Lecture 14: Order Statistics; Conditional Expectation

Relevant textbook passages:

Pitman [5]: Section 4.6

Larsen–Marx [2]: Section 3.11

14.1 Order statistics

Pitman [5]: § 4.6

Given a random vector $(X_1, \dots, X_n)$ on the probability space $(S, \mathcal{E}, P)$, for each $s \in S$, sort the components into a vector $\big(X_{(1)}(s), \dots, X_{(n)}(s)\big)$ satisfying

$$X_{(1)}(s) \leqslant X_{(2)}(s) \leqslant \cdots \leqslant X_{(n)}(s).$$

The vector $(X_{(1)}, \dots, X_{(n)})$ is called the vector of **order statistics** of $(X_1, \dots, X_n)$.

Equivalently,

$$X_{(k)} = \min\Big\{\max\{X_j : j \in J\} \,:\, J \subset \{1, \dots, n\} \;\&\; |J| = k\Big\}.$$

Order statistics play an important role in the study of auctions, among other things.

14.2 Marginal Distribution of Order Statistics

From here on, we shall assume that the original random variables $X_1, \dots, X_n$ are independent and identically distributed.

Let $X_1, \dots, X_n$ be independent and identically distributed random variables with common cumulative distribution function F, and let $(X_{(1)}, \dots, X_{(n)})$ be the vector of order statistics of $X_1, \dots, X_n$. By breaking the event $\big(X_{(k)} \leqslant x\big)$ into simple disjoint subevents, we get

$$\begin{aligned}
\big(X_{(k)} \leqslant x\big) = {} & \\
& \big(X_{(n)} \leqslant x\big) \\
& \bigcup \big(X_{(n)} > x,\ X_{(n-1)} \leqslant x\big) \\
& \qquad \vdots \\
& \bigcup \big(X_{(n)} > x, \dots, X_{(j+1)} > x, X_{(j)} \leqslant x\big) \\
& \qquad \vdots \\
& \bigcup \big(X_{(n)} > x, \quad \dots, \quad X_{(k+1)} > x, X_{(k)} \leqslant x\big)\,.
\end{aligned}$$

Each of these subevents is disjoint from the ones above it, and each has a binomial probability:

$$\big(X_{(n)} > x, \dots, X_{(j+1)} > x, X_{(j)} \leqslant x\big) = (n-j \text{ of the random variables are } > x \text{ and } j \text{ are } \leqslant x),$$

so

$$P\left(X_{(n)} > x, \dots, X_{(j+1)} > x, X_{(j)} \leqslant x\right) = \binom{n}{j}\big(1 - F(x)\big)^{n-j} F(x)^j.$$

Thus:

The cdf of the k^{th} order statistic from a sample of n is:

$$F_{(k,n)}(x) = P\left(X_{(k)} \leqslant x\right) = \sum_{j=k}^{n} \binom{n}{j}\big(1 - F(x)\big)^{n-j} F(x)^j. \tag{1}$$

14.3 Marginal Density of Order Statistics

Pitman [5]: p. 326

If F has a density $f = F'$, then we can calculate the marginal density of $X_{(k)}$ by differentiating the CDF $F_{(k,n)}$:

