Bruce K. Driver

180B Lecture Notes, W2011

January 11, 2011 *File:180Lec.tex*

Contents

Part

180B Notes

Course Notation List

1. (Ω, P) will denote a probability spaces and S will denote a set which is called **state space**.
2. If S is a discrete set, i.e. finite or countable and $X : \Omega \to S$ we let

$$\rho_X(s) := P(X = s).$$

   More generally if $X_i : \Omega \to S_i$ for $1 \leq i \leq n$ we let

$$\rho_{X_1,\ldots,X_n}(\mathbf{s}) := P(X_1 = s_1, \ldots, X_n = s_n)$$

   for all $\mathbf{s} = (s_1, \ldots, s_n) \in S_1 \times \cdots \times S_n$.
3. If S is $\mathbb{R}$ or $\mathbb{R}^n$ and $X : \Omega \to S$ is a continuous random variable, we let $\rho_X(x)$ be the operability density function of X, namely,

$$\mathbb{E}[f(X)] = \int_S f(x)\,\rho_X(x)\,dx.$$

4. Given random variables X and Y we let;
   a) $\mu_X := \mathbb{E}X$ be the mean of X.
   b) $\mathrm{Var}(X) := \mathbb{E}\left[(X - \mu_X)^2\right] = \mathbb{E}X^2 - \mu_X^2$ be the variance of X.
   c) $\sigma_X = \sigma(X) := \sqrt{\mathrm{Var}(X)}$ be the standard deviation of X.
   d) $\mathrm{Cov}(X, Y) := \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY] - \mu_X\mu_Y$ be the covariance of X and Y.
   e) $Corr(X, Y) := \mathrm{Cov}(X, Y) / (\sigma_X\sigma_Y)$ be the **correlation** of X and Y.

1

Course Overview and Plan

This course is an introduction to some basic topics in the theory of stochastic processes. After finishing the discussion of multivariate distributions and conditional probabilities initiated in Math 180A, we will study Markov chains in discrete time. We then begin our investigation of stochastic processes in continuous time with a detailed discussion of the Poisson process. These two topics will be combined in Math 180C when we study Markov chains in continuous time and renewal processes.

In the next two quarters we will study some aspects of Stochastic Processes. Stochastic (from the Greek $\sigma\tau\acute{o}\chi o\xi$ for aim or guess) means random. A stochastic process is one whose behavior is non-deterministic, in that a system's subsequent state is determined both by the process's predictable actions and by a random element. However, according to M. Kac[1] and E. Nelson[2], any kind of time development (be it deterministic or essentially probabilistic) which is analyzable in terms of probability deserves the name of stochastic process.

Mathematically we will be interested in collection of random variables or vectors $\{X_t\}_{t\in T}$ with $X_t : \Omega \to S$ (S is the **state space**) on some probability space, (Ω, P). Here T is typically in $\mathbb{R}_+$ or $\mathbb{Z}_+$ but not always.

Example 1.1. 1. X_t is the value of a spinner at times $t \in \mathbb{Z}_+$.
2. X_t denotes the prices of a stock (or stocks) on the stock market.
3. X_t denotes the value of your portfolio at time t.
4. X_t is the position of a dust particle like in Brownian motion.
5. X_A is the number of stars in a region A contained in space or the number of raisins in a region of a cake, etc.
6. $X_n \in S = Perm(\{1,\dots,52\})$ is the ordering of cards in a deck of cards after the n^{th} shuffle.

Our goal in this course is to introduce and analyze models for such random objects. This is clearly going to require that we make assumptions on $\{X_t\}$ which will typically be some sort of dependency structures. This is where we will begin our study – namely heading towards conditional expectations and related topics.

[1] M. Kac & J. Logan, in Fluctuation Phenomena, eds. E.W. Montroll & J.L. Lebowitz, North-Holland, Amsterdam, 1976.
[2] E. Nelson, Quantum Fluctuations, Princeton University Press, Princeton, 1985.

1.1 180B Course Topics:

1. Review the linear algebra of orthogonal projections in the context of least squares approximations in the context of Probability Theory.
2. Use the least squares theory to interpret covariance and correlations.
3. Review of conditional probabilities for discrete random variables.
4. Introduce conditional expectations as least square approximations.
5. Develop conditional expectation relative to discrete random variables.
6. Give a short introduction to martingale theory.
7. Study in some detail discrete time Markov chains.
8. Review of conditional probability densities for continuous random variables.
9. Develop conditional expectations relative to continuous random variables.
10. Begin our study of the Poisson process.

The bulk of this quarter will involve the study of Markov chains and processes. These are processes for which the past and future are independent given the present. This is a typical example of a dependency structure that we will consider in this course. For an example of such a process, let $S = \mathbb{Z}$ and place a coin at each site of S (perhaps the coins are biased with different probabilities of heads at each site of S.) Let $X_0 = s_0$ be some point in S be fixed and then flip the coin at s_0 and move to the right on step if the result is heads and to left one step if the result is tails. Repeat this process to determine the position X_{n+1} from the position X_n along with a flip of the coin at X_n. This is a typical example of a Markov process.

Before going into these and other processes in more detail we are going to develop the extremely important concept of **conditional expectation.** The idea is as follows. Suppose that X and Y are two random variables with $\mathbb{E}|Y|^2 < \infty$. We wish to find the function h such that $h(X)$ is the minimizer of $\mathbb{E}(Y - f(X))^2$ over all functions f such that $\mathbb{E}\left[f(X)^2\right] < \infty$, that is $h(X)$ is a least squares approximation to Y among random variables of the form $f(X)$, i.e.

$$\mathbb{E}(Y - h(X))^2 = \min_f \mathbb{E}(Y - f(X))^2. \tag{1.1}$$

Fact: a minimizing function h always exist and is "essentially unique." We denote $h(X)$ as $\mathbb{E}[Y|X]$ and call it the **conditional expectation of Y given**

X. We are going to spend a fair amount of time filling in the details of this construction and becoming familiar with this concept.

As a warm up to conditional expectation, we are going to consider a simpler problem of best linear approximations. The goal now is to find $a_0, b_0 \in \mathbb{R}$ such that

$$\mathbb{E}\left(Y - a_0 X + b_0\right)^2 = \min_{a,b\in\mathbb{R}} \mathbb{E}\left(Y - aX + b\right)^2. \tag{1.2}$$

This is the same sort of problem as finding conditional expectations except we now only allow consider functions of the form $f(x) = ax + b$. (You should be able to find a_0 and b_0 using the first derivative test from calculus! We will carry this out using linear algebra ideas below.) It turns out the answer to finding (a_0, b_0) solving Eq. (1.2) only requires knowing the first and second moments of X and Y and $\mathbb{E}[XY]$. On the other hand finding $h(X)$ solving Eq. (1.1) require full knowledge of the joint distribution of (X, Y).

By the way, you are asked to show on your first homework that $\min_{c\in\mathbb{R}} \mathbb{E}\left(Y - c\right)^2 = \mathrm{Var}(Y)$ which occurs for $c = \mathbb{E}Y$. Thus $\mathbb{E}Y$ is the least squares approximation to Y by a constant function and $\mathrm{Var}(Y)$ is the least square error associated with this problem.

Covariance and Correlation

Suppose that (Ω, P) is a probability space. We say that $X : \Omega \to \mathbb{R}$ is **integrable** if $\mathbb{E}|X| < \infty$ and X is **square integrable** if $\mathbb{E}|X|^2 < \infty$. We denote the set of integrable random variables by $L^1(P)$ and the square integrable random variables by $L^2(P)$. When X is integrable we let $\mu_X := \mathbb{E}X$ be the **mean** of X. If Ω is a finite set, then

$$\mathbb{E}[|X|^p] = \sum_{\omega \in \Omega} |X(\omega)|^p P(\{\omega\}) < \infty$$

for any $0 < p < \infty$. So when the sample space is finite requiring integrability or square integrability is no restriction at all. On the other hand when Ω is infinite life can become a little more complicated.

Example 2.1. Suppose that N is a geometric with parameter p so that $P(N = k) = p(1-p)^{k-1}$ for $k \in \mathbb{N} = \{1, 2, 3, \dots\}$. If $X = f(N)$ for some function $f : \mathbb{N} \to \mathbb{R}$, then

$$\mathbb{E}[f(N)] = \sum_{k=1}^{\infty} p(1-p)^{k-1} f(k)$$

when the sum makes sense. So if $X_\lambda = \lambda^N$ for some $\lambda > 0$ we have

$$\mathbb{E}\left[X_\lambda^2\right] = \sum_{k=1}^{\infty} p(1-p)^{k-1} \lambda^{2k} = p\lambda^2 \sum_{k=1}^{\infty} \left[(1-p)\lambda^2\right]^{k-1} < \infty$$

iff $(1-p)\lambda^2 < 1$, i.e. $\lambda < 1/\sqrt{1-p}$. Thus we see that $X_\lambda \in L^2(P)$ iff $\lambda < 1/\sqrt{1-p}$.

Lemma 2.2. *$L^2(P)$ is a subspace of the vector space of random variables on (Ω, P). Moreover if $X, Y \in L^2(P)$, then $XY \in L^1(P)$ and in particular (take $Y = 1$) it follows that $L^2(P) \subset L^1(P)$.*

Proof. If $X, Y \in L^2(P)$ and $c \in \mathbb{R}$ then $\mathbb{E}|cX|^2 = c^2\mathbb{E}|X|^2 < \infty$ so that $cX \in L^2(P)$. Since

$$0 \le (|X| - |Y|)^2 = |X|^2 + |Y|^2 - 2|X||Y|,$$

it follows that

$$|XY| \le \frac{1}{2}|X|^2 + \frac{1}{2}|Y|^2 \in L^1(P).$$

Moreover,

$$(X+Y)^2 = X^2 + Y^2 + 2XY \le X^2 + Y^2 + 2|XY| \le 2\left(X^2 + Y^2\right)$$

from which it follows that $\mathbb{E}(X+Y)^2 < \infty$, i.e. $X + Y \in L^2(P)$. ∎

Definition 2.3. *The **covariance,** $\mathrm{Cov}(X, Y)$, of two square integrable random variables, X and Y, is defined by*

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY] - \mathbb{E}X \cdot \mathbb{E}Y$$

*where $\mu_X := \mathbb{E}X$ and $\mu_Y := \mathbb{E}Y$. The **variance** of X,*

$$\mathrm{Var}(X) = \mathrm{Cov}(X, X) = \mathbb{E}\left[X^2\right] - (\mathbb{E}X)^2 \tag{2.1}$$

$$= \mathbb{E}\left[(X - \mu_X)^2\right] \tag{2.2}$$

*We say that X and Y are **uncorrelated** if $\mathrm{Cov}(X, Y) = 0$, i.e. $\mathbb{E}[XY] = \mathbb{E}X \cdot \mathbb{E}Y$. More generally we say $\{X_k\}_{k=1}^n \subset L^2(P)$ are **uncorrelated** iff $\mathrm{Cov}(X_i, X_j) = 0$ for all $i \neq j$.*

Definition 2.4 (Correlation). *Given two non-constant random variables we define $Corr(X, Y) := \frac{\mathrm{Cov}(X,Y)}{\sigma(X)\cdot\sigma(Y)}$ to be the **correlation** of X and Y.*

It follows from Eqs. (2.1) and (2.2) that

$$0 \le \mathrm{Var}(X) \le \mathbb{E}\left[X^2\right] \text{ for all } X \in L^2(P). \tag{2.3}$$

Exercise 2.1. Let X, Y be two random variables on $(\Omega, \mathcal{B}, P)$;

1. Show that X and Y are independent iff $\mathrm{Cov}(f(X), g(Y)) = 0$ (i.e. $f(X)$ and $g(Y)$ are **uncorrelated**) for bounded measurable functions, $f, g : \mathbb{R} \to \mathbb{R}$. (In this setting X and Y may take values in some arbitrary state space, S.)
2. If $X, Y \in L^2(P)$ and X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$. Note well: we will see in examples below that $\mathrm{Cov}(X, Y) = 0$ does **not** necessarily imply that X and Y are independent.

