Chapter 6. Expectation for discrete random variables.

6.1 Expected value.

If X is a discrete r.v. with finite sample space $\Omega_X = \{x_1, \ldots, x_N\}$ and p.m.f. f_X, then as the name suggests we can think of the probabilities $f_X(x_1), \ldots, f_X(x_N)$ as masses located at points $x_1, \ldots, x_N$ on a segment of the number line. With this interpretation the center of gravity of the distribution of X is located at $E(X) = x_1 f_X(x_1) + \cdots + x_N f_X(x_N)$ (read this as the expected value of X), *i.e.*, $E(X)$ is the weighted average of the possible values of X with weights given by their probabilities. If an individual is selected at random from a population of people and X represents the individual's height rounded to the nearest whole inch so that Ω_X represents the collection of distinct heights for this population and $f_X(x_i)$ is the proportion of individuals in the population with height x_i, then $E(X)$ is the population mean (average) height. If we think of X as the winnings of a player (with a negative value indicating a loss) in one play of a game of chance with f_X indicating the probabilities of the possible outcomes of the game, then $E(X)$ is the player's expected winnings. For example in the game of craps discussed earlier if we let the dichotomous variable X indicate the player's winnings ($X = 1$ indicating the player wins \$1 and $X = -1$ indicating the player loses \$1), then the player's expected winnings is $E(X) = (1)\left(\frac{244}{495}\right) + (-1)\left(\frac{251}{495}\right) = \frac{-7}{495} \approx -.01414$. This means that on average in the long run the player loses 1.414 cents per game and the house (casino) wins an average of 1.414 cents per game. Thus if the house offers equal odds on this bet at craps (meaning that if the player bets \$1 then he either wins \$1 or loses \$1), then the house can make money provided a large number of such bets are made.

Definition. *If X is a discrete r.v. with finite sample space $\Omega_X = \{x_1, \ldots, x_N\}$ and p.m.f. f_X, then the expected value of X is defined by*

$$E(X) = \sum_{x_i \in \Omega_X} x_i f_X(x_i).$$

The expected value of X is also known as the mean of X and the expectation of X. Furthermore, these terms are also applied to the distribution of X, *e.g.*, $E(X)$ is the mean of the distribution of X.

If Ω_X is countably infinite, then the series $\sum_{x_i \in \Omega_X} x_i f_X(x_i)$ may not converge. Thus some technicalities need to be considered before we provide a general definition of the expected value of a discrete r.v. For example, it can be shown that $\sum_{x=1}^{\infty} \frac{1}{x^2} = \frac{\pi^2}{6}$, thus $f_X(x) = \frac{6}{\pi^2 x^2}\mathbf{1}_{\{1,2,\ldots\}}(x)$ is a valid p.m.f. on $\Omega_X = \{1, 2, \ldots\}$. However, for this distribution the series $\sum_{x=1}^{\infty} x\frac{6}{\pi^2 x^2} = \frac{6}{\pi^2}\sum_{x=1}^{\infty}\frac{1}{x}$ clearly diverges, since this is a multiple of the harmonic series. If Ω_X is countably infinite and Ω_X contains positive and negative values,

then the series $\sum_{x_i \in \Omega_X} x_i f_X(x_i)$ may fail to exist because it is not well defined in the sense that the value of the sum may depend on the order in which the terms are summed. An infinite series $\sum_{x=1}^{\infty} a_x$ is said to be absolutely convergent if $\sum_{x=1}^{\infty} |a_x|$ exists. If the series $\sum_{x_i \in \Omega_X} x_i f_X(x_i)$ is absolutely convergent, *i.e.*, if $\sum_{x_i \in \Omega_X} |x_i| f_X(x_i)$ converges, then the series $\sum_{x_i \in \Omega_X} x_i f_X(x_i)$ is well defined and converges.

Definition. *If X is a discrete r.v. with sample space Ω_X and p.m.f. f_X and if the series $\sum_{x_i \in \Omega_X} x_i f_X(x_i)$ is absolutely convergent, then the expected value of X is defined by*

$$E(X) = \sum_{x_i \in \Omega_X} x_i f_X(x_i).$$

If the series $\sum_{x_i \in \Omega_X} x_i f_X(x_i)$ is not absolutely convergent, then the expected value of X does not exist.