$$\begin{aligned}
&\frac{d}{dx}F_{(k,n)}(x) \\
&= \frac{d}{dx}\sum_{j=k}^{n}\binom{n}{j}F(x)^j\big(1-F(x)\big)^{n-j} \\
&= \sum_{j=k}^{n}\binom{n}{j}\frac{d}{dx}F(x)^j\big(1-F(x)\big)^{n-j} \\
&= \sum_{j=k}^{n}\binom{n}{j}\Big(jF(x)^{j-1}\big(1-F(x)\big)^{n-j}F'(x) - (n-j)F(x)^j\big(1-F(x)\big)^{n-j-1}F'(x)\Big) \\
&= \sum_{j=k}^{n}\binom{n}{j}\Big(jF(x)^{j-1}\big(1-F(x)\big)^{n-j} - (n-j)F(x)^j\big(1-F(x)\big)^{n-j-1}\Big)f(x) \\
&= \sum_{j=k}^{n}\frac{n!}{(j-1)!(n-j)!}F(x)^{j-1}\big(1-F(x)\big)^{n-j}f(x) \\
&\quad - \sum_{j=k}^{n-1}\frac{n!}{j!(n-j-1)!}\big(1-F(x)\big)^{n-j-1}F(x)^j f(x) \\
&= \frac{n!}{(k-1)!(n-k)!}\big(1-F(x)\big)^{n-j}F(x)^{k-1}f(x) \\
&\quad + \sum_{j=k+1}^{n}\frac{n!}{(j-1)!(n-j)!}\big(1-F(x)\big)^{n-j}F(x)^{j-1}f(x) \\
&\quad - \sum_{j=k}^{n-1}\frac{n!}{j!(n-j-1)!}\big(1-F(x)\big)^{n-j-1}F(x)^j f(t).
\end{aligned}$$

The last two terms above cancel, since using the change of variables $i = j - 1$,

$$\sum_{j=k+1}^{n} \frac{n!}{(j-1)!(n-j)!}\big(1-F(x)\big)^{n-j}F(x)^{j-1} = \sum_{i=k}^{n-1} \frac{n!}{i!(n-i-1)!}\big(1-F(x)\big)^{n-i}F(x)^{i}.$$

So the density of the k^{th} order statistic from a sample of size n is:

$$f_{(k,n)}(x) = \frac{n!}{(k-1)!(n-k)!}\big(1-F(x)\big)^{n-k}F(x)^{k-1}f(x) = n\binom{n-1}{k-1}\big(1-F(x)\big)^{(n-1)-(k-1)}F(x)^{k-1}f(x). \quad (2)$$

Note: I could have written $n-k$ instead of $(n-1)-(k-1)$, but then the binomial coefficient would have seemed more mysterious.

14.4 Some special order statistics

The 1^{st} order statistic $X_{(1)}$ from a sample of size n is just the minimum

$$X_{(1)} = \min\{X_1, \dots, X_n\}.$$

The event that $\min\{X_1, \dots, X_n\} \leqslant x$ is just the event that *at least one* $X_i \leqslant x$. The complement of this event is that all $X_i > x$, which has probability $\big(1-F(x)\big)^n$, so the cdf is

$$F_{(1,n)}(x) = 1 - \big(1-F(x)\big)^n$$

and its density is

$$f_{(1,n)}(x) = n\big(1-F(x)\big)^{(n-1)}f(x).$$

The n^{th} order statistic $X_{(n)}$ from a sample of size n is just the maximum

$$X_{(n)} = \max\{X_1, \dots, X_n\}.$$

The event that $max\{X_1, \dots, X_n\} \leqslant x$ is just the event that *all* $X_i \leqslant x$, so its cdf is

$$F_{(n,n)}(x) = F(x)^n$$

and its density is

$$f_{(n,n)}(x) = nF(x)^{(n-1)}f(x).$$

The $(n-1)^{\text{st}}$ order statistic is the second-highest value. Its cdf is

$$F_{(n-1,n)}(x) = n\big(1-F(x)\big)F(x)^{n-2} + F(x)^n = nF(x)^{n-1} - (n-1)F(x)^n,$$

and its density is

$$n(n-1)\big(1-F(x)\big)F(x)^{n-2}f(x).$$

In the study of auctions, the second-highest bid is of special interest. Indeed a **second-price auction** awards the item to the highest bidder, but the price is the second-highest bid. A standard auction provides incentives for bidders to bid less than their values, so the distribution of bids and the distribution of values is not the same. Nevertheless it can be shown (see my online notes) that the expected revenue to a seller in an auction with n bidders with independent and identically distributed values is just the expectation of the second-highest order statistic for the distribution of values.

