}{p} \\ &= \frac{2(1-p)+p}{p^2} = \frac{2-p}{p^2} \end{aligned}$$

Thus

$$\mathrm{Var}(X) = \frac{2-p}{p^2} - \left(\frac{1}{p}\right)^2 = \frac{1-p}{p^2}$$

Example 2

Consider a negative binomial X with parameters r and p. Suppose also that r is an integer.

As with the geometric distribution there are two equivalent r.v.s called the negative binomial r.v. Here we consider X to be the number of trials, for iid Bernoulli p, till the r-th success. Thus X has pmf

$$P(X=k) = \binom{k-1}{r-1} p^{r-1}(1-p)^{k-r} \ , \ k = r, r+1, r+2, \ldots$$

X has an equivalent representation of

$$X = Y_1 + Y_2 + \ldots + Y_r$$

where Y_j are iid geometric p, where Y_j is the number of trials till the first success in iid Bernoulli p trials. See the example above.

Thus

$$\mathrm{E}(X) = \mathrm{E}(Y_1) + \mathrm{E}(Y_2) + \ldots + \mathrm{E}(Y_r)$$

and

$$\begin{aligned} \mathrm{Var}(X) &= \mathrm{Var}(Y_1 + Y_2 + \ldots + Y_r) \\ &= \sum_{j=1}^{r} \mathrm{Var}(Y_j) + \sum_{j \neq k} \mathrm{Cov}(Y_j, Y_k) \\ &= r\mathrm{Var}(Y_1) \end{aligned}$$

The student should finish this example at home.

Example 3 See the bivariate normal handout

Example 4 Gamma distribution

Suppose $X \sim \mathrm{Gamma}(\alpha, \lambda)$

$$\begin{aligned} \mathrm{E}(X) &= \int_{-\infty}^{\infty} x f(x) dx \\ &= \int_0^{\infty} x \frac{1}{\Gamma(\alpha)} x^{\alpha-1} \lambda^{\alpha} e^{-\lambda x} dx \\ &= \frac{1}{\Gamma(\alpha)} \int_0^{\infty} x^{\alpha+1-1} \lambda^{\alpha} e^{-\lambda x} dx \end{aligned}$$

$$\begin{aligned} &= \frac{\Gamma(\alpha+1)}{\lambda\Gamma(\alpha)}\int_0^\infty \frac{1}{\Gamma(\alpha+1)}x^{\alpha+1-1}\lambda^{\alpha+1}e^{-\lambda x}dx \\ &= \frac{\alpha}{\lambda}\times 1 \\ &= \frac{\alpha}{\lambda} \end{aligned}$$

At home find a formula for $\mathrm{E}(X^2)$, $\mathrm{Var}(X)$ and $\mathrm{E}(X^k)$ for integers $k \geq 1$.

Example 5 Lognormal

Y is said to have a log normal distribution if it has pdf

$$f_Y(y) = \frac{1}{y\sigma\sqrt{2\pi}}e^{-\frac{(\log(y)-\mu)^2}{2\sigma^2}}I(y>0)$$

In fact Y is said to have a log normal distribution if and only if it is of the form $Y = e^X$ where $X \sim N(\mu, \sigma^2)$.

We may calculate

$$\mathrm{E}(Y) = \int_0^\infty y\frac{1}{y\sigma\sqrt{2\pi}}e^{-\frac{(\log(y)-\mu)^2}{2\sigma^2}}dy = \int_0^\infty \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(\log(y)-\mu)^2}{2\sigma^2}}dy$$

but on initial inspection this is not straightforward, or at least requires some change of variables.

We may also calculate

$$\begin{aligned} \mathrm{E}(Y) &= \mathrm{E}\left(e^X\right) \\ &= \int_{-\infty}^\infty e^x\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx \\ &= \int_{-\infty}^\infty \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left\{\left(\frac{(x-\mu}{\sigma}\right)^2-2x\right\}}dx \quad (\text{now set } w = \frac{x-\mu}{\sigma}) \\ &= \int_{-\infty}^\infty \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left\{w^2-2(\sigma w+\mu)\right\}}\sigma dw \\ &= \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left\{w^2-2\sigma w+\sigma^2-\sigma^2-2\mu)\right\}}dw \\ &= \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\{w-\sigma\}^2}dwe^{-\frac{1}{2}\left(-\sigma^2-2\mu\right)} \\ &= e^{\frac{1}{2}\sigma^2+\mu} \end{aligned}$$

since the integral evaluates to 1, because it has integrand which is the pdf of a $N(\sigma, 1)$ distribution.

The calculation above is also a special case of moment generating function calculations, and the MGF for normals is known and the student will calculate it at that time.

