MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.436J/15.085J Fall 2018
Lecture 6

MORE ON DISCRETE RANDOM VARIABLES AND THEIR EXPECTATIONS

Contents

1 COMMENTS ON EXPECTED VALUES

(a) Recall that $\mathbb{E}[X]$ is well defined unless both sums $\sum_{x:x<0} xp_X(x)$ and $\sum_{x:x>0} xp_X(x)$ are infinite. Furthermore, $\mathbb{E}[X]$ is well-defined and finite if and only if both sums are finite. This is the same as requiring that

$$\mathbb{E}[|X|] = \sum_x |x| p_X(x) < \infty.$$

Random variables that satisfy this condition are called **integrable**.

(b) Note that for any random variable X, $\mathbb{E}[X^2]$ is always well-defined (whether finite or infinite), because all the terms in the sum $\sum_x x^2 p_X(x)$ are nonnegative. If we have $\mathbb{E}[X^2] < \infty$, we say that X is **square integrable**.

(c) Using the inequality $|x| \leq 1 + x^2$, we have $\mathbb{E}[|X|] \leq 1 + \mathbb{E}[X^2]$, which shows that a square integrable random variable is always integrable. Similarly, for every positive integer r, if $\mathbb{E}[|X|^r]$ is finite then it is also finite for every $l < r$ (fill details).

Exercise 1. *Recall that the r-the central moment of a random variable X is $\mathbb{E}[(X - \mathbb{E}[X])^r]$. Show that if the r-th central moment of an almost surely non-negative random variable X is finite, then its l-th central moment is also finite for every $l < r$.*

(d) Because of the formula $\mathrm{var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$, we see that: (i) if X is square integrable, the variance is finite; (ii) if X is integrable, but not square integrable, the variance is infinite; (iii) if X is not integrable, the variance is undefined.

2 EXPECTED VALUES OF SOME COMMON RANDOM VARIABLES

In this section, we use either the definition or the properties of expectations to calculate the mean and variance of a few common discrete random variables.

(a) **Bernoulli**(p)**.** Let X be a Bernoulli random variable with parameter p. Then,

$$\begin{aligned}\mathbb{E}[X] &= 1 \cdot p + 0 \cdot (1-p) = p,\\ \mathrm{var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = 1^2 \cdot p + 0^2 \cdot (1-p) - p^2 = p(1-p).\end{aligned}$$

(b) **Binomial**(n, p)**.** Let X be a binomial random variable with parameters n and p. We note that X can be expressed in the form $X = \sum_{i=1}^n X_i$, where $X_1, \ldots, X_n$ are independent Bernoulli random variables with a common parameter p. It follows that

$$\mathbb{E}[X] = \sum_{i=1}^n \mathbb{E}[X_i] = np.$$

Furthermore, using the independence of the random variables X_i, we have

$$\mathrm{var}(X) = \sum_{i=1}^n \mathrm{var}(X_i) = np(1-p).$$

(c) **Geometric**(p)**.** Let X be a geometric random variable with parameter p. We will use the formula $\mathbb{E}[X] = \sum_{n=0}^\infty \mathbb{P}(X > n)$. We observe that

$$\mathbb{P}(X > n) = \sum_{j=n+1}^\infty (1-p)^{j-1} p = (1-p)^n,$$

which implies that

$$\mathbb{E}[X] = \sum_{n=0}^{\infty} (1-p)^n = \frac{1}{p}.$$

The variance of X is given by

$$\text{var}(X) = \frac{1-p}{p^2},$$

but we defer the derivation to a later section.

(d) **Poisson**(λ)**.** Let X be a Poisson random variable with parameter λ. A direct calculation yields

$$\begin{aligned}
\mathbb{E}[X] &= e^{-\lambda} \sum_{n=0}^{\infty} n \frac{\lambda^n}{n!} \\
&= e^{-\lambda} \sum_{n=1}^{\infty} n \frac{\lambda^n}{n!} \\
&= e^{-\lambda} \sum_{n=1}^{\infty} \frac{\lambda^n}{(n-1)!} \\
&= \lambda e^{-\lambda} \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} \\
&= \lambda.
\end{aligned}$$

The variance of X turns out to satisfy $\text{var}(X) = \lambda$, but we defer the derivation to a later section. We note, however, that the mean and the variance of a Poisson random variable are exactly what one would expect, on the basis of the formulae for the mean and variance of a binomial random variable, and taking the limit as $n \to \infty$, $p \to 0$, while keeping np fixed at λ.