Solution to Exercise (2.1). (Only roughly sketched the proof of this in class.)

1. Since

$$\operatorname{Cov}(f(X), g(Y)) = \mathbb{E}[f(X) g(Y)] - \mathbb{E}[f(X)] \mathbb{E}[g(Y)]$$

it follows that $\operatorname{Cov}(f(X), g(Y)) = 0$ iff

$$\mathbb{E}[f(X) g(Y)] = \mathbb{E}[f(X)] \mathbb{E}[g(Y)]$$

from which item 1. easily follows.

2. Let $f_M(x) = x 1_{|x| \le M}$, then by independence,

$$\mathbb{E}[f_M(X) g_M(Y)] = \mathbb{E}[f_M(X)] \mathbb{E}[g_M(Y)]. \tag{2.4}$$

Since

$$|f_M(X) g_M(Y)| \le |XY| \le \frac{1}{2}\left(X^2 + Y^2\right) \in L^1(P),$$
$$|f_M(X)| \le |X| \le \frac{1}{2}\left(1 + X^2\right) \in L^1(P), \text{ and}$$
$$|g_M(Y)| \le |Y| \le \frac{1}{2}\left(1 + Y^2\right) \in L^1(P),$$

we may use the DCT three times to pass to the limit as $M \to \infty$ in Eq. (2.4) to learn that $\mathbb{E}[XY] = \mathbb{E}[X] \mathbb{E}[Y]$, i.e. $\operatorname{Cov}(X, Y) = 0$. (These technical details were omitted in class.)

End of 1/3/2011 Lecture.

Example 2.5. Suppose that $P(X \in dx, Y \in dy) = e^{-y} 1_{0<x<y} dx dy$. Recall that

$$\int_0^\infty y^k e^{-\lambda y} dy = \left(-\frac{d}{d\lambda}\right)^k \int_0^\infty e^{-\lambda y} dy = \left(-\frac{d}{d\lambda}\right)^k \frac{1}{\lambda} = k! \frac{1}{\lambda^{k+1}}.$$

Therefore,

$$\mathbb{E}Y = \int\int y e^{-y} 1_{0<x<y} dx dy = \int_0^\infty y^2 e^{-y} dy = 2,$$
$$\mathbb{E}Y^2 = \int\int y^2 e^{-y} 1_{0<x<y} dx dy = \int_0^\infty y^3 e^{-y} dy = 3! = 6$$

$$\mathbb{E}X = \int\int x e^{-y} 1_{0<x<y} dx dy = \frac{1}{2}\int_0^\infty y^2 e^{-y} dy = 1,$$
$$\mathbb{E}X^2 = \int\int x^2 e^{-y} 1_{0<x<y} dx dy = \frac{1}{3}\int_0^\infty y^3 e^{-y} dy = \frac{1}{3} 3! = 2$$

and

$$\mathbb{E}[XY] = \int\int xy e^{-y} 1_{0<x<y} dx dy = \frac{1}{2}\int_0^\infty y^3 e^{-y} dy = \frac{3!}{2} = 3.$$

Therefore $\operatorname{Cov}(X, Y) = 3 - 2 \cdot 1 = 1$, $\sigma^2(X) = 2 - 1^2 = 1$, $\sigma^2(Y) = 6 - 2^2 = 2$,

$$Corr(X, Y) = \frac{1}{\sqrt{2}}.$$

Lemma 2.6. *The covariance function,* $\operatorname{Cov}(X, Y)$ *is bilinear in* X *and* Y *and* $\operatorname{Cov}(X, Y) = 0$ *if either* X *or* Y *is constant. For any constant* k, $\operatorname{Var}(X + k) = \operatorname{Var}(X)$ *and* $\operatorname{Var}(kX) = k^2 \operatorname{Var}(X)$. *If* $\{X_k\}_{k=1}^n$ *are uncorrelated* $L^2(P)$ *– random variables, then*

$$\operatorname{Var}(S_n) = \sum_{k=1}^n \operatorname{Var}(X_k).$$

Proof. We leave most of this simple proof to the reader. As an example of the type of argument involved, let us prove $\operatorname{Var}(X + k) = \operatorname{Var}(X)$;

$$\begin{aligned}\operatorname{Var}(X + k) &= \operatorname{Cov}(X + k, X + k) = \operatorname{Cov}(X + k, X) + \operatorname{Cov}(X + k, k) \\ &= \operatorname{Cov}(X + k, X) = \operatorname{Cov}(X, X) + \operatorname{Cov}(k, X) \\ &= \operatorname{Cov}(X, X) = \operatorname{Var}(X),\end{aligned}$$

wherein we have used the bilinearity of $\operatorname{Cov}(\cdot, \cdot)$ and the property that $\operatorname{Cov}(Y, k) = 0$ whenever k is a constant. ∎

Example 2.7. Suppose that X and Y are distributed as follows;

$$\begin{array}{cccccc} & \rho_Y & 1/4 & \frac{1}{2} & 1/4 \\ \rho_X & X\backslash Y & -1 & 0 & 1 \\ 1/4 & 1 & 0 & 1/4 & 0 \\ 3/4 & 0 & 1/4 & 1/4 & 1/4 \end{array}$$

so that $\rho_{X,Y}(1,-1) = P(X=1,Y=-1) = 0$, $\rho_{X,Y}(1,0) = P(X=1,Y=0) = 1/4$, etc. In this case $XY = 0$ a.s. so that $\mathbb{E}[XY] = 0$ while

$$\mathbb{E}[X] = 1 \cdot \frac{1}{4} + 0 \cdot \frac{3}{4} = \frac{1}{4}, \text{ and}$$
$$\mathbb{E}Y = (-1)\,1/4 + 0\frac{1}{2} + 1\frac{1}{4} = 0$$

so that $\operatorname{Cov}(X,Y) = 0 - \frac{1}{4} \cdot 0 = 0$. Again X and Y are not independent since $\rho_{X,Y}(x,y) \neq \rho_X(x)\,\rho_Y(y)$.

Example 2.8. Let X have an even distribution and let $Y = X^2$, then

$$\operatorname{Cov}(X,Y) = \mathbb{E}\left[X^3\right] - \mathbb{E}\left[X^2\right] \cdot \mathbb{E}X = 0$$

since,

$$\mathbb{E}\left[X^{2k+1}\right] = \int_{-\infty}^{\infty} x^{2k+1} \rho(x)\,dx = 0 \text{ for all } k \in \mathbb{N}.$$

On the other hand $\operatorname{Cov}\left(Y, X^2\right) = \operatorname{Cov}(Y,Y) = \operatorname{Var}(Y) \neq 0$ in general so that Y is not independent of X.

Example 2.9 (Not done in class.). Let X and Z be independent with $P(Z = \pm 1) = \frac{1}{2}$ and take $Y = XZ$. Then $\mathbb{E}Z = 0$ and

$$\begin{aligned} \operatorname{Cov}(X,Y) &= \mathbb{E}\left[X^2 Z\right] - \mathbb{E}[X]\,\mathbb{E}[XZ] \\ &= \mathbb{E}\left[X^2\right] \cdot \mathbb{E}Z - \mathbb{E}[X]\,\mathbb{E}[X]\,\mathbb{E}Z = 0. \end{aligned}$$

On the other hand it should be intuitively clear that X and Y are not independent since knowledge of X typically will give some information about Y. To verify this assertion let us suppose that X is a discrete random variable with $P(X=0) = 0$. Then

$$P(X = x, Y = y) = P(X = x, xZ = y) = P(X = x) \cdot P(X = y/x)$$

while

$$P(X = x)\,P(Y = y) = P(X = x) \cdot P(XZ = y).$$

Thus for X and Y to be independent we would have to have,

$$P(xX = y) = P(XZ = y) \text{ for all } x, y.$$

This is clearly not going to be true in general. For example, suppose that $P(X=1) = \frac{1}{2} = P(X=0)$. Taking $x = y = 1$ in the previously displayed equation would imply

$$\frac{1}{2} = P(X=1) = P(XZ=1) = P(X=1, Z=1) = P(X=1)\,P(Z=1) = \frac{1}{4}$$

which is false.

Presumably you saw the following exercise in Math 180A.

Exercise 2.2 (A Weak Law of Large Numbers). Assume $\{X_n\}_{n=1}^{\infty}$ is a sequence if uncorrelated square integrable random variables which are identically distributed, i.e. $X_n \overset{d}{=} X_m$ for all $m, n \in \mathbb{N}$. Let $S_n := \sum_{k=1}^n X_k$, $\mu := \mathbb{E}X_k$ and $\sigma^2 := \operatorname{Var}(X_k)$ (these are independent of k). Show;

$$\mathbb{E}\left[\frac{S_n}{n}\right] = \mu,$$
$$\mathbb{E}\left(\frac{S_n}{n} - \mu\right)^2 = \operatorname{Var}\left(\frac{S_n}{n}\right) = \frac{\sigma^2}{n}, \text{ and}$$
$$P\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}$$

for all $\varepsilon > 0$ and $n \in \mathbb{N}$.

3

Geometric aspects of $L^2(P)$

Definition 3.1 (Inner Product). *For $X, Y \in L^2(P)$, let $(X,Y) := \mathbb{E}[XY]$ and $\|X\| := \sqrt{(X,X)} = \sqrt{\mathbb{E}|X^2|}$.*

Example 3.2 (This was already mentioned in Lecture 1 with $N = 4$.). Suppose that $\Omega = \{1, \dots, N\}$ and $P(\{i\}) = \frac{1}{N}$ for $1 \le i \le N$. Then

$$(X,Y) = \mathbb{E}[XY] = \frac{1}{N}\sum_{i=1}^{N} X(i)\,Y(i) = \frac{1}{N}\mathbf{X}\cdot\mathbf{Y}$$

where

$$\mathbf{X} := \begin{bmatrix} X(1) \\ X(2) \\ \vdots \\ X(N) \end{bmatrix} \text{ and } \mathbf{Y} := \begin{bmatrix} Y(1) \\ Y(2) \\ \vdots \\ Y(N) \end{bmatrix}.$$

Thus the inner product we have defined in this case is essentially the dot product that you studied in math 20F.