14.5 Uniform order statistics and the Beta function

Pitman [5]: pp. 327–328

For a Uniform[0,1] distribution, $F(t) = t$ and $f(t) = 1$ on $[0, 1]$. In this case equation (2) tells us:

> The density $f_{(k,n)}$ of the k^{th} order statistic for n independent Uniform$[0, 1]$ random variables is
> $$f_{(k,n)}(t) = n\binom{n-1}{k-1}(1-t)^{n-k}t^{k-1}.$$

Since $f_{(k,n)}$ is a density,

$$\int_0^1 f_{(k,n)}(t)\,dt = n\binom{n-1}{k-1}\int_0^1 (1-t)^{n-k}t^{k-1} = 1,$$

or

$$\int_0^1 (1-t)^{n-k}t^{k-1} = \frac{1}{n\binom{n-1}{k-1}} = \frac{(k-1)!(n-k)!}{n!}. \tag{3}$$

Now change variables by setting

$$r = k \qquad \text{and} \qquad s = n - r + 1 \text{ (so } s - 1 = n - r \text{ and } n = r + s - 1).$$

Then rewrite (3) as

$$\int_0^1 (1-t)^{s-1}t^{r-1} = \frac{(r-1)!(s-1)!}{(s+r-1)!} = \frac{\Gamma(s)\Gamma(r)}{\Gamma(r+s)}.$$

Recall that the **Gamma function** is a continuous version of the factorial, and has the property that $\Gamma(s+1) = s\Gamma(s)$ for every $s > 0$, and $\Gamma(m) = (m-1)!$ for every natural number m. See Definition 13.1.1.

This fact suggests (to at least some people) the following definition:

> **14.5.1 Definition** *The **Beta function** is defined for $r, s > 0$ (not necessarily integers), by*
> $$B(r,s) = \int_0^1 t^{r-1}(1-t)^{s-1}\,dt = \frac{\Gamma(s)\Gamma(r)}{\Gamma(r+s)}.$$

So for integers r and s, we see that

$$B(r+1, s+1) = \frac{\Gamma(s+1)\Gamma(r+1)}{\Gamma(r+s+2)} = \frac{(s)!(r)!}{(r+s-1)!} = (r+s)\frac{(s)!(r)!}{(r+s)!} = (r+s)\frac{1}{\binom{r+s}{r}}.$$

> **14.5.2 Definition** *The **beta(r, s) distribution** has the density*
> $$f(x) = \frac{1}{B(r,s)}x^{r-1}(1-x)^{s-1}$$
> *on the interval $[0, 1]$ and zero elsewhere.*

Note that for integer $r$and s, the density of the beta$(r+1, s+1)$ distribution is

$$f(x) = \frac{1}{r+s}\binom{r+s}{r}x^r(1-x)^s,$$

which is $1/(r+s)$ times the Binomial probability of r successes and s failures in $r+s$ trials, where the probability of success is x.

The mean of a beta(r, s) distribution is

$$\frac{r}{r+s}.$$

Proof:

$$\begin{aligned}
\int_0^1 x f(x)\,dx &= \frac{1}{B(r,s)}\int_0^1 x\, x^{r-1}(1-x)^{s-1}\,dx \\
&= \frac{1}{B(r,s)}\int_0^1 x^{r+1-1}(1-x)^{s-1} \\
&= \frac{B(r+1,s)}{B(r,s)} \\
&= \frac{\Gamma(s)\Gamma(r+1)}{\Gamma(r+1+s)}\,\frac{\Gamma(r+s)}{\Gamma(s)\Gamma(r)} \\
&= \frac{\Gamma(s)r\Gamma(r)}{(r+s)\Gamma(r+s)}\,\frac{\Gamma(r+s)}{\Gamma(s)\Gamma(r)} \\
&= \frac{r}{r+s}.
\end{aligned}$$

∎

Thus for a Uniform[0,1] distribution, the (k, n) order statistic has a beta$(k, n-k+1)$ distribution and so has mean

$$\frac{k}{n+1}.$$

[Application to breaking a unit length bar into $n+1$ pieces by choosing n breaking points. The expectation of the k^{th} breaking point is at $k/(n+1)$, so each piece has expected length $1/n$.]