(e) **Power**(α)**.** Let X be a random variable with a power law distribution with parameter α. We have

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} \mathbb{P}(X > k) = \sum_{k=0}^{\infty} \frac{1}{(k+1)^{\alpha}}.$$

If $\alpha \leq 1$, the expected value is seen to be infinite. For $\alpha > 1$, the sum is finite, but a closed form expression is not available; it is known as the Riemann zeta function, and is denoted by $\zeta(\alpha)$.

3 COVARIANCE AND CORRELATION

3.1 Covariance

The **covariance** of two square integrable random variables X and Y is denoted by $\text{cov}(X, Y)$, and is defined by

$$\text{cov}(X,Y) = \mathbb{E}\Big[\ X - \mathbb{E}[X] \quad Y - \mathbb{E}[Y] \quad .$$

When $\text{cov}(X, Y) = 0$, we say that X and Y are **uncorrelated**.

Note that, under the square integrability assumption, the covariance is always well-defined and finite. This is a consequence of the fact that $|XY| \leq (X^2 + Y^2)/2$, which implies that XY, as well as $(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])$, are integrable.

Roughly speaking, a positive or negative covariance indicates that the values of $X - \mathbb{E}[X]$ and $Y - \mathbb{E}[Y]$ obtained in a single experiment "tend" to have the same or the opposite sign, respectively. Thus, the sign of the covariance provides an important qualitative indicator of the relation between X and Y.

We record a few properties of the covariance, which are immediate consequences of its definition:

(a) $\text{cov}(X, X) = \text{var}(X)$;

(b) $\text{cov}(X, Y + a) = \text{cov}(X, Y)$;

(c) $\text{cov}(X, Y) = \text{cov}(Y, X)$;

(d) $\text{cov}(X, aY + bZ) = a \cdot \text{cov}(X, Y) + b \cdot \text{cov}(X, Z)$.

An alternative formula for the covariance is

$$\text{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y],$$

as can be verified by a simple calculation. Recall from last lecture that if X and Y are independent, we have $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$, which implies that $\text{cov}(X, Y) = 0$. Thus, if X and Y are independent, they are also uncorrelated. However, the reverse is not true, as illustrated by the following example.

Example. Suppose that the pair of random variables (X, Y) takes the values $(1, 0)$, $(0, 1)$, $(-1, 0)$, and $(0, -1)$, each with probability 1/4. Thus, the marginal PMFs of X and Y are symmetric around 0, and $\mathbb{E}[X] = \mathbb{E}[Y] = 0$. Furthermore, for all possible value pairs (x, y), either x or y is equal to 0, which implies that $XY = 0$ and $\mathbb{E}[XY] = 0$. Therefore,

$$\text{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = 0,$$

and X and Y are uncorrelated. However, X and Y are not independent since, for example, a nonzero value of X fixes the value of Y to zero.

3.2 Variance of the sum of random variables

The covariance can be used to obtain a formula for the variance of the sum of several (not necessarily independent) random variables. In particular, if X_1, X_2, ..., X_n are random variables with finite variance, we have

$$\text{var}(X_1 + X_2) = \text{var}(X_1) + \text{var}(X_2) + 2\text{cov}(X_1, X_2),$$

and, more generally,

$$\text{var}\ \ \sum_{i=1}^{n} X_i \ \ = \sum_{i=1}^{n} \text{var}(X_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \text{cov}(X_i, X_j).$$

This can be seen from the following calculation, where for brevity, we denote $\tilde{X}_i = X_i - \mathbb{E}[X_i]$:

$$\begin{aligned}
\text{var}\ \ \sum_{i=1}^{n} X_i \ \ &= \mathbb{E}\left[\ \ \sum_{i=1}^{n} \tilde{X}_i \ \ {}^{2}\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} \tilde{X}_i \tilde{X}_j\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{E}[\tilde{X}_i \tilde{X}_j] \\
&= \sum_{i=1}^{n} \mathbb{E}\left[\tilde{X}_i^2\right] + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \mathbb{E}[\tilde{X}_i \tilde{X}_j] \\
&= \sum_{i=1}^{n} \text{var}(X_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \text{cov}(X_i, X_j).
\end{aligned}$$

3.3 Correlation coefficient

The **correlation coefficient** $\rho(X, Y)$ of two random variables X and Y that have nonzero and finite variances is defined as

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}.$$

(The simpler notation ρ will also be used when X and Y are clear from the context.) It may be viewed as a normalized version of the covariance $\text{cov}(X, Y)$.

Theorem 1. *Let X and Y be discrete random variables with positive variance, and correlation coefficient equal to ρ.*

(a) We have $-1 \leq \rho \leq 1$.