Remark 3.3. The inner product on $H := L^2(P)$ satisfies,

1. $(aX + bY, Z) = a(X,Z) + b(Y,Z)$ i.e. $X \to (X,Z)$ is linear.
2. $(X,Y) = (Y,X)$ (symmetry).
3. $\|X\|^2 := (X,X) \ge 0$ with $\|X\|^2 = 0$ iff $X = 0$.

Notice that combining properties (1) and (2) that $X \to (Z,X)$ is linear for fixed $Z \in H$, i.e.

$$(Z, aX + bY) = \bar{a}(Z,X) + \bar{b}(Z,Y).$$

The following identity will be used frequently in the sequel without further mention,

$$\begin{aligned} \|X+Y\|^2 &= (X+Y, X+Y) = \|X\|^2 + \|Y\|^2 + (X,Y) + (Y,X) \\ &= \|X\|^2 + \|Y\|^2 + 2(X,Y). \end{aligned} \tag{3.1}$$

Theorem 3.4 (Schwarz Inequality). *Let $(H, (\cdot,\cdot))$ be an inner product space, then for all $X, Y \in H$*

$$|(X,Y)| \le \|X\|\|Y\|$$

and equality holds iff X and Y are linearly dependent. Applying this result to $|X|$ and $|Y|$ shows,

$$\mathbb{E}[|XY|] \le \|X\| \cdot \|Y\|.$$

Proof. If $Y = 0$, the result holds trivially. So assume that $Y \neq 0$ and observe; if $X = \alpha Y$ for some $\alpha \in \mathbb{C}$, then $(X,Y) = \alpha\|Y\|^2$ and hence

$$|(X,Y)| = |\alpha|\,\|Y\|^2 = \|X\|\|Y\|.$$

Now suppose that $X \in H$ is arbitrary, let $Z := X - \|Y\|^{-2}(X,Y)Y$. (So $\|Y\|^{-2}(X,Y)Y$ is the "orthogonal projection" of X along Y, see Figure 3.1.)

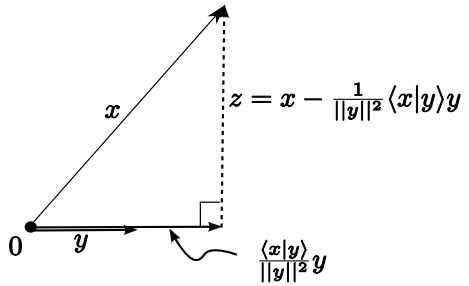

Fig. 3.1. The picture behind the proof of the Schwarz inequality.

Then

$$\begin{aligned} 0 \le \|Z\|^2 &= \left\| X - \frac{(X,Y)}{\|Y\|^2}Y \right\|^2 = \|X\|^2 + \frac{|(X|Y)|^2}{\|Y\|^4}\|Y\|^2 - 2(X|\frac{(X|Y)}{\|Y\|^2}Y) \\ &= \|X\|^2 - \frac{|(X|Y)|^2}{\|Y\|^2} \end{aligned}$$

from which it follows that $0 \le \|Y\|^2\|X\|^2 - |(X|Y)|^2$ with equality iff $Z = 0$ or equivalently iff $X = \|Y\|^{-2}(X|Y)Y$.

Alternative argument: Let $c \in \mathbb{R}$ and $Z := X - cY$, then

$$0 \le \|Z\|^2 = \|X - cY\|^2 = \|X\|^2 - 2c(X,Y) + c^2\|Y\|^2.$$

The right side of this equation is minimized at $c = (X,Y)/\|Y\|^2$ and for this valued of c we find,

$$0 \le \|X - cY\|^2 = \|X\|^2 - (X,Y)^2/\|Y\|^2$$

with equality iff $X = cY$. Solving this last inequality for $|(X,Y)|$ gives the result. ∎

Corollary 3.5. *The norm, $\|\cdot\|$, satisfies the triangle inequality and $(\cdot,\cdot)$ is continuous on $H \times H$.*

Proof. If $X, Y \in H$, then, using Schwarz's inequality,

$$\begin{aligned}\|X+Y\|^2 &= \|X\|^2 + \|Y\|^2 + 2(X,Y)\\ &\le \|X\|^2 + \|Y\|^2 + 2\|X\|\|Y\| = (\|X\| + \|Y\|)^2.\end{aligned}$$

Taking the square root of this inequality shows $\|\cdot\|$ satisfies the triangle inequality. (The rest of this proof may be skipped.)

Checking that $\|\cdot\|$ satisfies the remaining axioms of a norm is now routine and will be left to the reader. If $X, Y, \Delta X, \Delta Y \in H$, then

$$\begin{aligned}|(X+\Delta X, Y+\Delta Y) - (X,Y)| &= |(X,\Delta Y) + (\Delta X, Y) + (\Delta X, \Delta Y)|\\ &\le \|X\|\|\Delta Y\| + \|Y\|\|\Delta X\| + \|\Delta X\|\|\Delta Y\|\\ &\to 0 \text{ as } \Delta X, \Delta Y \to 0,\end{aligned}$$

from which it follows that $(\cdot,\cdot)$ is continuous. ∎

Definition 3.6. *Let $(H,(\cdot,\cdot))$ be an inner product space, we say $X, Y \in H$ are **orthogonal** and write $X \perp Y$ iff $(X,Y) = 0$. More generally if $A \subset H$ is a set, $X \in H$ is **orthogonal to** A (write $X \perp A$) iff $(X,Y) = 0$ for all $Y \in A$. Let $A^\perp = \{X \in H : X \perp A\}$ be the set of vectors orthogonal to A. A subset $S \subset H$ is an **orthogonal set** if $X \perp Y$ for all distinct elements $X, Y \in S$. If S further satisfies, $\|X\| = 1$ for all $X \in S$, then S is said to be an **orthonormal set.***

Proposition 3.7. *Let $(H,(\cdot,\cdot))$ be an inner product space then*

1. ***(Pythagorean Theorem)** If $S \subset\subset H$ is a finite orthogonal set, then*

$$\left\|\sum_{X\in S} X\right\|^2 = \sum_{X\in S}\|X\|^2. \tag{3.2}$$

2. ***(Parallelogram Law)** (Skip this one.) For all $X, Y \in H$,*

$$\|X+Y\|^2 + \|X-Y\|^2 = 2\|X\|^2 + 2\|Y\|^2 \tag{3.3}$$

Proof. Items 1. and 2. are proved by the following elementary computations;and

$$\begin{aligned}\left\|\sum_{X\in S} X\right\|^2 &= \left(\sum_{X\in S} X, \sum_{Y\in S} Y\right) = \sum_{X,Y\in S}(X,Y)\\ &= \sum_{X\in S}(X,X) = \sum_{X\in S}\|X\|^2\end{aligned}$$

and

$$\begin{aligned}&\|X+Y\|^2 + \|X-Y\|^2\\ &\quad= \|X\|^2 + \|Y\|^2 + 2(X,Y) + \|X\|^2 + \|Y\|^2 - 2(X,Y)\\ &\quad= 2\|X\|^2 + 2\|Y\|^2.\end{aligned}$$

∎

Theorem 3.8 (Least Squares Approximation Theorem). *Suppose that V is a subspace of $H := L^2(P)$, $X \in V$, and $Y \in L^2(P)$. Then the following are equivalent;*

1. *$\|Y - X\| \ge \|Y - Z\|$ for all $Z \in V$ (i.e. X is a least squares approximation to Y by an element from V) and*
2. *$(Y - X) \perp V$.*

*Moreover there is "essentially" at most one $X \in V$ satisfying 1. or equivalently 2. We denote random variable by $Q_V Y$ and call it **orthogonal projection of Y along V.***

Proof. 1 $\implies$ 2. If 1. holds then $f(t) := \|Y - (X + tZ)\|^2$ has a minimum at $t = 0$ and therefore $\dot f(0) = 0$. Since

$$f(t) := \|Y - X - tZ\|^2 = \|Y - X\|^2 + t^2\|Z\|^2 - 2t(Y - X, Z),$$

we may conclude that

$$0 = \dot f(0) = -2(Y - X, Z).$$

As $Z \in V$ was arbitrary we may conclude that $(Y - X) \perp V$.

2 $\implies$ 1. Now suppose that $(Y - X) \perp V$ and $Z \in V$, then $(Y - X) \perp (X - Z)$ and so

$$\|Y - Z\|^2 = \|Y - X + X - Z\|^2 = \|Y - X\|^2 + \|X - Z\|^2 \ge \|Y - X\|^2. \tag{3.4}$$

Moreover if Z is another best approximation to Y then $\|Y - Z\|^2 = \|Y - X\|^2$ which happens according to Eq. (3.4) iff

$$\|X - Z\|^2 = \mathbb{E}(X - Z)^2 = 0,$$

i.e. iff $X = Z$ a.s. ∎

End of Lecture 3: 1/07/2011 (Given by Tom Laetsch)

Corollary 3.9 (Orthogonal Projection Formula). *Suppose that V is a subspace of $H := L^2(P)$ and $\{X_i\}_{i=1}^N$ is an orthogonal basis for V. Then*

$$Q_V Y = \sum_{i=1}^{N} \frac{(Y, X_i)}{\|X_i\|^2} X_i \text{ for all } Y \in H.$$

Proof. The best approximation $X \in V$ to Y is of the form $X = \sum_{i=1}^N c_i X_i$ where $c_i \in \mathbb{R}$ need to be chosen so that $(Y - X) \perp V$. Equivalently put we must have

$$0 = (Y - X, X_j) = (Y, X_j) - (X, X_j) \text{ for } 1 \leq j \leq N.$$

Since

$$(X, X_j) = \sum_{i=1}^{N} c_i (X_i, X_j) = c_j \|X_j\|^2,$$

we see that $c_j = (Y, X_j) / \|X_j\|^2$, i.e.

$$Q_V Y = X = \sum_{i=1}^{N} \frac{(Y, X_i)}{\|X_i\|^2} X_i.$$

■

Example 3.10. Given $Y \in L^2(P)$ the best approximation to Y by a constant function c is given by

$$c = \frac{\mathbb{E}[Y1]}{\mathbb{E}1^2} 1 = \mathbb{E}Y.$$

You already proved this on your first homework by a direct calculus exercise.