14.6 The war of attrition

In the 1970s, the ethologist John Maynard Smith [3] began to apply game theory to problems of animal behavior. One application was to the settlement of intraspecies conflict. In some species (e.g., peafowl), conflicts are not settled by violent means, but by means of *displays*. The rivals will fan their tails, and eventually one will depart, leaving to the other whatever was the source of the conflict. Maynard Smith modeled this as a "war of attrition."

In the war of attrition game there are two rival contestants $i = 1, 2$ for a prize of value v. Each chooses a length of time t_i at random according to a common probability distribution with cumulative distribution function F. Waiting is costly, and the cost of waiting a length of time t is ct. The rivals continue their displays, until the lesser time elapses and that animal leaves. The distribution is an *symmetric equilibrium distribution* if it has the following properties. (i.)

Each rival, knowing that the opponent has drawn a time t_i from the distribution specified by F, is also willing to choose a time specified by F. (ii.) When the time t_i has elapsed, and contestant i's opponent has not left, then i does not have an incentive to stay longer, and so will leave.

Suppose contestant 2 chooses a waiting time s at random according to an exponential distribution with parameter λ. Now consider contestant 1's decision. Suppose contestant 1 chooses to wait a length of time t. If $s < t$, which happens with probability $1 - e^{-\lambda t}$, hew ins the prize and receives V, but he also incurs a waiting cost cs. If $s > t$, which happens with probability $e^{-\lambda t}$, then his cost is ct and he doe not get the prize. The expected total payoff $\varphi(t)$ to 1 is then

$$\varphi(t) = V(1 - e^{-\lambda t}) - \Big[\int_0^t cse^{-\lambda s}\,ds + cte^{-\lambda t}\Big].$$

Rather than integrate this to find out its value, let's see how it depends on t by computing its derivative. Recalling the Fundamental Theorem of Calculus, we see that

$$\varphi'(t) = v\lambda e^{-\lambda t} - \Big[ct\lambda e^{-\lambda t} + \big(ce^{-\lambda t} - c\lambda te^{-\lambda t}\big)\Big] = (v\lambda - c)e^{-\lambda t}.$$

Now if λ is chosen so that the expected waiting cost is equal to the value of the prize,

$$\frac{c}{\lambda} = v \implies v\lambda = c,$$

then $\varphi'(t) = 0$. That is, contestant 1 receives the same expected payoff regardless of when he leaves. As a result he is perfectly content to choose his waiting time at random according to the same exponential distribution. Thus an exponentially distributed waiting time with parameter $\lambda = c/v$ satisfies property (i) of a symmetric equilibrium distribution. It is easy to see that $\varphi(0) = 0$, so the expected payoff is zero. (This makes sense, as the expected payoff is the same for both players, and they both can't win.)

To verify property (ii) of a symmetric equilibrium distribution, suppose some length of time passes and neither contestant has dropped out. Since the exponential distribution is memoryless, each contestant can redo the calculation above, and conclude there is no advantage to choosing a different time to leave. Thus contestant i is happy to leave at time t_i. If both contestants choose t_i according to this λ, then we have an equilibrium.

The length of the contest is $\min\{T_1, T_2\}$. Now $\min\{T_1, T_2\} \leqslant t$ if and only if it is not the case that $T_1 > t$ and $T_2 > t$. Thus

$$\begin{aligned} P\left(\min\{T_1, T_2\} \leqslant t\right) &= 1 - P\left(T_1 > t \;\&\; T_2 > t\right) \\ &= 1 - P\left(T_1 > t\right) P\left(t_2 > T\right) = 1 - e^{\lambda t}e^{\lambda t} \\ &= 1 - e^{2\lambda t}, \end{aligned}$$

which is an exponential survival function with parameter 2λ. Thus the expected length of the contest is $1/(2\lambda)$. So the winner and loser both expect to wait $1/(2\lambda)$ and the expected total cost incurred is equal to v, the value of the prize.