The definition of the expected value of X can be extended to expected values of functions of X. If X is a discrete r.v. with sample space Ω_X and g is a function mapping Ω_X onto Ω_Y, then $Y = g(X)$ is a discrete r.v. and if $E(Y)$ exists, then it is easy to see that

$$E(Y) = \sum_{y \in \Omega_Y} y f_Y(y) = \sum_{x \in \Omega_X} g(x) f_X(x) = E(g(X)).$$

We can also consider expected values of functions of two or more r.v.'s. For example, if X and Y are jointly discrete r.v.'s with joint sample space $\Omega_{X,Y}$ and g is a function mapping $\Omega_{X,Y}$ onto Ω_W, then $W = g(X, Y)$ is a discrete r.v. and if $E(W)$ exists, then it is easy to see that

$$E(W) = \sum_{w \in \Omega_W} w f_W(w) = \sum_{(x,y) \in \Omega_{X,Y}} g(x, y) f_{X,Y}(x, y) = E(g(X, Y)).$$

Theorem 6.1. *If a is a constant, X and Y are discrete r.v.'s, and $E(X)$ and $E(Y)$ exist, then*
1) $E(a) = a$, i.e., if $Pr(X = a) = 1$, then $E(X) = a$.
2) If $Pr(X \geq a) = 1$, then $E(X) \geq a$. Similarly, if $Pr(X \leq a) = 1$, then $E(X) \leq a$.
3) $E(aX) = aE(X)$.
4) $E(a + X) = a + E(X)$.
5) $E(X + Y) = E(X) + E(Y)$.

Proof. Let a, X, and Y as specified be given.
1) Assume that $\Pr(X = a) = 1$. Then $\Omega_X = \{a\}$ and $E(X) = a f_X(a) = a$.

2) Assume that $\Pr(X \geq a) = 1$. Then $E(X) = \sum_{x \in \Omega_X} x f_X(x) \geq \sum_{x \in \Omega_X} a f_X(x) = a$.

3) $E(aX) = \sum_{x\in\Omega_X} axf_X(x) = a\sum_{x\in\Omega_X} xf_X(x) = aE(X)$.

4) $E(a+X) = \sum_{x\in\Omega_X}(a+x)f_X(x) = \sum_{x\in\Omega_X} af_X(x) + \sum_{x\in\Omega_X} xf_X(x) = a + E(X)$.

5) $E(X+Y) = \sum_{(x,y)\in\Omega_{X,Y}}(x+y)f_{X,Y}(x,y)$
$= \sum_{(x,y)\in\Omega_{X,Y}} xf_{X,Y}(x,y) + \sum_{(x,y)\in\Omega_{X,Y}} yf_{X,Y}(x,y)$
$= \sum_{x\in\Omega_X} x\sum_{y\in\Omega_Y} f_{X,Y}(x,y) + \sum_{y\in\Omega_Y} y\sum_{x\in\Omega_X} f_{X,Y}(x,y)$
$= \sum_{x\in\Omega_X} xf_X(x) + \sum_{y\in\Omega_Y} yf_Y(y) = E(X) + E(Y)$. □

6.2 Moments.

For a r.v. X and a nonnegative integer k, the k^{th} (raw) moment of X (k^{th} moment of the distribution of X) is $E(X^k)$, and the k^{th} factorial moment of X is $E[X(X-1)\cdots(X-k+1)]$ provided these expectations exist. If $\mu_X = E(X)$ exists, then the k^{th} central moment of X is $E[(X-\mu_X)^k]$ provided this expectation exists.

Definition. *If X is a discrete r.v. with sample space Ω_X and p.m.f. f_X, then for a nonnegative integer k:*
1) If the indicated expectation exists, then the k^{th} raw moment of X is

$$E(X^k) = \sum_{x\in\Omega_X} x^k f_X(x);$$

2) If the indicated expectation exists, then the k^{th} factorial moment of X is

$$E[X(X-1)\cdots(X-k+1)] = \sum_{x\in\Omega_X} x(x-1)\cdots(x-k+1)f_X(x);$$

3) If $\mu_X = E(X)$ exists and the indicated expectation exists, then the k^{th} central moment of X is

$$E[(X-\mu_X)^k] = \sum_{x\in\Omega_X} (x-\mu_X)^k f_X(x).$$

Note that if the k^{th} moment (raw,factorial or central) exists, then every moment of lower order also exists.