Linear prediction and a canonical form

Theorem 4.1 (Linear Prediction Theorem). *Let X and Y be two square integrable random variables, then*

$$\sigma(Y)\sqrt{1-Corr^2(X,Y)} = \min_{a,b\in\mathbb{R}} \|Y-(aX+b)\| = \|Y-W\|$$

where

$$W = \mu_Y + \frac{\operatorname{Cov}(X,Y)}{\operatorname{Var}(X)}(X-\mu_X) = \frac{\operatorname{Cov}(X,Y)}{\operatorname{Var}(X)}X + \left(\mathbb{E}Y - \mu_X\frac{\operatorname{Cov}(X,Y)}{\operatorname{Var}(X)}\right).$$

Proof. Let $\mu = \mathbb{E}X$ and $\bar{X} = X-\mu$. Then $\left\{1,\bar{X}\right\}$ is an orthogonal set and $V := \operatorname{span}\{1,X\} = \operatorname{span}\left\{1,\bar{X}\right\}$. Thus best approximation of Y by random variable of the form $aX+b$ is given by

$$W = (Y,1)\,1 + \frac{\left(Y,\bar{X}\right)}{\left\|\bar{X}\right\|^2}\bar{X} = \mathbb{E}Y + \frac{\operatorname{Cov}(X,Y)}{\operatorname{Var}(X)}(X-\mu_X).$$

The **root mean square error** of this approximation is

$$\begin{aligned}\|Y-W\|^2 &= \left\|\bar{Y} - \frac{\operatorname{Cov}(X,Y)}{\operatorname{Var}(X)}\bar{X}\right\|^2 = \sigma^2(Y) - \frac{\operatorname{Cov}^2(X,Y)}{\sigma^2(X)}\\ &= \sigma^2(Y)\left(1-Corr^2(X,Y)\right),\end{aligned}$$

so that

$$\|Y-W\| = \sigma(Y)\sqrt{1-Corr^2(X,Y)}.$$

■

Example 4.2. Suppose that $P(X\in dx, Y\in dy) = e^{-y}1_{0<x<y}dxdy$. Recall from Example 2.5 that

$$\begin{aligned}\mathbb{E}X = 1, &\qquad \mathbb{E}Y = 2,\\ \mathbb{E}X^2 = 2, &\qquad \mathbb{E}Y^2 = 6\\ \sigma(X) = 1, &\qquad \sigma(Y) = \sqrt{2},\end{aligned}$$

$$\operatorname{Cov}(X,Y) = 1, \text{ and } Corr(X,Y) = \frac{1}{\sqrt{2}}.$$

So in this case

$$W = 2 + \frac{1}{1}(X-1) = X+1$$

is the best linear predictor of Y and the root mean square error in this prediction is

$$\|Y-W\| = \sqrt{2}\sqrt{1-\frac{1}{2}} = 1.$$

Corollary 4.3 (Correlation Bounds). *For all square integrable random variables, X and Y,*

$$|\operatorname{Cov}(X,Y)| \le \sigma(X)\cdot\sigma(Y)$$

or equivalently,

$$|Corr(X,Y)| \le 1.$$

.

Proof. This is a simply application of Schwarz's inequality (Theorem 3.4);

$$|\operatorname{Cov}(X,Y)| = |\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]| \le \|X-\mu_X\|\cdot\|Y-\mu_Y\| = \sigma(X)\cdot\sigma(Y).$$

■

Theorem 4.4 (Canonical form). *If $X,Y\in L^2(P)$, then there are two mean zero uncorrelated Random variables $\{Z_1,Z_2\}$ such that $\|Z_1\| = \|Z_2\| = 1$ and*

$$\begin{aligned}X &= \mu_X + \sigma(X)Z_1, \text{ and}\\ Y &= \mu_Y + \sigma(Y)\left[\cos\theta\cdot Z_1 + \sin\theta\cdot Z_2\right],\end{aligned}$$

where $0\le\theta\le\pi$ is chosen such that $\cos\theta := Corr(X,Y)$.

Proof. (Just sketch the main ideal in class!). The proof amounts to applying the Gram-Schmidt procedure to $\left\{\bar{X} := X-\mu_X, \bar{Y} := Y-\mu_Y\right\}$ to find Z_1 and Z_2 followed by expressing X and Y in uniquely in terms of the linearly independent set, $\{1,Z_1,Z_2\}$. The details follow.

Performing Gram-Schmidt on $\left\{\bar{X},\bar{Y}\right\}$ gives $Z_1 = \bar{X}/\sigma(X)$ and

$$\tilde{Z}_2 = \bar{Y} - \frac{\left(\bar{Y},\bar{X}\right)}{\sigma(X)^2}\bar{X}.$$

To get Z_2 we need to normalize $\tilde{Z}_2$ using;

$$\begin{aligned}\mathbb{E}\tilde{Z}_2^2 &= \sigma(Y)^2 - 2\frac{\left(\bar{Y},\bar{X}\right)}{\sigma(X)^2}\left(\bar{X},\bar{Y}\right) + \frac{\left(\bar{Y},\bar{X}\right)^2}{\sigma(X)^4}\sigma(X)^2 \\ &= \sigma(Y)^2 - \frac{\left(\bar{X},\bar{Y}\right)^2}{\sigma(X)^2} = \sigma(Y)^2\left(1 - Corr^2(X,Y)\right) \\ &= \sigma(Y)^2 \sin^2\theta.\end{aligned}$$

Therefore $Z_1 = \bar{X}/\sigma(X)$ and

$$\begin{aligned}Z_2 &:= \frac{\tilde{Z}_2}{\left\|\tilde{Z}_2\right\|} = \frac{\bar{Y} - \frac{\left(\bar{Y},\bar{X}\right)}{\sigma(X)^2}\bar{X}}{\sigma(Y)\sin\theta} = \frac{\bar{Y} - \frac{\sigma(X)\sigma(Y)Corr(X,Y)}{\sigma(X)^2}\bar{X}}{\sigma(Y)\sin\theta} \\ &= \frac{\bar{Y} - \frac{\sigma(Y)}{\sigma(X)}\cos\theta\cdot\bar{X}}{\sigma(Y)\sin\theta} = \frac{\bar{Y} - \sigma(Y)\cos\theta\cdot Z_1}{\sigma(Y)\sin\theta}\end{aligned}$$

Solving for $\bar{X}$ and $\bar{Y}$ shows,

$$\bar{X} = \sigma(X) Z_1 \text{ and } \bar{Y} = \sigma(Y)\left[\sin\theta\cdot Z_2 + \cos\theta\cdot Z_1\right]$$

which is equivalent to the desired result. ∎

Corollary 4.5. *If* $Corr(X,Y) = 1$, *then*

$$\bar{Y} = \sigma(Y) Z_1 = \frac{\sigma(Y)}{\sigma(X)}\bar{X}.$$

If $Corr(X,Y) = -1$ *then*

$$\bar{Y} = -\sigma(Y) Z_1 = -\frac{\sigma(Y)}{\sigma(X)}\bar{X}.$$

Exercise 4.1 (A correlation inequality). Suppose that X is a random variable and $f, g : \mathbb{R}\to\mathbb{R}$ are two increasing functions such that both $f(X)$ and $g(X)$ are square integrable, i.e. $\mathbb{E}|f(X)|^2 + \mathbb{E}|g(X)|^2 < \infty$. Show $\mathrm{Cov}(f(X), g(X)) \geq 0$. **Hint:** let Y be another random variable which has the same law as X and is independent of X. Then consider

$$\mathbb{E}\left[(f(Y) - f(X))\cdot(g(Y) - g(X))\right].$$

5

Conditional Expectation

Notation 5.1 (Conditional Expectation 1) *Given $Y \in L^1(P)$ and $A \subset \Omega$ let*

$$\mathbb{E}[Y:A] := \mathbb{E}[1_A Y]$$

and

$$\mathbb{E}[Y|A] = \begin{cases} \mathbb{E}[Y:A]/P(A) & \text{if } P(A) > 0 \\ 0 & \text{if } P(A) = 0 \end{cases}. \tag{5.1}$$

(In point of fact, when $P(A) = 0$ we could set $\mathbb{E}[Y|A]$ to be any real number. We choose 0 for definiteness and so that $Y \to \mathbb{E}[Y|A]$ is always linear.)

Lemma 5.2. *If $P(A) > 0$ then $\mathbb{E}[Y|A] = \mathbb{E}_{P(\cdot|A)} Y$ for all $Y \in L^1(P)$.*

Proof. I will only prove $\mathbb{E}[Y|A] = \mathbb{E}_{P(\cdot|A)} Y$ when Y is discrete although the result does hold in general. In the discrete case,

$$\begin{aligned} \mathbb{E}_{P(\cdot|A)} Y &= \mathbb{E}_{P(\cdot|A)} \sum_{y \in \mathbb{R}} y 1_{Y=y} = \sum_{y \in \mathbb{R}} y \mathbb{E}_{P(\cdot|A)} 1_{Y=y} = \sum_{y \in \mathbb{R}} y P(Y = y|A) \\ &= \sum_{y \in \mathbb{R}} y P(Y = y|A) = \sum_{y \in \mathbb{R}} y \frac{P(Y = y, A)}{P(A)} = \frac{1}{P(A)} \sum_{y \in \mathbb{R}} y \mathbb{E}[1_A 1_{Y=y}] \\ &= \frac{1}{P(A)} \mathbb{E}\left[1_A \sum_{y \in \mathbb{R}} y 1_{Y=y} \right] = \frac{1}{P(A)} \mathbb{E}[1_A Y] = \mathbb{E}[Y|A]. \end{aligned}$$

■

Lemma 5.3. *No matter whether $P(A) > 0$ or $P(A) = 0$ we always have,*

$$|\mathbb{E}[Y|A]| \le \mathbb{E}[|Y||A] \le \sqrt{\mathbb{E}\left[|Y|^2 |A\right]}. \tag{5.2}$$

Proof. If $P(A) = 0$ then all terms in Eq. (5.2) are zero and so the inequalities hold. For $P(A) > 0$ we have, using the Schwarz inequality in Theorem 3.4), that

$$|\mathbb{E}[Y|A]| = \left|\mathbb{E}_{P(\cdot|A)} Y\right| \le \mathbb{E}_{P(\cdot|A)} |Y| \le \sqrt{\mathbb{E}_{P(\cdot|A)} |Y|^2 \cdot \mathbb{E}_{P(\cdot|A)} 1} = \sqrt{\mathbb{E}_{P(\cdot|A)} |Y|^2}.$$

This completes that proof as $\mathbb{E}_{P(\cdot|A)} |Y| = \mathbb{E}[|Y||A]$ and $\mathbb{E}_{P(\cdot|A)} |Y|^2 = \mathbb{E}\left[|Y|^2 |A\right]$. ■

Notation 5.4 *Let S be a set (often $S = \mathbb{R}$ or $S = \mathbb{R}^N$) and suppose that $X : \Omega \to S$ is a function. (So X is a random variable if $S = \mathbb{R}$ and a random vector when $S = \mathbb{R}^N$.) Further let V_X denote those random variables $Z \in L^2(P)$ which may be written as $Z = f(X)$ for some function $f : S \to \mathbb{R}$. (This is a subspace of $L^2(P)$ and we let $\mathcal{F}_X := \left\{ f : S \to \mathbb{R} : f(X) \in L^2(P) \right\}$.)*

Definition 5.5 (Conditional Expectation 2). *Given a function $X : \Omega \to S$ and $Y \in L^2(P)$, we define $\mathbb{E}[Y|X] := Q_{V_X} Y$ where Q_{V_X} is orthogonal projection onto V_X. (**Fact:** $Q_{V_X} Y$ always exists. The proof requires technical details beyond the scope of this course.)*

Remark 5.6. By definition, $\mathbb{E}[Y|X] = h(X)$ where $h \in \mathcal{F}_X$ is chosen so that $[Y - h(X)] \perp V_X$, i.e. $\mathbb{E}[Y|X] = h(X)$ iff $(Y - h(X), f(X)) = 0$ for all $f \in \mathcal{F}_X$. So in summary, $\mathbb{E}[Y|X] = h(X)$ iff

$$\mathbb{E}[Y f(X)] = \mathbb{E}[h(X) f(X)] \text{ for all } f \in \mathcal{F}_X. \tag{5.3}$$

Corollary 5.7 (Law of total expectation). *For all random variables $Y \in L^2(P)$, we have $\mathbb{E}Y = \mathbb{E}(\mathbb{E}(Y|X))$.*

Proof. Take $f = 1$ in Eq. (5.3). ■

This notion of conditional expectation is rather abstract. It is now time to see how to explicitly compute conditional expectations. (In general this can be quite tricky to carry out in concrete examples!)