(b) We have $\rho = 1$ (respectively, $\rho = -1$) if and only if there exists a positive (respectively, negative) constant a such that $Y - \mathbb{E}[Y] = a(X - \mathbb{E}[X])$, with probability 1.

The proof of Theorem 1 relies on the Schwarz (or Cauchy-Schwarz) inequality, given below.

Proposition 1. (Cauchy-Schwarz inequality) *For any two random variables, X and Y, with finite variance, we have*

$$\mathbb{E}[XY] \;\; ^2 \leq \mathbb{E}[X^2]\,\mathbb{E}[Y^2].$$

Proof: Let us assume that $\mathbb{E}[Y^2] \neq 0$; otherwise, we have $Y = 0$ with probability 1, and hence $\mathbb{E}[XY] = 0$, so the inequality holds. We have

$$\begin{aligned}
0 &\leq \mathbb{E}\left[\left(X - \frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]}Y\right)^2\right] \\
&= \mathbb{E}\left[X^2 - 2\frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]}XY + \frac{\big(\mathbb{E}[XY]\big)^2}{\big(\mathbb{E}[Y^2]\big)^2}Y^2\right] \\
&= \mathbb{E}[X^2] - 2\frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]}\mathbb{E}[XY] + \frac{\big(\mathbb{E}[XY]\big)^2}{\big(\mathbb{E}[Y^2]\big)^2}\mathbb{E}[Y^2] \\
&= \mathbb{E}[X^2] - \frac{\big(\mathbb{E}[XY]\big)^2}{\mathbb{E}[Y^2]},
\end{aligned}$$

i.e., $\mathbb{E}[XY] \;\; ^2 \leq \mathbb{E}[X^2]\,\mathbb{E}[Y^2]$. □

Proof of Theorem 1:

(a) Let $\tilde{X} = X - \mathbb{E}[X]$ and $\tilde{Y} = Y - \mathbb{E}[Y]$. Using the Schwarz inequality, we get

$$\big(\rho(X,Y)\big)^2 = \frac{\big(\mathbb{E}[\tilde{X}\tilde{Y}]\big)^2}{\mathbb{E}[\tilde{X}^2]\,\mathbb{E}[\tilde{Y}^2]} \leq 1,$$

and hence $|\rho(X,Y)| \leq 1$.

(b) One direction is straightforward. If $\tilde{Y} = a\tilde{X}$, then

$$\rho(X,Y) = \frac{\mathbb{E}[\tilde{X}a\tilde{X}]}{\sqrt{\mathbb{E}[\tilde{X}^2]\,\mathbb{E}[(a\tilde{X})^2]}} = \frac{a}{|a|},$$

which equals 1 or -1 depending on whether a is positive or negative.

To establish the reverse direction, let us assume that $(\rho(X,Y))^2 = 1$, which implies that $\mathbb{E}[\tilde{X}^2]\mathbb{E}[\tilde{Y}^2] = (\mathbb{E}[\tilde{X}\tilde{Y}])^2$. Using the inequality established in the proof of Proposition 1, we conclude that the random variable

$$\tilde{X} - \frac{\mathbb{E}[\tilde{X}\tilde{Y}]}{\mathbb{E}[\tilde{Y}^2]}\tilde{Y}$$

is equal to zero, with probability 1. It follows that, with probability 1,

$$\tilde{X} = \frac{\mathbb{E}[\tilde{X}\tilde{Y}]}{\mathbb{E}[\tilde{Y}^2]}\tilde{Y} = \sqrt{\frac{\mathbb{E}[\tilde{X}^2]}{\mathbb{E}[\tilde{Y}^2]}}\rho(X,Y)\tilde{Y}.$$

Note that the sign of the constant ratio of $\tilde{X}$ and $\tilde{Y}$ is determined by the sign of $\rho(X,Y)$, as claimed. □

Example. Consider n independent tosses of a coin with probability of a head equal to p. Let X and Y be the numbers of heads and of tails, respectively, and let us look at the correlation coefficient of X and Y. Here, we have $X + Y = n$, and also $\mathbb{E}[X] + \mathbb{E}[Y] = n$. Thus,

$$X - \mathbb{E}[X] = - \;\; Y - \mathbb{E}[Y] \;\; .$$

We will calculate the correlation coefficient of X and Y, and verify that it is indeed equal to -1.