Note that this model implies that the observed length of contest durations should be exponentially distributed, provided c and v are the same in each contest. The noted game theorist Robert Rosenthal once told me that in fact the duration of display contests among dung flies is exponentially distributed, but he didn't mention any citations. A little digging found a paper by Parker and Thompson [4] where they find qualified support for this distribution, but argue that an asymmetric model would fit better.

14.7 The Winner's Curse

The Winner's Curse is a phenomenon that was first observed in oil leases. The Federal government claims all land on the continental shelf up to 200 miles offshore. It would auction off the

right to drill for oil on tracts, and oil companies found that despite their best efforts to employ scientific methods to estimating the value of the lease, they systematically lost money on their leases.

The explanation is straightforward. The value of a lease is an unknown number V, and each company i got an estimate X_i of the value of V. Let's suppose that each estimate is an independent draw from the same distribution F. If the company bids based on the distribution F, they will subject to the winner's curse. Namely, the winner will be the company with the largest X_i. But the distribution of the largest value of X_i is the n^{th} order statistic, the maximum, which has cdf F^n, not F. In order to avoid the winner's curse you have to take into account the fact that you win only when your $X_i = X_{(n)}$.

14.8 ⋆ The σ-algebra of events generated by random variables

Recall that the Borel sets of the real numbers are the members of the smallest σ-algebra that contains all the intervals. Given a set $X_1, \dots, X_n$ of random variables defined on the probability space $(S, \mathcal{E}, P)$ and intervals $I_1, \dots, I_n$,

$$(X_1 \in I_1, X_2 \in I_2, \dots X_n \in I_n) \text{ is an event.}$$

The smallest σ-algebra that contains all these events, as the intervals I_j range over all intervals, is called the **σ-algebra of events generated by $X_1, \dots, X_n$**, and is denoted

$$\sigma(X_1, \dots, X_n).$$

A function $g\colon S \to \boldsymbol{R}$ is $\sigma(X_1, \dots, X_n)$-measurable if for every interval I, the set $g^{-1}(I)$ belongs to $\mathcal{E}$. The following theorem is beyond the scope of this course, but may be found, for instance, in Aliprantis–Border [1, Theorem 4.41] or M. M. Rao [6, Theorem 1.2.3, p. 4].

14.8.1 Theorem *Let $\boldsymbol{X} = (X_1, \dots, X_n)$ be a random vector on $(S, \mathcal{E}, P)$ and let $g\colon S \to \boldsymbol{R}$. Then the function g is $\sigma(X_1, \dots, X_n)$-measurable if and only if there exists a Borel function $h\colon \boldsymbol{R}^{\mathrm{n}} \to \boldsymbol{R}$ such that $g = h \circ \boldsymbol{X}$.*

This means there is a one-to-one correspondence between $\boldsymbol{X}$-measurable functions and functions that depend only on $\boldsymbol{X}$. As a corollary we have:

14.8.2 Corollary *The set of $\boldsymbol{X}$-measurable functions is a vector subspace of the space of random variables.*

14.9 Conditioning on the value of a Random Variable: The discrete case

Pitman [5]: Section 6.1

Let X and Y be discrete random variables with joint pmf $p(x, y)$, and let p_X and p_Y be the respective marginals. Then $(Y = y)$ and $(X = x)$ are events so the conditional probability of $(Y = y)$ given $(X = x)$ is

$$P\left(Y = y \,\middle|\, X = x\right) = \frac{P\left(Y = y,\ X = x\right)}{P\left(X = x\right)} = \frac{p(x, y)}{\sum_{y'} p(x, y')} = \frac{p(x, y)}{p_X(x)}. \tag{4}$$

This is a function of x, known as the **conditional pmf of Y given $X = x$** and it defines the **conditional distribution of Y given $X = x$**

14.9.1 Example (Pitman [5, Exercise 6.1.5, p. 399]) Let X and Y be independent Poisson random variables with parameters μ and λ. What is the distribution of X given $X + Y = n$?