5.1 Conditional Expectation for Discrete Random Variables

Recall that if A and B are events with $P(A) > 0$, then we define $P(B|A) := \frac{P(B \cap A)}{P(A)}$. By convention we will set $P(B|A) = 0$ if $P(A) = 0$.

Example 5.8. If Ω is a finite set with N elements, P is the uniform distribution on Ω, and A is a non-empty subset of Ω, then $P(\cdot|A)$ restricted to events contained in A is the uniform distribution on A. Indeed, $a = \#(A)$ and $B \subset A$, we have

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(B)}{P(A)} = \frac{\#(B)/N}{\#(A)/N} = \frac{\#(B)}{\#(A)} = \frac{\#(B)}{a}.$$

Theorem 5.9. *Suppose that S is a finite or countable set and $X:\Omega\to S$, then $\mathbb{E}[Y|X]=h(X)$ where $h(s):=\mathbb{E}[Y|X=s]$ for all $s\in S$.*

Proof. First Proof. Our goal is to find $h(s)$ such that

$$\mathbb{E}[Yf(X)]=\mathbb{E}[h(X)f(X)] \text{ for all bounded } f.$$

Let $S'=\{s\in S:P(X=s)>0\}$, then

$$\begin{aligned}\mathbb{E}[Yf(X)]&=\sum_{s\in S}\mathbb{E}[Yf(X):X=s]=\sum_{s\in S'}\mathbb{E}[Yf(X):X=s]\\&=\sum_{s\in S'}f(s)\,\mathbb{E}[Y|X=s]\cdot P(X=s)\\&=\sum_{s\in S'}f(s)\,h(s)\cdot P(X=s)\\&=\sum_{s\in S}f(s)\,h(s)\cdot P(X=s)=\mathbb{E}[h(X)f(X)]\end{aligned}$$

where $h(s):=\mathbb{E}[Y|X=s]$.

Second Proof. If S is a finite set, such that $P(X=s)>0$ for all $s\in S$. Then

$$f(X)=\sum_{s\in S}f(s)\,1_{X=s}$$

which shows that $V_X=\operatorname{span}\{1_{X=s}:s\in S\}$. As $\{1_{X=s}\}_{s\in S}$ is an orthogonal set, we may compute

$$\begin{aligned}\mathbb{E}[Y|X]&=\sum_{s\in S}\frac{(Y,1_{X=s})}{\|1_{X=s}\|^2}1_{X=s}=\sum_{s\in S}\frac{\mathbb{E}[Y:X=s]}{P(X=s)}1_{X=s}\\&=\sum_{s\in S}\mathbb{E}[Y|X=s]\cdot 1_{X=s}=h(X).\end{aligned}$$

■

Example 5.10. Suppose that X and Y are discrete random variables with joint distribution given as;

	ρ_Y	$1/4$	$\frac{1}{2}$	$1/4$
ρ_X	$X\backslash Y$	-1	0	1
$1/4$	1	0	$1/4$	0
$3/4$	0	$1/4$	$1/4$	$1/4$

We then have

$$\mathbb{E}[Y|X=1]=\frac{1}{1/4}\left(-1\cdot 0+0\cdot\frac{1}{4}+1\cdot 0\right)=0 \text{ and}$$

$$\mathbb{E}[Y|X=0]=\frac{1}{3/4}\left(-1\cdot 1/4+0\cdot\frac{1}{4}+1\cdot 1/4\right)=0$$

and therefore $\mathbb{E}[Y|X]=0$. On the other hand,

$$\begin{aligned}\mathbb{E}[X|Y=-1]&=\frac{1}{1/4}\left(1\cdot 0+0\cdot\frac{1}{4}\right)=0,\\\mathbb{E}[X|Y=0]&=\frac{1}{1/2}\left(1\cdot 1/4+0\cdot\frac{1}{4}\right)=\frac{1}{2}, \text{ and}\\\mathbb{E}[X|Y=1]&=\frac{1}{1/4}\left(1\cdot 0+0\cdot\frac{1}{4}\right)=0.\end{aligned}$$

Therefore

$$\mathbb{E}[X|Y]=\frac{1}{2}1_{Y=0}.$$

Example 5.11. Let X and Y be discrete random variables with values in $\{1,2,3\}$ whose joint distribution and marginals are given by

	ρ_X	$.3$	$.35$	$.35$
ρ_Y	$Y\backslash X$	1	2	3
$.6$	1	$.1$	$.2$	$.3$
$.3$	2	$.15$	$.15$	0
$.1$	3	$.05$	0	$.05$

.

Then

$$\begin{aligned}\rho_{X|Y}(1,3)&=P(X=1|Y=3)=\frac{.05}{.1}=\frac{1}{2},\\\rho_{X|Y}(2,3)&=P(X=2|Y=3)=\frac{0}{.1}=0, \text{ and}\\\rho_{X|Y}(3,3)&=P(X=3|Y=3)=\frac{.05}{.1}=\frac{1}{2}.\end{aligned}$$

Therefore,

$$\mathbb{E}[X|Y=3]=1\cdot\frac{1}{2}+2\cdot 0+3\cdot\frac{1}{2}=2$$

or

$$h(3):=\mathbb{E}[X|Y=3]=\frac{1}{.1}(1\cdot.05+2\cdot 0+3\cdot.05)=2$$

Similarly,

$$\begin{aligned}h(1):=\mathbb{E}[X|Y=1]&=\frac{1}{.6}(1\cdot.1+2\cdot.2+3\cdot.3)=2\frac{1}{3},\\h(2):=\mathbb{E}[X|Y=2]&=\frac{1}{.3}(1\cdot.15+2\cdot.15+3\cdot 0)=1.5\end{aligned}$$

and so

$$\mathbb{E}[X|Y]=h(Y)=2\frac{1}{3}\cdot 1_{Y=1}+1.5\cdot 1_{Y=2}+2\cdot 1_{Y=3}.$$

Example 5.12 (Number of girls in a family). Suppose the number of children in a family is a random variable X with mean μ, and given $X = n$ for $n \geq 1$, each of the n children in the family is a girl with probability p and a boy with probability $1-p$. Problem. What is the expected number of girls in a family?

Solution. Intuitively, the answer should be $p\mu$. To show this is correct let G be the random number of girls in a family. Then,

$$\mathbb{E}[G|X=n] = p \cdot n$$

as $G = 1_{A_1} + \cdots + 1_{A_n}$ on $X = n$ where A_i is the event the i^{th} – child is a girl. We are given $P(A_i|X=n) = p$ so that $\mathbb{E}[1_{A_i}|X=n] = p$ and so $\mathbb{E}[G|X=n] = p \cdot n$. Therefore, $\mathbb{E}[G|X] = p \cdot X$ and

$$\mathbb{E}[G] = \mathbb{E}\mathbb{E}[G|X] = \mathbb{E}[p \cdot X] = p\mu.$$

$$\begin{aligned}\mathbb{E}Y = \mathbb{E}[\mathbb{E}[Y|X]] &= 1 \cdot \frac{12}{36} + \frac{45}{9} \cdot \frac{6}{36} + \frac{46}{10} \cdot \frac{8}{36} + \frac{47}{11} \cdot \frac{10}{36} \\ &= \frac{557}{165} \cong 3.376 \text{ rolls.}\end{aligned}$$

Example 5.13. Suppose that X and Y are i.i.d. random variables with the geometric distribution,

$$P(X=k) = P(Y=k) = (1-p)^{k-1} p \text{ for } k \in \mathbb{N}.$$

We compute, for $n > m$,

$$\begin{aligned}P(X=m|X+Y=n) &= \frac{P(X=m, X+Y=n)}{P(X+Y=n)} \\ &= \frac{P(X=m, Y=n-m)}{\sum_{k+l=n} P(X=k, Y=l)}\end{aligned}$$

where

$$\begin{aligned}P(X=m, Y=n-m) &= p^2 (1-p)^{m-1} (1-p)^{n-m-1} \\ &= p^2 (1-p)^{n-2}\end{aligned}$$

and

$$\begin{aligned}\sum_{k+l=n} P(X=k, Y=l) &= \sum_{k+l=n} (1-p)^{k-1} p (1-p)^{l-1} p \\ &= \sum_{k+l=n} p^2 (1-p)^{n-2} = p^2 (1-p)^{n-2} \sum_{k=1}^{n-1} 1.\end{aligned}$$

Thus we have shown,

$$P(X=m|X+Y=n) = \frac{1}{n-1} \text{ for } 1 \leq m < n.$$

From this it follows that

$$\mathbb{E}[f(X)|X+Y=n] = \frac{1}{n-1} \sum_{m=1}^{n-1} f(m)$$

and so

$$\mathbb{E}[f(X)|X+Y] = \frac{1}{X+Y-1} \sum_{m=1}^{X+Y-1} f(m).$$

As a check if $f(m) = m$ we have

$$\begin{aligned}\mathbb{E}[X|X+Y] &= \frac{1}{X+Y-1} \sum_{m=1}^{X+Y-1} m \\ &= \frac{1}{X+Y-1} \frac{1}{2} (X+Y-1)(X+Y-1+1) \\ &= \frac{1}{2}(X+Y)\end{aligned}$$

as we will see hold in fair generality, see Example 5.19 below.

Example 5.14 (Durrett Example 4.6.2, p. 205). Suppose we want to determine the expected value of $Y =$ the number of rolls it takes to complete a game of craps. (In this game, if the sum of the dice is $2, 3$, or 12 on his first roll, he loses; if the sum is 7 or 11, he wins; if the sum is $4, 5, 6, 8, 9$, or 10, this number becomes his "point" and he wins if he "makes his point," i.e., his number comes up again before he throws a 7.)