We have

$$\begin{aligned}\mathrm{cov}(X,Y) &= \mathbb{E}\Big[\big(X - \mathbb{E}[X]\big)\big(Y - \mathbb{E}[Y]\big)\Big] \\ &= -\mathbb{E}\left[\big(X - \mathbb{E}[X]\big)^2\right] \\ &= -\mathrm{var}(X).\end{aligned}$$

Hence, the correlation coefficient is

$$\rho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}} = \frac{-\mathrm{var}(X)}{\sqrt{\mathrm{var}(X)\mathrm{var}(X)}} = -1.$$

4 INDICATOR VARIABLES AND THE INCLUSION-EXCLUSION FORMULA

Indicator functions are special discrete random variables that can be useful in simplifying certain derivations or proofs. In this section, we develop the inclusion-exclusion formula and apply it to a matching problem.

Recall that with every event A, we can associate its **indicator function**, which is a discrete random variable $I_A : \Omega \to \{0, 1\}$, defined by $I_A(\omega) = 1$ if $\omega \in A$, and $I_A(\omega) = 0$ otherwise. Note that $I_{A^c} = 1 - I_A$ and that $\mathbb{E}[I_A] = \mathbb{P}(A)$. These simple observations, together with the linearity of expectations turn out to be quite useful.

4.1 The inclusion-exclusion formula

Note that $I_{A \cap B} = I_A I_B$, for every $A, B \in \mathcal{F}$. Therefore,

$$\begin{aligned} I_{A \cup B} &= 1 - I_{(A \cup B)^c} = 1 - I_{A^c \cap B^c} = 1 - I_{A^c} I_{B^c} \\ &= 1 - (1 - I_A)(1 - I_B) = I_A + I_B - I_A I_B. \end{aligned}$$

Taking expectations of both sides, we obtain

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B),$$

an already familiar formula.

We now derive a generalization, known as the inclusion-exclusion formula. Suppose we have a collection of events A_j, $j = 1, \ldots, n$, and that we are interested in the probability of the event $B = \cup_{j=1}^n A_j$. Note that

$$I_B = 1 - \prod_{j=1}^n (1 - I_{A_j}).$$

We begin with the easily verifiable fact that for any real numbers $a_1, \ldots, a_n$, we have

$$\begin{aligned} \prod_{j=1}^n (1 - a_j) =& 1 - \sum_{1 \le j \le n} a_j + \sum_{1 \le i < j \le n} a_i a_j - \sum_{1 \le i < j < k \le n} a_i a_j a_k \\ & + \cdots + (-1)^n a_1 \cdots a_n. \end{aligned}$$

We replace a_j by I_{A_j}, and then take expectations of both sides, to obtain

$$\mathbb{P}(B) = \sum_{1\le j\le n} \mathbb{P}(A_j) - \sum_{1\le i<j\le n} \mathbb{P}(A_i \cap A_j) + \sum_{1\le i<j<k\le n} \mathbb{P}(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n+1}\mathbb{P}(A_1 \cap \cdots \cap A_n).$$

4.2 The matching problem

Suppose that n people throw their hats in a box, where $n \geq 2$, and then each person picks one hat at random. (Each hat will be picked by exactly one person.) We interpret "at random" to mean that every permutation of the n hats is equally likely, and therefore has probability $1/n!$.

In an alternative model, we can visualize the experiment sequentially: the first person picks one of the n hats, with all hats being equally likely; then, the second person picks one of the remaining $n-1$ remaining hats, with every remaining hat being equally likely, etc. It can be verified that under this second model, every permutation has probability $1/n!$, so the two models are equivalent.

We are interested in the mean, variance, and PMF of a random variable X, defined as the number of people that get back their own hat.[1] This problem is best approached using indicator variables.

For the ith person, we introduce a random variable X_i that takes the value 1 if the person selects his/her own hat, and takes the value 0 otherwise. Note that

$$X = X_1 + X_2 + \cdots + X_n.$$

Since $\mathbb{P}(X_i = 1) = 1/n$ and $\mathbb{P}(X_i = 0) = 1 - 1/n$, the mean of X_i is

$$\mathbb{E}[X_i] = 1 \cdot \frac{1}{n} + 0 \cdot \left(1 - \frac{1}{n}\right) = \frac{1}{n},$$

which implies that

$$\mathbb{E}[X] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n] = n \cdot \frac{1}{n} = 1.$$

In order to find the variance of X, we first find the variance and covariances of the random variables X_i. We have

$$\text{var}(X_i) = \frac{1}{n}\left(1 - \frac{1}{n}\right).$$

[1] For more results on various extensions of the matching problem, see L.A. Zager and G.C. Verghese, "Caps and robbers: what can you expect?," *College Mathematics Journal*, v. 38, n. 3, 2007, pp. 185-191.