For a nonnegative integer valued r.v. X there is an interesting and useful connection between the factorial moments of X and the p.g.f. of X. Recall that, letting $p_x = f_X(x)$ for $x = 0, 1, \ldots$, the probability generating function of the nonnegative integer valued r.v. X is

$$P_X(t) = \sum_{x=0}^{\infty} t^x p_x = p_0 + tp_1 + t^2 p_2 + \cdots$$

and this series converges absolutely at least for $-1 \leq t \leq 1$. Note that $P_X(t) = E(t^X)$ and, as mentioned earlier, the derivative of this series is

$$P'_X(t) = \sum_{x=1}^{\infty} x t^{x-1} p_x = p_1 + 2tp_2 + 3t^2 p_3 + \cdots$$

and $P'_X(1) = \sum_{x=1}^{\infty} x p_x$. If $E(X)$ exists, then this series expansion of $P'_X(t)$ is a continuous function of t for $-1 \leq t \leq 1$ and $P'_X(1) = E(X)$. If $E(X)$ does not exist, then the series $P'_X(1)$ diverges. The second derivative of $P_X(t)$ is

$$P''_X(t) = \sum_{x=2}^{\infty} x(x-1) t^{x-2} p_x = 2p_2 + 6tp_3 + 12t^3 p_4 + \cdots$$

and if $E[X(X-1)]$ exists, then $P''_X(1) = E[X(X-1)]$. This differentiation process leads to the following result. If the k^{th} factorial moment $E[X(X-1)\cdots(X-k+1)]$ exists, then $P_X^{(k)}(1) = E[X(X-1)\cdots(X-k+1)]$. This approach to computing factorial moments from the p.g.f. is often the easiest way to compute the moments and related expectations of positive integer valued X. For example, $E[X(X-1)] = E(X^2) - E(X)$ thus $E(X^2) = P''_X(1) + P'_X(1)$.

6.3 Variance.

Letting $E(X) = \mu_X$ (the mean of the distribution of X) and assuming that this mean exists, the random variable $g(X) = (X - \mu_X)^2$ (the squared deviation of X from the mean of the distribution of X) can be used as a measure of the variability in the distribution of X. For example, in our example with X denoting the height of an individual selected at random from a finite population of people the r.v. $(X - \mu_X)^2$ is one way to measure the extent to which the individual's height differs from the population mean height. The expected value of this r.v. is the variance of the distribution of X which is defined more formally below. Note that the variance of the distribution of X is the second central moment of the distribution of X.

Definition. *If X is a discrete r.v. with sample space Ω_X, p.m.f. f_X, and mean $\mu_X = E(X)$, then the variance of X is defined by*

$$\text{var}(X) = \sum_{x \in \Omega_X} (x - \mu_X)^2 f_X(x),$$

provided this series converges. If the series does not converge, then the variance of X does not exist. The principal square root of $\text{var}(X)$ is know as the standard deviation of X. Note that, if $\text{var}(X)$ exists, then $\text{var}(X) = E(X^2) - [E(X)]^2$.

Theorem 6.2. *If a is a constant, X is a discrete r.v., and var(X) exists, then*
1) var$(a) = 0$, i.e., if $Pr(X = a) = 1$, then var$(X) = 0$.
2) var$(aX) = a^2$var(X).
3) var$(a + X) =$ var(X).
4) var$(X) = 0$ if and only if there is a constant a such that $Pr(X = a) = 1$.

Proof. Let a and X as specified be given.
1) Since $E(a) = a$ we have $\text{var}(a) = E[(a - a)^2] = E(0) = 0$.

2) Since $E(aX) = aE(X)$ we have $\text{var}(aX) = E([aX - aE(X)]^2) = E(a^2[X - E(X)]^2) = a^2\text{var}(X)$.

3) Since $E(a + X) = a + E(X)$ we have $\text{var}(a + X) = E[(a + X - [a + E(X)])^2] = E([X - E(X)]^2) = \text{var}(X)$.
4) Suppose that there is a constant a such that $\Pr(X = a) = 1$, then $\text{var}(X) = \text{var}(a) = 0$ by part 1). Now suppose that $\text{var}(X) = 0$, then $\sum_{x\in\Omega_X}[x - E(X)]^2 f_X(x) = 0$ but this requires that $x = E(X)$ for all $x \in \Omega_X$ which is equivalent to saying that the only value of x for which $f_X(x) > 0$ is $E(X)$. This clearly implies that $\Pr[X = E(X)] = 1$ and the result holds with $a = E(X)$. □

6.4 Examples.

Binomial distribution. If $X \sim$ binomial(n, p), then
$E(X) = np$ and $\text{var}(X) = npq$.