Let X be the sum we obtain on the first roll. If $X = 2, 3, 7, 11, 12$, then the outcome is determined by the first roll so in these cases $\mathbb{E}(Y|X=x) = 1$. If $X = 4$ then the game is completed when a 4 or a 7 appears. So we are waiting for an event with probability $9/36$ and the formula for the mean of the geometric tells us that the expected number of rolls is $36/9 = 4$. Adding the first roll we have $\mathbb{E}[Y|X=4] = 45/9 = 5$. Similar calculations give us

$x \in$	$\{2,3,7,11,12\}$	$\{4,10\}$	$\{5,9\}$	$\{6,8\}$
$\mathbb{E}[Y\|X=x]$	1	$\frac{45}{9}$	$\frac{46}{10}$	$\frac{47}{11}$
probability	$\frac{12}{36}$	$\frac{6}{36}$	$\frac{8}{36}$	$\frac{10}{36}$

.

(For example, there are 5 ways to get a 6 and 6 ways to get a 7 so when $X = 6$ we are waiting for an event with probability $11/36$ and the mean of this geometric random variables is $36/11$ and adding the first roll to this implies, $\mathbb{E}[Y|X=6] = 47/11$. Similarly for $x = 8$ and $P(X = 6 \text{ or } 8) = (5+5)/36$.) Putting the pieces together and using the law of total expectation gives,

5.2 General Properties of Conditional Expectation

Let us pause for a moment to record a few basic general properties of conditional expectations.

Proposition 5.15 (Contraction Property). *For all $Y \in L^2(P)$, we have $\mathbb{E}|\mathbb{E}[Y|X]| \le \mathbb{E}|Y|$. Moreover if $Y \ge 0$ then $\mathbb{E}[Y|X] \ge 0$ (a.s.).*

Proof. Let $\mathbb{E}[Y|X] = h(X)$ (with $h : S \to \mathbb{R}$) and then define

$$f(x) = \begin{cases} 1 & \text{if } h(x) \ge 0 \\ -1 & \text{if } h(x) < 0 \end{cases}.$$

Since $h(x) f(x) = |h(x)|$, it follows from Eq. (5.3) that

$$\mathbb{E}[|h(X)|] = \mathbb{E}[Y f(X)] = |\mathbb{E}[Y f(X)]| \le \mathbb{E}[|Y f(X)|] = \mathbb{E}|Y|.$$

For the second assertion take $f(x) = 1_{h(x)<0}$ in Eq. (5.3) in order to learn

$$\mathbb{E}\left[h(X) 1_{h(X)<0}\right] = \mathbb{E}\left[Y 1_{h(X)<0}\right] \ge 0.$$

As $h(X) 1_{h(X)<0} \le 0$ we may conclude that $h(X) 1_{h(X)<0} = 0$ a.s. ∎

Because of this proposition we may extend the notion of conditional expectation to $Y \in L^1(P)$ as stated in the following theorem which we do not bother to prove here.

Theorem 5.16. *Given $X : \Omega \to S$ and $Y \in L^1(P)$, there exists an "essentially unique" function $h : S \to \mathbb{R}$ such that Eq. (5.3) holds for all bounded functions, $f : S \to \mathbb{R}$. (As above we write $\mathbb{E}[Y|X]$ for $h(X)$.) Moreover the contraction property, $\mathbb{E}|\mathbb{E}[Y|X]| \le \mathbb{E}|Y|$, still holds.*

Theorem 5.17 (Basic properties). *Let Y, Y_1, and Y_2 be integrable random variables and $X : \Omega \to S$ be given. Then:*

1. $\mathbb{E}(Y_1 + Y_2|X) = \mathbb{E}(Y_1|X) + \mathbb{E}(Y_2|X)$.
2. $\mathbb{E}(aY|X) = a\mathbb{E}(Y|X)$ *for all constants* a.
3. $\mathbb{E}(g(X)Y|X) = g(X)\mathbb{E}(Y|X)$ *for all bounded functions* g.
4. $\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}Y$. ***(Law of total expectation.)***
5. *If Y and X are independent then* $\mathbb{E}(Y|X) = \mathbb{E}Y$.

Proof. 1. Let $h_i(X) = \mathbb{E}[Y_i|X]$, then for all bounded f,

$$\begin{aligned} \mathbb{E}[Y_1 f(X)] &= \mathbb{E}[h_1(X) f(X)] \text{ and} \\ \mathbb{E}[Y_2 f(X)] &= \mathbb{E}[h_2(X) f(X)] \end{aligned}$$

and therefore adding these two equations together implies

$$\begin{aligned} \mathbb{E}[(Y_1 + Y_2) f(X)] &= \mathbb{E}[(h_1(X) + h_2(X)) f(X)] \\ &= \mathbb{E}[(h_1 + h_2)(X) f(X)] \\ \mathbb{E}[Y_2 f(X)] &= \mathbb{E}[h_2(X) f(X)] \end{aligned}$$

for all bounded f. Therefore we may conclude that

$$\mathbb{E}(Y_1 + Y_2|X) = (h_1 + h_2)(X) = h_1(X) + h_2(X) = \mathbb{E}(Y_1|X) + \mathbb{E}(Y_2|X).$$

2. The proof is similar to 1 but easier and so is omitted.

3. Let $h(X) = \mathbb{E}[Y|X]$, then $\mathbb{E}[Y f(X)] = \mathbb{E}[h(X) f(X)]$ for all bounded functions f. Replacing f by $g \cdot f$ implies

$$\mathbb{E}[Y g(X) f(X)] = \mathbb{E}[h(X) g(X) f(X)] = \mathbb{E}[(h \cdot g)(X) f(X)]$$

for all bounded functions f. Therefore we may conclude that

$$\mathbb{E}[Y g(X)|X] = (h \cdot g)(X) = h(X) g(X) = g(X)\mathbb{E}(Y|X).$$

4. Take $f \equiv 1$ in Eq. (5.3).

5. If X and Y are independent and $\mu := \mathbb{E}[Y]$, then

$$\mathbb{E}[Y f(X)] = \mathbb{E}[Y]\mathbb{E}[f(X)] = \mu\mathbb{E}[f(X)] = \mathbb{E}[\mu f(X)]$$

from which it follows that $\mathbb{E}[Y|X] = \mu$ as desired. ∎

The next theorem says that conditional expectations essentially only depends on the distribution of (X, Y) and nothing else.

Theorem 5.18. *Suppose that (X, Y) and $\left(\tilde{X}, \tilde{Y}\right)$ are random vectors such that $(X, Y) \overset{d}{=} \left(\tilde{X}, \tilde{Y}\right)$, i.e. $\mathbb{E}[f(X, Y)] = \mathbb{E}\left[f\left(\tilde{X}, \tilde{Y}\right)\right]$ for all bounded (or non-negative) functions f. If $h(X) = \mathbb{E}[u(X, Y)|X]$, then $\mathbb{E}\left[u\left(\tilde{X}, \tilde{Y}\right)|\tilde{X}\right] = h\left(\tilde{X}\right)$.*

Proof. By assumption we know that

$$\mathbb{E}[u(X, Y) f(X)] = \mathbb{E}[h(X) f(X)] \text{ for all bounded } f.$$

Since $(X, Y) \overset{d}{=} \left(\tilde{X}, \tilde{Y}\right)$, this is equivalent to

$$\mathbb{E}\left[u\left(\tilde{X}, \tilde{Y}\right) f\left(\tilde{X}\right)\right] = \mathbb{E}\left[h\left(\tilde{X}\right) f\left(\tilde{X}\right)\right] \text{ for all bounded } f$$

which is equivalent to $\mathbb{E}\left[u\left(\tilde{X}, \tilde{Y}\right)|\tilde{X}\right] = h\left(\tilde{X}\right)$. ∎

Example 5.19. Let $\{X_i\}_{i=1}^{\infty}$ be i.i.d. random variables with $\mathbb{E}|X_i| < \infty$ for all i and let $S_m := X_1 + \cdots + X_m$ for $m = 1, 2, \ldots$. We wish to show,

$$\mathbb{E}[S_m|S_n] = \frac{m}{n} S_n \text{ for all } m \le n.$$

for all $m \le n$. To prove this first observe by symmetry[1] that

$$\mathbb{E}(X_i|S_n) = h(S_n) \text{ independent of } i.$$

Therefore

$$S_n = \mathbb{E}(S_n|S_n) = \sum_{i=1}^{n} \mathbb{E}(X_i|S_n) = \sum_{i=1}^{n} h(S_n) = n \cdot h(S_n).$$

Thus we see that

$$\mathbb{E}(X_i|S_n) = \frac{1}{n} S_n$$

and therefore

$$\mathbb{E}(S_m|S_n) = \sum_{i=1}^{m} \mathbb{E}(X_i|S_n) = \sum_{i=1}^{m} \frac{1}{n} S_n = \frac{m}{n} S_n.$$

If $m > n$, then $S_m = S_n + X_{n+1} + \cdots + X_m$. Since X_i is independent of S_n for $i > n$, it follows that

$$\begin{aligned}\mathbb{E}(S_m|S_n) &= \mathbb{E}(S_n + X_{n+1} + \cdots + X_m|S_n) \\ &= \mathbb{E}(S_n|S_n) + \mathbb{E}(X_{n+1}|S_n) + \cdots + \mathbb{E}(X_m|S_n) \\ &= S_n + (m-n)\mu \text{ if } m \ge n\end{aligned}$$

where $\mu = \mathbb{E}X_i$.

Example 5.20 (See Durrett, #8, p. 213). Suppose that X and Y are two integrable random variables such that

$$\mathbb{E}[X|Y] = 18 - \frac{3}{5}Y \text{ and } \mathbb{E}[Y|X] = 10 - \frac{1}{3}X.$$

We would like to find $\mathbb{E}X$ and $\mathbb{E}Y$. To do this we use the law of total expectation to find,

$$\mathbb{E}X = \mathbb{E}\mathbb{E}[X|Y] = \mathbb{E}\left(18 - \frac{3}{5}Y\right) = 18 - \frac{3}{5}\mathbb{E}Y \text{ and}$$

$$\mathbb{E}Y = \mathbb{E}\mathbb{E}[Y|X] = \mathbb{E}\left(10 - \frac{1}{3}X\right) = 10 - \frac{1}{3}\mathbb{E}X.$$

Solving this pair or linear equations shows $\mathbb{E}X = 15$ and $\mathbb{E}Y = 5$.

[1] Apply Theorem 5.18 using $(X_1, S_n) \overset{d}{=} (X_i, S_n)$ for $1 \le i \le n$.

5.3 Conditional Expectation for Continuous Random Variables

(This section will be covered later in the course when first needed.)