For $i \neq j$, we have

$$\begin{aligned}
\mathrm{cov}(X_i, X_j) &= \mathbb{E}\Big[\; X_i - \mathbb{E}[X_i] \quad X_j - \mathbb{E}[X_j] \\
&= \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\,\mathbb{E}[X_j] \\
&= \mathbb{P}(X_i = 1 \text{ and } X_j = 1) - \mathbb{P}(X_i = 1)\mathbb{P}(X_j = 1) \\
&= \mathbb{P}(X_i = 1)\mathbb{P}(X_j = 1 \mid X_i = 1) - \mathbb{P}(X_i = 1)\mathbb{P}(X_j = 1) \\
&= \frac{1}{n} \cdot \frac{1}{n-1} - \frac{1}{n^2} \\
&= \frac{1}{n^2(n-1)}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathrm{var}(X) &= \mathrm{var}\left(\sum_{i=1}^{n} X_i\right) \\
&= \sum_{i=1}^{n} \mathrm{var}(X_i) + 2\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathrm{cov}(X_i, X_j) \\
&= n \cdot \frac{1}{n}\left(1 - \frac{1}{n}\right) + 2 \cdot \frac{n(n-1)}{2} \cdot \frac{1}{n^2(n-1)} \\
&= 1.
\end{aligned}$$

Finding the PMF of X is a little harder. Let us first dispense with some easy cases. We have $\mathbb{P}(X = n) = 1/n!$, because there is only one (out of the $n!$ possible) permutations under which every person receives their own hat. Furthermore, the event $X = n - 1$ is impossible: if $n - 1$ persons have received their own hat, the remaining person must also have received their own hat.

Let us continue by finding the probability that $X = 0$. Let A_i be the event that the ith person gets their own hat, i.e., $X_i = 1$. Note that the event $X = 0$ is the same as the event $\cap_i A_i^c$. Thus, $\mathbb{P}(X = 0) = 1 - \mathbb{P}(\cup_{i=1}^{n} A_i)$. Using the inclusion-exclusion formula, we have

$$\mathbb{P}(\cup_{i=1}^{n} A_i) = \sum_i \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbb{P}(A_i \cap A_j \cap A_k) + \cdots .$$

Observe that for every fixed distinct indices $i_1, i_2, \ldots, i_k$, we have

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = \frac{1}{n} \cdot \frac{1}{n-1} \cdots \frac{1}{n-k+1} = \frac{(n-k)!}{n!}. \qquad (1)$$

Thus,

$$\begin{aligned}\mathbb{P}(\cup_{i=1}^{n} A_i) &= n \cdot \frac{1}{n} - \genfrac{}{}{0pt}{}{n}{2} \frac{(n-2)!}{n!} + \genfrac{}{}{0pt}{}{n}{3} \frac{(n-3)!}{n!} + \cdots + (-1)^{n+1} \genfrac{}{}{0pt}{}{n}{n} \frac{(n-n)!}{n!} \\ &= 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + (-1)^{n+1} \frac{1}{n!}.\end{aligned}$$

We conclude that

$$\mathbb{P}(X = 0) = \frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^n \frac{1}{n!}. \qquad (2)$$

Note that $\mathbb{P}(X = 0) \to e^{-1}$, as $n \to \infty$.

To conclude, let us now fix some integer r, with $0 < r \leq n-2$, and calculate $\mathbb{P}(X = r)$. The event $\{X = r\}$ can only occur as follows: for some subset S of $\{1, \ldots, n\}$, of cardinality r, the following two events, B_S and C_S, occur:

B_S : for every $i \in S$, person i receives their own hat;

C_S : for every $i \notin S$, person i does not receive their own hat.

We then have

$$\{X = r\} = \bigcup_{S:\,|S|=r} B_S \cap C_S.$$

The events $B_S \cap C_S$ for different subsets S are disjoint. Furthermore, by symmetry, $\mathbb{P}(B_S \cap C_S)$ is the same for every S of cardinality r. Thus,

$$\begin{aligned}\mathbb{P}(X = r) &= \sum_{S:\,|S|=r} \mathbb{P}(B_S \cap C_S) \\ &= \binom{n}{r} \mathbb{P}(B_S)\, \mathbb{P}(C_S \mid B_S).\end{aligned}$$

Note that

$$\mathbb{P}(B_S) = \frac{(n-r)!}{n!},$$

by the same argument as in Eq. (1). Conditioned on the event that the r persons in the set S have received their own hats, the event C_S will materialize if and only if none of the remaining $n - r$ persons receive their own hat. But this is the same situation as the one analyzed when we calculated the probability that $X = 0$, except that n needs to be replaced by $n - r$. We conclude that

$$P(C_S \mid B_S) = \frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^{n-r} \frac{1}{(n-r)!}.$$