Chebyshev's Inequality

This inequality is useful since it gives an upper bound for tail probabilities that uses no special properties of the distribution but only the second moment. An exact calculation would involve calculating quantiles. This upper bound is easier to calculate and may be much larger than the true probability. However for some purposes it gives a good enough answer. Some of these ideas are discussed later.

X is a random variable with mean μ and variance σ^2.

In the continuous case with $X \sim f$, and $a > 0$

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{E}\left((X-\mu)^2\right) \\
&= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx \\
&\geq \int_{\{x:|x-\mu|>a\}} (x-\mu)^2 f(x)dx \\
&\geq \int_{\{x:|x-\mu|>a\}} a^2 f(x)dx \\
&= a^2 \int_{\{x:|x-\mu|>a\}} f(x)dx \\
&= a^2 P(|X-\mu|>a)
\end{aligned}$$

Thus we obtain

$$\frac{\mathrm{Var}(X)}{a^2} \geq P(|X-\mu|>a) \tag{1}$$

Relation (1) is also true for any other random (discrete or continuous) with a finite mean and variance. It is known as Chebyshev's inequality; see text Section 4.2.

One of its important immediate consequences is in the special case when $\sigma^2 = \mathrm{Var}(X) = 0$. See Corollary A, p 134.

Below is a proof for this Corollary.

For any number $\epsilon > 0$ we have

$$P(|X-\mu|>\epsilon) \leq \frac{\sigma^2}{\epsilon^2} = \frac{0}{\epsilon^2} = 0 \ .$$

Therefore

$$\begin{aligned}
P(X=\mu) &= P(|X-\mu|=0) \\
&= 1 - P(|X-\mu|>0) \\
&= 1 - \lim_{\epsilon \to 0^+} P(|X-\mu|>\epsilon) \\
&= 1 - \lim_{\epsilon \to 0^+} 0 \\
&= 1 - 0 = 1 \ .
\end{aligned}$$

Covariance and correlation are useful and important notions.

Covariance

$$\mathrm{Cov}(X,Y) = \mathrm{E}\left((X-\mu_X)(Y-\mu_Y)\right)$$

Correlation

Consider the standardized rv's

$$Z_1 = \frac{X-\mu_X}{\sigma_X} \ , \ Z_2 = \frac{Y-\mu_Y}{\sigma_Y}$$

Then

$$\mathrm{corr}(X,Y) = \rho(X,Y) = \mathrm{Cov}(Z_1, Z_2)$$

that is correlation is the covariance of the two standardized r.v.s. Some texts use the notation $\rho(X,Y)$ for correlation and some use $\mathrm{corr}(X,Y)$. From this we can also find that

$$\mathrm{corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} \ .$$

Proposition 3 *Suppose X, Y have finite second moments. The correlation of r.v.s X, Y, say ρ, then satisfies $-1 \leq \rho \leq 1$.*

Corollary 1 *Under the conditions of Proposition 3 then*

- *If $\rho = 1$ then*

  $$Y = aX + b$$

  for constants a, b and $a > 0$

- *If $\rho = -1$ then*

  $$Y = aX + b$$

  for constants a, b and $a < 0$

Proof We only do the first part, and the student should do the second part.

$$\mathrm{Var}\left\{\frac{X-\mu_X}{\sigma_X} - \frac{Y-\mu_Y}{\sigma_Y}\right\} = \mathrm{Var}\left\{\frac{X-\mu_X}{\sigma_X}\right\} - 2\rho(X,Y) + \mathrm{Var}\left\{\frac{Y-\mu_Y}{\sigma_Y}\right\} = 2 - 2\rho = 0$$

Therefore noting that

$$0 = \mathrm{E}\left\{\frac{X-\mu_X}{\sigma_X} - \frac{Y-\mu_Y}{\sigma_Y}\right\}$$

then by Proposition 3 we have, with probability 1,

$$\frac{X-\mu_X}{\sigma_X} - \frac{Y-\mu_Y}{\sigma_Y} = 0$$

Rearranging this gives

$$Y = \mu_Y + \frac{\sigma_Y}{\sigma_X}X - \frac{\sigma_Y}{\sigma_X}\mu_X = \mu_Y - \frac{\sigma_Y}{\sigma_X}\mu_X + \frac{\sigma_Y}{\sigma_X}X \ .$$

This is of the required form with

$$a = \frac{\sigma_Y}{\sigma_X} \ , \ \mu_Y - \frac{\sigma_Y}{\sigma_X}\mu_X \ .$$

End of Proof

The student should work through the correlation $\rho = -1$ case.

If X, Y are independent and have finite second moments then $\text{Cov}(X, Y) = 0$ and $\text{corr}(X, Y) = 0$. However if $\text{Cov}(X, Y) = 0$ ($\text{corr}(X, Y) = 0$) it does not imply that X, Y are independent. We have various examples to show this. However in the normal case, we have independence if and only if the correlation equals 0. See bivariate normal handout.