Suppose that Y and X are continuous random variables which have a joint density, $\rho_{(Y,X)}(y,x)$. Then by definition of $\rho_{(Y,X)}$, we have, for all bounded or non-negative, U, that

$$\mathbb{E}[U(Y,X)] = \int\int U(y,x)\,\rho_{(Y,X)}(y,x)\,dydx. \tag{5.4}$$

The marginal density associated to X is then given by

$$\rho_X(x) := \int \rho_{(Y,X)}(y,x)\,dy \tag{5.5}$$

and recall from Math 180A that the conditional density $\rho_{(Y|X)}(y,x)$ is defined by

$$\rho_{(Y|X)}(y,x) = \begin{cases} \frac{\rho_{(Y,X)}(y,x)}{\rho_X(x)} & \text{if } \rho_X(x) > 0 \\ 0 & \text{if } \rho_X(x) = 0 \end{cases}.$$

Observe that if $\rho_{(Y,X)}(y,x)$ is continuous, then

$$\rho_{(Y,X)}(y,x) = \rho_{(Y|X)}(y,x)\,\rho_X(x) \text{ for all } (x,y). \tag{5.6}$$

Indeed, if $\rho_X(x) = 0$, then

$$0 = \rho_X(x) = \int \rho_{(Y,X)}(y,x)\,dy$$

from which it follows that $\rho_{(Y,X)}(y,x) = 0$ for all y. If $\rho_{(Y,X)}$ is not continuous, Eq. (5.6) still holds for "a.e." (x,y) which is good enough.

Lemma 5.21. *In the notation above,*

$$\rho(x,y) = \rho_{(Y|X)}(y,x)\,\rho_X(x) \text{ for a.e. } (x,y). \tag{5.7}$$

Proof. By definition Eq. (5.7) holds when $\rho_X(x) > 0$ and $\rho(x,y) \ge \rho_{(Y|X)}(y,x)\,\rho_X(x)$ for all (x,y). Moreover,

$$\begin{aligned}\int\int \rho_{(Y|X)}(y,x)\,\rho_X(x)\,dxdy &= \int\int \rho_{(Y|X)}(y,x)\,\rho_X(x)\,1_{\rho_X(x)>0}dxdy \\ &= \int\int \rho(x,y)\,1_{\rho_X(x)>0}dxdy \\ &= \int \rho_X(x)\,1_{\rho_X(x)>0}dx = \int \rho_X(x)\,dx \\ &= 1 = \int\int \rho(x,y)\,dxdy,\end{aligned}$$

or equivalently,

$$\int\int \left[\rho(x,y) - \rho_{(Y|X)}(y,x)\,\rho_X(x)\right] dxdy = 0$$

which implies the result. ∎

Theorem 5.22. *Keeping the notation above,for all or all bounded or non-negative, U, we have $\mathbb{E}[U(Y,X)|X] = h(X)$ where*

$$h(x) = \int U(y,x)\,\rho_{(Y|X)}(y,x)\,dy \tag{5.8}$$

$$= \begin{cases} \frac{\int U(y,x)\rho_{(Y,X)}(y,x)dy}{\int \rho_{(Y,X)}(y,x)dy} & \text{if} \quad \int \rho_{(Y,X)}(y,x)\,dy > 0 \\ 0 & \text{otherwise} \end{cases}. \tag{5.9}$$

In the future we will usually denote $h(x)$ informally by $\mathbb{E}[U(Y,x)|X=x]$,[2] so that

$$\mathbb{E}[U(Y,x)|X=x] := \int U(y,x)\,\rho_{(Y|X)}(y,x)\,dy. \tag{5.10}$$

Proof. We are looking for $h: S \to \mathbb{R}$ such that

$$\mathbb{E}[U(Y,X)\,f(X)] = \mathbb{E}[h(X)\,f(X)] \text{ for all bounded } f.$$

Using Lemma 5.21, we find

$$\begin{aligned}
\mathbb{E}[U(Y,X)\,f(X)] &= \int\int U(y,x)\,f(x)\,\rho_{(Y,X)}(y,x)\,dydx \\
&= \int\int U(y,x)\,f(x)\,\rho_{(Y|X)}(y,x)\,\rho_X(x)\,dydx \\
&= \int \left[\int U(y,x)\,\rho_{(Y|X)}(y,x)\,dy\right] f(x)\,\rho_X(x)\,dx \\
&= \int h(x)\,f(x)\,\rho_X(x)\,dx \\
&= \mathbb{E}[h(X)\,f(X)]
\end{aligned}$$

where h is given as in Eq. (5.8). ∎

Example 5.23 (Durrett 8.15, p. 145). Suppose that X and Y have joint density $\rho(x,y) = 8xy \cdot 1_{0<y<x<1}$. We wish to compute $\mathbb{E}[u(X,Y)|Y]$. To this end we compute

$$\rho_Y(y) = \int_{\mathbb{R}} 8xy \cdot 1_{0<y<x<1}dx = 8y\int_{x=y}^{x=1} x\cdot dx = 8y\cdot\frac{x^2}{2}|_y^1 = 4y\cdot\left(1-y^2\right).$$

Therefore,

$$\rho_{X|Y}(x,y) = \frac{\rho(x,y)}{\rho_Y(y)} = \frac{8xy\cdot 1_{0<y<x<1}}{4y\cdot(1-y^2)} = \frac{2x\cdot 1_{0<y<x<1}}{(1-y^2)}$$

and so

$$\mathbb{E}[u(X,Y)|Y=y] = \int_{\mathbb{R}} \frac{2x\cdot 1_{0<y<x<1}}{(1-y^2)}u(x,y)\,dx = 2\frac{1_{0<y<1}}{1-y^2}\int_y^1 u(x,y)\;xdx$$

and so

$$\mathbb{E}[u(X,Y)|Y] = 2\frac{1}{1-Y^2}\int_Y^1 u(x,Y)\,xdx.$$

is the best approximation to $u(X,Y)$ be a function of Y alone.

Proposition 5.24. *Suppose that X, Y are independent random functions, then*

$$\mathbb{E}[U(Y,X)|X] = h(X)$$

where

$$h(x) := \mathbb{E}[U(Y,x)].$$

Proof. I will prove this in the continuous distribution case and leave the discrete case to the reader. (The theorem is true in general but requires measure theory in order to prove it in full generality.) The independence assumption is equivalent to $\rho_{(Y,X)}(y,x) = \rho_Y(y)\,\rho_X(x)$. Therefore,

$$\rho_{(Y|X)}(y,x) = \begin{cases} \rho_Y(y) \text{ if } \rho_X(x) > 0 \\ 0 \quad \text{if } \rho_X(x) = 0 \end{cases}$$

and therefore $\mathbb{E}[U(Y,X)|X] = h_0(X)$ where

$$\begin{aligned}
h_0(x) &= \int U(y,x)\,\rho_{(Y|X)}(y,x)\,dy \\
&= 1_{\rho_X(x)>0}\int U(y,x)\,\rho_Y(y)\,dy = 1_{\rho_X(x)>0}\mathbb{E}[U(Y,x)] \\
&= 1_{\rho_X(x)>0}h(x).
\end{aligned}$$

If f is a bounded function of x, then

$$\begin{aligned}
\mathbb{E}[h_0(X)\,f(X)] &= \int h_0(x)\,f(x)\,\rho_X(x)\,dx = \int_{\{x:\rho_X(x)>0\}} h_0(x)\,f(x)\,\rho_X(x)\,dx \\
&= \int_{\{x:\rho_X(x)>0\}} h(x)\,f(x)\,\rho_X(x)\,dx = \int h(x)\,f(x)\,\rho_X(x)\,dx \\
&= \mathbb{E}[h(X)\,f(X)].
\end{aligned}$$

[2] **Warning:** this is **not** consistent with Eq. (5.1) as $P(X=x) = 0$ for continuous distributions.

So for all practical purposes, $h(X) = h_0(X)$, i.e. $h(X) = h_0(X)$ – a.s. (Indeed, take $f(x) = \operatorname{sgn}(h(x) - h_0(x))$ in the above equation to learn that $\mathbb{E}|h(X) - h_0(X)| = 0$. ∎

5.4 Conditional Variances

Definition 5.25 (Conditional Variance). *Suppose that $Y \in L^2(P)$ and $X : \Omega \to S$ are given. We define*

$$\operatorname{Var}(Y|X) = \mathbb{E}\left[Y^2|X\right] - (\mathbb{E}[Y|X])^2 \tag{5.11}$$

$$= \mathbb{E}\left[(Y - \mathbb{E}[Y|X])^2 |X\right] \tag{5.12}$$

*to be the **conditional variance of** Y **given** X.*

Theorem 5.26. *Suppose that $Y \in L^2(P)$ and $X : \Omega \to S$ are given, then*

$$\operatorname{Var}(Y) = \mathbb{E}[\operatorname{Var}(Y|X)] + \operatorname{Var}(\mathbb{E}[Y|X]).$$

Proof. Taking expectations of Eq. (5.11) implies,

$$\begin{aligned}\mathbb{E}[\operatorname{Var}(Y|X)] &= \mathbb{E}\mathbb{E}\left[Y^2|X\right] - \mathbb{E}(\mathbb{E}[Y|X])^2 \\ &= \mathbb{E}Y^2 - \mathbb{E}(\mathbb{E}[Y|X])^2 = \operatorname{Var}(Y) + (\mathbb{E}Y)^2 - \mathbb{E}(\mathbb{E}[Y|X])^2.\end{aligned}$$

The result follows from this identity and the fact that

$$\operatorname{Var}(\mathbb{E}[Y|X]) = \mathbb{E}(\mathbb{E}[Y|X])^2 - (\mathbb{E}\mathbb{E}[Y|X])^2 = \mathbb{E}(\mathbb{E}[Y|X])^2 - (\mathbb{E}Y)^2.$$

∎

5.5 Random Sums

Suppose that $\{X_i\}_{i=1}^\infty$ is a collection of random variables and let

$$S_n := \begin{cases} X_1 + \cdots + X_n & \text{if } n \geq 1 \\ 0 & \text{if } n = 0 \end{cases}.$$

Given a $\mathbb{Z}_+$ – valued random variable, N, we wish to consider the random sum;

$$S_N = X_1 + \cdots + X_N.$$

We are now going to suppose for the rest of this subsection that N is independent of $\{X_i\}_{i=1}^\infty$ and for $f \geq 0$ we let

$$Tf(n) := \mathbb{E}[f(S_n)] \text{ for all } n \in \mathbb{N}_0.$$

Theorem 5.27. *Suppose that N is independent of $\{X_i\}_{i=1}^\infty$ as above. Then for any positive function f, we have,*

$$\mathbb{E}[f(S_N)] = \mathbb{E}[Tf(N)].$$

Moreover this formula holds for any f such that

$$\mathbb{E}[|f(S_N)|] = \mathbb{E}[T|f|(N)] < \infty.$$

Proof. If $f \geq 0$ we have,

$$\begin{aligned}\mathbb{E}[f(S_N)] &= \sum_{n=0}^\infty \mathbb{E}[f(S_N) : S_N = n] = \sum_{n=0}^\infty \mathbb{E}[f(S_n) : S_N = n] \\ &= \sum_{n=0}^\infty \mathbb{E}[f(S_n)] P(S_N = n) = \sum_{n=0}^\infty (Tf)(n) P(S_N = n) \\ &= \mathbb{E}[Tf(N)].\end{aligned}$$