Putting everything together, we conclude that

$$\begin{aligned}\mathbb{P}(X = r) &= \binom{n}{r}\frac{(n-r)!}{n!}\Big(\frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^{n-r}\frac{1}{(n-r)!}\Big)\\ &= \frac{1}{r!}\Big(\frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^{n-r}\frac{1}{(n-r)!}\Big).\end{aligned}$$

Note that for each fixed r, the probability $\mathbb{P}(X = r)$ converges to $e^{-1}/r!$, as $n \to \infty$, which corresponds to a Poisson distribution with parameter 1. An intuitive justification is as follows. The random variables X_i are not independent (in particular, their covariance is nonzero). On the other hand, as $n \to \infty$, they are "approximately independent". Furthermore, the success probability for each person is $1/n$, and the situation is similar to the one in our earlier proof that the binomial PMF approaches the Poisson PMF.

5 CONDITIONAL EXPECTATIONS

We have already defined the notion of a conditional PMF, $p_{X\,|\,Y}(\,\cdot\,|\,y)$, given the value of a random variable Y. Similarly, given an event A, we can define a conditional PMF $p_{X|A}$, by letting $p_{X|A}(x) = \mathbb{P}(X = x\,|\,A)$. In either case, the conditional PMF, as a function of x, is a bona fide PMF (a nonnegative function that sums to one). As such, it is natural to associate a (conditional) expectation to the (conditional) PMF.

Definition 1. *Given an event A, such that $\mathbb{P}(A) > 0$, and a discrete random variable X, the **conditional expectation** of X given A is defined as*

$$\mathbb{E}[X\,|\,A] = \sum_x x p_{X\,|\,A}(x),$$

provided that the sum is well-defined.

Note that the preceding also provides a definition for a conditional expectation of the form $\mathbb{E}[X\,|\,Y = y]$, for any y such that $p_Y(y) > 0$: just let A be the event $\{Y = y\}$, which yields

$$\mathbb{E}[X\,|\,Y = y] = \sum_x x p_{X\,|\,Y}(x\,|\,y).$$

We note that the conditional expectation is always well defined when either the random variable X is nonnegative, or when the random variable X is integrable. In particular, whenever $\mathbb{E}[|X|] < \infty$, we also have $\mathbb{E}[|X|\,|\,Y = y] < \infty$, for every y such that $p_Y(y) > 0$. To verify the latter assertion, note that for every y such that $p_Y(y) > 0$, we have

$$\sum_x |x| p_{X|Y}(x \mid y) = \sum_x |x| \frac{p_{X,Y}(x,y)}{p_Y(y)} \leq \frac{1}{p_Y(y)} \sum_x |x| p_X(x) = \frac{\mathbb{E}[|X|]}{p_Y(y)}.$$

The converse, however, is not true: it is possible that $\mathbb{E}[|X| \mid Y = y]$ is finite for every y that has positive probability, while $\mathbb{E}[|X|] = \infty$. This is left as an exercise.

The conditional expectation is essentially the same as an ordinary expectation, except that the original PMF is replaced by the conditional PMF. As such, the conditional expectation inherits all the properties of ordinary expectations (cf. Proposition 4 in the notes for Lecture 6).

5.1 The total expectation theorem

A simple calculation yields

$$\begin{aligned} \sum_y \mathbb{E}[X \mid Y = y] p_Y(y) &= \sum_y \sum_x x p_{X|Y}(x \mid y) p_Y(y) \\ &= \sum_y \sum_x x p_{X,Y}(x,y) \\ &= \mathbb{E}[X]. \end{aligned}$$

Note that this calculation is rigorous if X is nonnegative or integrable.

Suppose now that $\{A_i\}$ is a countable family of disjoint events that forms a partition of the probability space Ω. Define a random variable Y by letting $Y = i$ if and only if A_i occurs. Then, $p_Y(i) = \mathbb{P}(A_i)$, and $\mathbb{E}[X \mid Y = i] = \mathbb{E}[X \mid A_i]$, which yields

$$\mathbb{E}[X] = \sum_i \mathbb{E}[X \mid A_i] \mathbb{P}(A_i).$$

Example. (The mean of the geometric.) Let X be a random variable with parameter p, so that $p_X(k) = (1-p)^{k-1}p$, for $p \in \mathbb{N}$. We first observe that the geometric distribution is memoryless: for $k \in \mathbb{N}$, we have

$$\begin{aligned} \mathbb{P}(X - 1 = k \mid X > 1) &= \frac{\mathbb{P}(X = k+1, X > 1)}{\mathbb{P}(X > 1)} \\ &= \frac{\mathbb{P}(X = k+1)}{\mathbb{P}(X > 1)} \\ &= \frac{(1-p)^k p}{1-p} = (1-p)^{k-1} p \\ &= \mathbb{P}(X = k). \end{aligned}$$

In words, in a sequence of repeated i.i.d., trials, given that the first trial was a failure, the distribution of the remaining trials, $X - 1$, until the first success is the same as the unconditional distribution of the number of trials, X, until the first success. In particular, $\mathbb{E}[X - 1 \mid X > 1] = \mathbb{E}[X]$.