Geometric distribution. If $X \sim$ geometric(p), then
$E(X) = \frac{q}{p}$ and $\text{var}(X) = \frac{q}{p^2}$.

Negative binomial distribution. If $X \sim$ negative binomial(n, p), then
$E(X) = \frac{rq}{p}$ and $\text{var}(X) = \frac{rq}{p^2}$.

Hypergeometric distribution. If $X \sim$ hypergeometric(N_1, N_2, n), then
$E(X) = n\left(\frac{N_1}{N_1+N_2}\right)$ and $\text{var}(X) = n\left(\frac{N_1}{(N_1+N_2)}\right)\left(\frac{N_2}{(N_1+N_2)}\right)\left(\frac{N_1+N_2-n}{N_1+N_2-1}\right)$.

Discrete uniform distribution. If $X \sim$ uniform$(\{1, 2, \ldots, N\})$, then
$E(X) = \frac{N+1}{2}$ and $\text{var}(X) = \frac{N^2-1}{12}$.

Poisson distribution. If $X \sim$ Poisson(λ), then
$E(X) = \lambda$ and $\text{var}(X) = \lambda$.

6.5 Covariance.

We begin by defining the covariance of the r.v.'s X and Y, $\text{cov}(X, Y) = \text{cov}(Y, X)$, which arises in the computation of the variance of their sum $X + Y$.

Definition. *If X and Y are discrete r.v.'s for which var(X) and var(Y) exist, then the covariance of X and Y is*

$$cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = E(XY) - \mu_X\mu_Y,$$

where $\mu_X = E(X)$ and $\mu_Y = E(Y)$.

Theorem 6.3. *If X and Y are independent discrete r.v.'s for which var(X) and var(Y) exist, then cov$(X,Y) = 0$.*

Proof. Let X and Y as specified be given. It is easy to see that $W = X - \mu_X$ and $Z = Y - \mu_Y$ are also independent. Therefore $\text{cov}(X,Y) = E(WZ) = E(W)E(Z) = 0$. □

It is important to note that the converse of this theorem is not true, *i.e.*, in general $\text{cov}(X,Y) = 0$ does not imply that X and Y are independent.

Theorem 6.4. *If X and Y are discrete r.v.'s for which var(X) and var(Y) exist, then*

$$var(X+Y) = var(X) + 2cov(X,Y) + var(Y).$$

Proof. Let X and Y as specified be given. Then

$$\begin{aligned}\text{var}(X+Y) &= E([(X+Y)-(\mu_X+\mu_Y)]^2) = E([(X-\mu_X)+(Y-\mu_Y)]^2)\\ &= E[(X-\mu_X)^2] + 2E[(X-\mu_X)(Y-\mu_Y)] + E[(Y-\mu_Y)^2]\\ &= \text{var}(X) + 2\text{cov}(X,Y) + \text{var}(Y).\square\end{aligned}$$

Theorem 6.5. *If $X_1, \ldots, X_n$ are discrete r.v.'s with variances $\sigma_1^2, \ldots, \sigma_n^2$, then*

$$var(X_1 + \cdots + X_n) = \sum_{i=1}^{n} \sigma_i^2 + 2\sum_{i<j} cov(X_i, X_j).$$

Corollary 6.5. *If $X_1, \ldots, X_n$ are independent discrete r.v.'s with variances $\sigma_1^2, \ldots, \sigma_n^2$, then*

$$var(X_1 + \cdots + X_n) = \sum_{i=1}^{n} \sigma_i^2.$$

Theorem 6.6 (Schwarz inequality). *If X and Y are discrete r.v.'s for which $E(X^2)$ and $E(Y^2)$ exist, then*

$$[E(XY)]^2 \le E(X^2)E(Y^2)$$

with equality if and only if there is a constant c such that $Pr(Y = cX) = 1$.