The moreover part follows from general non-sense not really covered in this course. ∎

Theorem 5.28. *Suppose that $\{X_i\}_{i=1}^\infty$ are uncorrelated $L^2(P)$ – random variables with $\mu = \mathbb{E}X_i$ and $\sigma^2 = \operatorname{Var}(X_i)$ independent of i. Assuming that $N \in L^2(P)$ is independent of the $\{X_i\}$, then*

$$\mathbb{E}[S_N] = \mu \cdot \mathbb{E}N \tag{5.13}$$

and

$$\operatorname{Var}(S_N) = \sigma^2 \mathbb{E}[N] + \mu^2 \operatorname{Var}(N). \tag{5.14}$$

Proof. Taking $f(x) = x$ in Theorem 5.27 using $Tf(n) = \mathbb{E}[S_n] = n \cdot \mu$ we find,

$$\mathbb{E}[S_N] = \mathbb{E}[\mu \cdot N] = \mu \cdot \mathbb{E}N$$

as claimed. Next take $f(x) = x^2$ in Theorem 5.27 using

$$Tf(n) = \mathbb{E}\left[S_n^2\right] = \operatorname{Var}(S_n) + (\mathbb{E}S_n)^2 = \sigma^2 n + (n \cdot \mu)^2,$$

we find that

$$\begin{aligned}\mathbb{E}\left[S_N^2\right] &= \mathbb{E}\left[\sigma^2 N + \mu^2 N^2\right] \\ &= \sigma^2 \mathbb{E}[N] + \mu^2 \mathbb{E}\left[N^2\right].\end{aligned}$$

Combining these results shows,

$$\begin{aligned}\operatorname{Var}(S_N) &= \sigma^2 \mathbb{E}[N] + \mu^2 \mathbb{E}\left[N^2\right] - \mu^2 (\mathbb{E}N)^2 \\ &= \sigma^2 \mathbb{E}[N] + \mu^2 \operatorname{Var}(N).\end{aligned}$$

∎

Example 5.29 (Karlin and Taylor E.3.1. p77). A six-sided die is rolled, and the number N on the uppermost face is recorded. Then a fair coin is tossed N times, and the total number Z of heads to appear is observed. Determine the mean and variance of Z by viewing Z as a random sum of N Bernoulli random variables. Determine the probability mass function of Z, and use it to find the mean and variance of Z.

We have $Z = S_N = X_1 + \cdots + X_N$ where $X_i = 1$ if heads on the ith toss and zero otherwise. In this case

$$\mathbb{E}X_1 = \frac{1}{2},$$
$$\operatorname{Var}(X_1) = \frac{1}{2} - \left(\frac{1}{2}\right)^2 = \frac{1}{4},$$
$$\mathbb{E}N = \frac{1}{6}(1 + \cdots + 6) = \frac{1}{6}\frac{7 \cdot 6}{2} = \frac{7}{2},$$
$$\mathbb{E}N^2 = \frac{1}{6}\left(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2\right) = \frac{91}{6}$$
$$\operatorname{Var}(N) = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}.$$

Therefore,

$$\mathbb{E}Z = \mathbb{E}X_1 \cdot \mathbb{E}N = \frac{1}{2} \cdot \frac{7}{2} = \frac{7}{4}$$
$$\operatorname{Var}(Z) = \frac{1}{4} \cdot \frac{7}{2} + \left(\frac{1}{2}\right)^2 \cdot \frac{35}{12} = \frac{77}{48} = 1.604\,2.$$

Alternatively, we have

$$\begin{aligned} P(Z = k) &= \sum_{n=1}^{6} P(Z = k|N = n)\, P(N = n) \\ &= \frac{1}{6} \sum_{n=k\vee 1}^{6} P(Z = k|N = n) \\ &= \frac{1}{6} \sum_{n=k\vee 1}^{6} \binom{n}{k} \left(\frac{1}{2}\right)^n. \end{aligned}$$

where

$$\begin{aligned} \mathbb{E}Z &= \sum_{k=0}^{6} kP(Z = k) = \sum_{k=1}^{6} kP(Z = k) \\ &= \sum_{k=1}^{6} k\frac{1}{6} \sum_{n=k}^{6} \binom{n}{k} \left(\frac{1}{2}\right)^n = \frac{7}{4} \end{aligned}$$

and

$$\mathbb{E}Z^2 = \sum_{k=0}^{6} k^2 P(Z = k) = \sum_{k=1}^{6} k^2 \frac{1}{6} \sum_{n=k}^{6} \binom{n}{k} \left(\frac{1}{2}\right)^n = \frac{14}{3}$$

so that

$$\operatorname{Var}(Z) = \frac{14}{3} - \left(\frac{7}{4}\right)^2 = \frac{77}{48}.$$

We have,

$$\begin{aligned} P(Z = 0) &= \frac{1}{6} \sum_{n=1}^{6} \binom{n}{0} \left(\frac{1}{2}\right)^n = \frac{21}{128} \\ P(Z = 1) &= \frac{1}{6} \sum_{n=1}^{6} \binom{n}{1} \left(\frac{1}{2}\right)^n = \frac{5}{16} \\ P(Z = 2) &= \frac{1}{6} \sum_{n=2}^{6} \binom{n}{2} \left(\frac{1}{2}\right)^n = \frac{33}{128} \\ P(Z = 3) &= \frac{1}{6} \sum_{n=3}^{6} \binom{n}{3} \left(\frac{1}{2}\right)^n = \frac{1}{6} \\ P(Z = 4) &= \frac{1}{6} \sum_{n=4}^{6} \binom{n}{4} \left(\frac{1}{2}\right)^n = \frac{29}{384} \\ P(Z = 5) &= \frac{1}{6} \sum_{n=5}^{6} \binom{n}{5} \left(\frac{1}{2}\right)^n = \frac{1}{48} \\ P(Z = 6) &= \frac{1}{6} \sum_{n=6}^{6} \binom{n}{6} \left(\frac{1}{2}\right)^n = \frac{1}{384}. \end{aligned}$$

Remark 5.30. If the $\{X_i\}$ are i.i.d., we may work out the moment generating function, $mgf_{S_N}(t) := \mathbb{E}\left[e^{tS_N}\right]$ as follows. Conditioning on $N = n$ shows,

$$\begin{aligned} \mathbb{E}\left[e^{tS_N}|N = n\right] &= \mathbb{E}\left[e^{tS_n}|N = n\right] = \mathbb{E}\left[e^{tS_n}\right] \\ &= \left[\mathbb{E}e^{tX_1}\right]^n = \left[mgf_{X_1}(t)\right]^n \end{aligned}$$

so that

$$\mathbb{E}\left[e^{tS_N}|N\right] = \left[mgf_{X_1}(t)\right]^N = e^{N\ln\left(mgf_{X_1}(t)\right)}.$$

Taking expectations of this equation using the law of total expectation gives,

$$mgf_{S_N}(t) = mgf_N\left(\ln\left(mgf_{X_1}(t)\right)\right).$$

5.6 Summary on Conditional Expectation Properties

Let Y and X be random variables such that $\mathbb{E}Y^2 < \infty$ and h be function from the range of X to $\mathbb{R}$. Then the following are equivalent:

1. $h(X) = \mathbb{E}(Y|X)$, i.e. $h(X)$ is the conditional expectation of Y given X.
2. $\mathbb{E}(Y - h(X))^2 \leq \mathbb{E}(Y - g(X))^2$ for all functions g, i.e. $h(X)$ is the best approximation to Y among functions of X.
3. $\mathbb{E}(Y \cdot g(X)) = \mathbb{E}(h(X) \cdot g(X))$ for all functions g, i.e. $Y - h(X)$ is orthogonal to all functions of X. Moreover, this condition uniquely determines $h(X)$.

The methods for computing $\mathbb{E}(Y|X)$ are given in the next two propositions.

Proposition 5.31 (Discrete Case). *Suppose that Y and X are discrete random variables and $p(y,x) := P(Y = y, X = x)$. Then $\mathbb{E}(Y|X) = h(X)$, where*

$$h(x) = \mathbb{E}(Y|X = x) = \frac{\mathbb{E}(Y : X = x)}{P(X = x)} = \frac{1}{p_X(x)} \sum_y y p(y,x) \tag{5.15}$$

and $p_X(x) = P(X = x)$ is the marginal distribution of X which may be computed as $p_X(x) = \sum_y p(y,x)$.

Proposition 5.32 (Continuous Case). *Suppose that Y and X are random variables which have a joint probability density $\rho(y,x)$ (i.e. $P(Y \in dy, X \in dx) = \rho(y,x)dydx$). Then $\mathbb{E}(Y|X) = h(X)$, where*

$$h(x) = \mathbb{E}(Y|X = x) := \frac{1}{\rho_X(x)} \int_{-\infty}^{\infty} y\rho(y,x)dy \tag{5.16}$$

and $\rho_X(x)$ is the marginal density of X which may be computed as

$$\rho_X(x) = \int_{-\infty}^{\infty} \rho(y,x)dy.$$

Intuitively, in all cases, $\mathbb{E}(Y|X)$ on the set $\{X = x\}$ is $\mathbb{E}(Y|X = x)$. This intuitions should help motivate some of the basic properties of $\mathbb{E}(Y|X)$ summarized in the next theorem.

Theorem 5.33. *Let Y, Y_1, Y_2 and X be random variables. Then:*

1. $\mathbb{E}(Y_1 + Y_2|X) = \mathbb{E}(Y_1|X) + \mathbb{E}(Y_2|X)$.
2. $\mathbb{E}(aY|X) = a\mathbb{E}(Y|X)$ *for all constants* a.
3. $\mathbb{E}(f(X)Y|X) = f(X)\mathbb{E}(Y|X)$ *for all functions* f.
4. $\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}Y$.
5. *If Y and X are independent then* $\mathbb{E}(Y|X) = \mathbb{E}Y$.
6. *If* $Y \geq 0$ *then* $\mathbb{E}(Y|X) \geq 0$.

Remark 5.34. Property 4 in Theorem 5.33 turns out to be a very powerful method for computing expectations. I will finish this summary by writing out Property 4 in the discrete and continuous cases:

$$\mathbb{E}Y = \sum_x \mathbb{E}(Y|X = x)p_X(x) \qquad (\textbf{Discrete Case})$$

where $\mathbb{E}(Y|X = x)$ is **defined** in Eq. (5.15)

$$\mathbb{E}U(Y,X) = \int \mathbb{E}(U(Y,X)|X = x)\rho_X(x)dx, \qquad (\textbf{Continuous Case})$$

where $\mathbb{E}(U(Y,X)|X = x)$ is now defined as in Eq. (5.10).

References