Using the total expectation theorem, we can write

$$\mathbb{E}[X] = \mathbb{E}[X \,|\, X > 1]\mathbb{P}(X > 1) + \mathbb{E}[X \,|\, X = 1]\mathbb{P}(X = 1) = (1 + \mathbb{E}[X])(1 - p) + 1 \cdot p.$$

We solve for $\mathbb{E}[X]$, and find that $\mathbb{E}[X] = 1/p$.

Similarly,

$$\mathbb{E}[X^2] = \mathbb{E}[X^2 \,|\, X > 1]\mathbb{P}(X > 1) + \mathbb{E}[X^2 \,|\, X = 1]\mathbb{P}(X = 1).$$

Note that

$$\mathbb{E}[X^2 \,|\, X > 1] = \mathbb{E}[(X-1)^2 \,|\, X > 1] + \mathbb{E}[2(X-1)+1 \,|\, X > 1] = \mathbb{E}[X^2] + (2/p) + 1.$$

Thus,

$$\mathbb{E}[X^2] = (1 - p)(\mathbb{E}[X^2] + (2/p) + 1) + p,$$

which yields

$$\mathbb{E}[X^2] = \frac{2}{p^2} - \frac{1}{p}.$$

We conclude that

$$\mathrm{var}(X) = \mathbb{E}[X^2] - \ \mathbb{E}[X] \ ^2 = \frac{2}{p^2} - \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2}.$$

Example. Suppose we flip a biased coin N times, independently, where N is a Poisson random variable with parameter λ. The probability of heads at each flip is p. Let X be the number of heads, and let Y be the number of tails. Then,

$$\mathbb{E}[X \,|\, N = n] = \sum_{m=0}^{\infty} m\mathbb{P}(X = m \,|\, N = n) = \sum_{m=0}^{n} m\binom{n}{m}p^m(1-p)^{n-m}.$$

But X is just the expected number of heads in n trials, so that $\mathbb{E}[X \,|\, N = n] = np$.

Let us now calculate $\mathbb{E}[N \,|\, X = m]$. We have

$$\begin{aligned}
\mathbb{E}[N \,|\, X = m] &= \sum_{n=0}^{\infty} n\mathbb{P}(N = n \,|\, X = m) \\
&= \sum_{n=m}^{\infty} n\frac{\mathbb{P}(N = n, X = m)}{\mathbb{P}(X = m)} \\
&= \sum_{n=m}^{\infty} n\frac{\mathbb{P}(X = m \mid N = n)\mathbb{P}(N = n)}{\mathbb{P}(X = m)} \\
&= \sum_{n=m}^{\infty} n\frac{\genfrac{}{}{0pt}{}{n}{m}\, p^m(1-p)^{n-m}(\lambda^n/n!)e^{-\lambda}}{\mathbb{P}(X = m)}.
\end{aligned}$$

Recall that $X \overset{d}{=} \text{Pois}(\lambda p)$, so that $\mathbb{P}(X = m) = e^{-\lambda p}(\lambda p)^m/m!$. Thus, after some cancellations, we obtain

$$\begin{aligned}
\mathbb{E}[N \mid X = m] &= \sum_{n=m}^{\infty} n \frac{(1-p)^{n-m}\lambda^{n-m}e^{-\lambda(1-p)}}{(n-m)!} \\
&= \sum_{n=m}^{\infty} (n-m) \frac{(1-p)^{n-m}\lambda^{n-m}e^{-\lambda(1-p)}}{(n-m)!} \\
&+ m \sum_{n=m}^{\infty} \frac{(1-p)^{n-m}\lambda^{n-m}e^{-\lambda(1-p)}}{(n-m)!} \\
&= \lambda(1-p) + m.
\end{aligned}$$

A faster way of obtaining this result is as follows. From Theorem 3 in the notes for Lecture 6, we have that X and Y are independent, and that Y is Poisson with parameter $\lambda(1-p)$. Therefore,

$$\mathbb{E}[N \mid X = m] = \mathbb{E}[X \mid X = m] + \mathbb{E}[Y \mid X = m] = m + \mathbb{E}[Y] = m + \lambda(1-p).$$

Exercise. (Simpson's "paradox") Let S be an event and X, Y discrete random variables, all defined on a common probability space. Show that

$$\mathbb{P}[S|X = 0, Y = y] > \mathbb{P}[S|X = 1, Y = y] \qquad \forall y$$

does <u>not</u> imply

$$\mathbb{P}[S|X = 0] \geq \mathbb{P}[S|X = 1] .$$

Thus in a clinical trial comparing two treatments (indexed by X) a drug can be more successful on each group of patients (indexed by Y) yet be less successful overall.