Proof. Let X and Y as specified be given. Let $g(t) = E[(tX - Y)^2]$. Since $g(t) \ge 0$ for all t the discriminant of the quadratic $g(t) = t^2E(X^2) + 2tE(XY) + E(Y^2)$ must be

nonpositive, *i.e.*, we must have $[E(XY)]^2 \leq E(X^2)E(Y^2)$. Furthermore, since $g(t) \geq 0$ for all t, there is a t_0 for which $g(t_0) = 0$ if and only if the discriminant is zero, *i.e.*, if and only if $[E(XY)]^2 = E(X^2)E(Y^2)$. This establishes the result since $g(t_0) = E[(t_0X - Y)^2] = 0$ if and only if $\Pr(Y = t_0X) = 1$. □

Theorem 6.7. *If X and Y are discrete r.v.'s for which var(X) and var(Y) exist, then*

$$|cov(X,Y)| \leq \sqrt{var(X)var(Y)}.$$

Proof. This follows from the Schwarz inequality applied to $X - \mu_X$ and $Y - \mu_Y$. □

If we standardize the r.v.'s X and Y to have mean zero and variance one (by subtracting the mean and dividing by the standard deviation) and then compute the covariance between these standardized r.v.'s we obtain the correlation of X and Y

$$\rho(X,Y) = E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right].$$

Note that $-1 \leq \rho(X,Y) \leq 1$. The r.v.'s X and Y are said to be uncorrelated when $\rho(X,Y) = 0$, which is equivalent to $\text{cov}(X,Y) = 0$. When $\rho(X,Y) = 1$, X and Y are said to be perfectly positively correlated and when $\rho(X,Y) = -1$, X and Y are said to be perfectly negatively correlated.

Theorem 6.8. *(Markov inequality) If a is a positive constant and X is a positive valued discrete r.v. for which $E(X)$ exists, then*

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

Proof. Let X and a which satisfy the hypothesis be given. Then
$E(X) = \sum_{\{x\in\Omega_X : x<a\}} x f_X(x) + \sum_{\{x\in\Omega_X : x\geq a\}} x f_X(x) \geq \sum_{\{x\in\Omega_X : x\geq a\}} x f_X(x)$
$\geq \sum_{\{x\in\Omega_X : x\geq a\}} a f_X(x) = a\Pr(X \geq a)$, since $x \geq a$ on this region.
Thus $\Pr(X \geq a) \leq \frac{E(X)}{a}$. □

Theorem 6.9 (Chebyshev's inequality). *If a is a positive constant and X is a discrete r.v. for which $E(X) = \mu_X$ and var$(X) = \sigma_X^2$ exist, then*

$$Pr\left(|X - \mu_X| \geq a\right) \leq \frac{\sigma_X^2}{a^2}.$$

Proof. Let X and a which satisfy the hypothesis be given and let $Y = (X - \mu_X)^2$. Note that Y is a positive valued r.v., $E(Y) = \sigma_X^2$, and

$$\Pr(Y \geq a^2) = \Pr\left((X-\mu_X)^2 \geq a^2\right) = \Pr\left(|X - \mu_X| \geq a\right)$$

Thus Theorem 6.8 implies that

$$\Pr\left(|X-\mu_X| \geq a\right) \leq \frac{E(Y)}{a^2} = \frac{\sigma_X^2}{a^2}. \ \square$$

Theorem 6.10 The weak law of large numbers. *Let $X_1, X_2, \ldots$ denote a sequence of independent discrete r.v.'s with a common distribution for which the mean μ and variance σ^2 exist. Let $S_n = X_1 + \cdots + X_n$ denote the sum of the first n r.v.'s. Then, for any $\epsilon > 0$, as $n \to \infty$*

$$Pr\left(\left|\frac{S_n}{n} - \mu\right| < \epsilon\right) \to 1.$$

In words this indicates that by increasing the number of terms n in the average $\frac{S_n}{n}$, the probability that this average is arbitrarily close to the mean μ approaches one.

Proof. Let $S_n = \sum_{i=1}^{n} X_i$ denote the sum of the first n terms of the sequence $\{X_i\}$ of independent discrete r.v.'s with a common distribution for which the mean μ and variance σ^2 exist and let $\epsilon > 0$ be given. Since $E(S_n) = n\mu$ and $\text{var}(S_n) = n\sigma^2$ application of the Chebyshev inequality with $a = \epsilon$ yields

$$\Pr\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}.$$

The result follows from the fact that the bound $\frac{\sigma^2}{n\epsilon^2}$ tends to zero as n tends to infinity. $\square$