5.2 The conditional expectation as a random variable

Let X and Y be two discrete random variables. For any fixed value of y, the expression $\mathbb{E}[X \mid Y = y]$ is a real number, which however depends on y, and can be used to define a function $\phi : \mathbb{R} \to \mathbb{R}$, by letting $\phi(y) = \mathbb{E}[X \mid Y = y]$. Consider now the random variable $\phi(Y)$; this random variable takes the value $\mathbb{E}[X \mid Y = y]$ whenever Y takes the value y, which happens with probability $\mathbb{P}(Y = y)$. This random variable will be denoted as $\mathbb{E}[X \mid Y]$. (Strictly speaking, one needs to verify that this is a measurable function, which is left as an exercise.)

Example. Let us return to the last example and find $\mathbb{E}[X \mid N]$ and $\mathbb{E}[N \mid X]$. We found that $\mathbb{E}[X \mid N = n] = np$. Thus $\mathbb{E}[X \mid N] = Np$, i.e., it is a random variable that takes the value np with probability $\mathbb{P}(N = n) = (\lambda^n/n!)e^{-\lambda}$. We found that $\mathbb{E}[N \mid X = m] = \lambda(1-p) + m$. Thus $\mathbb{E}[N \mid X] = \lambda(1-p) + X$.

Note further that

$$\mathbb{E}[\mathbb{E}[X \mid N]] = \mathbb{E}[Np] = \lambda p = \mathbb{E}[X],$$

and

$$\mathbb{E}[\mathbb{E}[N \mid X]] = \lambda(1-p) + \mathbb{E}[X] = \lambda(1-p) + \lambda p = \lambda = \mathbb{E}[N].$$

This is not a coincidence; the equality $\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X]$ is always true, as we shall now see. In fact, this is just the total expectation theorem, written in more abstract notation.

Theorem 2. *Let $g : \mathbb{R} \to \mathbb{R}$ be a measurable function such that $Xg(Y)$ is either nonnegative or integrable. Then,*

$$\mathbb{E}\ \mathbb{E}[X \mid Y]g(Y)\ \ = \mathbb{E}[Xg(Y)].$$

In particular, by letting $g(y) = 1$ for all y, we obtain $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$.

Proof: We have

$$\begin{aligned} \mathbb{E}\ \mathbb{E}[X|Y]g(Y)\ &= \sum_y \mathbb{E}[X \mid Y = y]g(y)p_Y(y) \\ &= \sum_y \sum_x x p_{X|Y}(x \mid y)g(y)p_Y(y) \\ &= \sum_{x,y} xg(y)p_{X,Y}(x,y) = \mathbb{E}[Xg(Y)]. \end{aligned}$$

□

The formula in Theorem 2 can be rewritten in the form

$$\mathbb{E}\big[(\mathbb{E}[X \mid Y] - X)g(Y)\big] = 0. \tag{3}$$

Here is an interpretation. We can think of $\mathbb{E}[X \mid Y]$ as an estimate of X, on the basis of Y, and $\mathbb{E}[X \mid Y] - X$ as an estimation error. The above formula says that the estimation error is uncorrelated with every function of the original data.

Equation (3) can be used as the basis for an abstract definition of conditional expectations. Namely, we define the conditional expectation as a random variable of the form $\phi(Y)$, where ϕ is a measurable function, that has the property

$$\mathbb{E}\big[(\phi(Y) - X)g(Y)\big] = 0,$$

for every measurable function g. The merits of this definition is that it can be used for all kinds of random variables (discrete, continuous, mixed, etc.). However, for this definition to be sound, there are two facts that need to be verified:

(a) Existence: It turns out that as long as X is integrable, a function ϕ with the above properties is guaranteed to exist. We already know that this is the case for discrete random variables: the conditional expectation as defined in the beginning of this section does have the desired properties. For general random variables, this is a nontrivial and deep result. It will be revisited later in this course.

(b) Uniqueness: It turns out that there is essentially only one function ϕ with the above properties. More precisely, any two functions with the above properties are equal with probability 1.