You may or may not recall what the distribution of $X + Y$ is, so let's just roll out the old convolution formula (recalling that X and Y are always nonnegative):

$$\begin{aligned}
P(X+Y=n) &= \sum_{k=0}^{n} p(k, n-k) && \text{convolution \& nonnegativity}\\
&= \sum_{k=0}^{n} e^{-\mu}\frac{\mu^k}{k!}e^{-\lambda}\frac{\lambda^{n-k}}{(n-k)!} && \text{independence}\\
&= \frac{e^{-(\mu+\lambda)}}{n!}\sum_{k=0}^{n}\frac{n!}{k!(n-k)!}\mu^k\lambda^{n-k} && \text{arithmetic}\\
&= \frac{e^{-(\mu+\lambda)}}{n!}(\mu+\lambda)^n && \text{Binomial Theorem}
\end{aligned}$$

which is a Poisson$(\mu + \lambda)$ distribution.

So

$$\begin{aligned}
P\big(X=k \,\big|\, X+Y=n\big) &= \frac{P(X=k,\ X+Y=n)}{P(X+Y=n)} && \text{defn.}\\
&= \frac{P(X=k,\ Y=n-k)}{P(X+Y=n)} && \text{arithmetic}\\
&= \frac{P(X=k)\ P(Y=n-k)}{P(X+Y=n)} && \text{independence}\\
&= \frac{\left(e^{-\mu}\frac{\mu^k}{k!}\right)\left(e^{-\lambda}\frac{\lambda^{n-k}}{(n-k)!}\right)}{\frac{e^{-(\mu+\lambda)}}{n!}(\mu+\lambda)^n} && \text{Poisson}\\
&= \binom{n}{k}\frac{\mu^k\lambda^{n-k}}{(\mu+\lambda)^n}\\
&= \binom{n}{k}\left(\frac{\mu}{\mu+\lambda}\right)^k\left(\frac{\lambda}{\mu+\lambda}\right)^{n-k}.
\end{aligned}$$

This is just the probability that a Binomial(n, p) random variable is equal to k, when $p = \mu/(\mu + \lambda)$. □

14.9.1 Application to the Poisson arrival process

In the Poisson process we discussed last time, N_t is the number of arrivals in the interval $[0, t]$ and it has Poisson(λt) distribution where λ is the arrival rate—the rate in the Exponential(λ) waiting time distribution.

Now let $0 < s < t$. What can we say about

$$P(N_s = k \,|\, N_t = n)?$$

Well N_s is Poisson(λs) random variable and it is independent of $N_t - N_s$, which is distributed according to the Poisson$\big(\lambda(t-s)\big)$ pmf. Moreover

$$N_t = N_s + (N_t - N_s),$$

so according to Example 14.9.1 that we just worked out, $P(N_s = k \,|\, N_t = n)$ is the Binomial

probability

$$P(N_s = k \,|\, N_t = n) = \binom{n}{k} \left(\frac{\lambda s}{\lambda s + \lambda(t-s)}\right)^k \left(\frac{\lambda(t-s)}{\lambda s + \lambda(t-s)}\right)^{n-k}$$
$$= \binom{n}{k} \left(\frac{s}{t}\right)^k \left(\frac{t-s}{t}\right)^{n-k}.$$

When you think about it, one interpretation of the Poisson process, that arrivals are uniformly scattered in an interval, so the probability of hitting $[0, s]$ given that $[0, t]$ has been hit is just s/t. (Remember $s < t$.) So the probability of getting k hits on $[0, s]$ given n hits on $[0, t]$ is given by the Binomial with probability of success s/t.

14.10 Conditional Expectation

In one way, conditional expectation is quite simple. It is the expectation of a random variable with respect to a conditional distribution.

For instance,

$$\boldsymbol{E}\left(Y \,|\, X = x\right) = \sum_y yP(Y = y \,|\, X = x) = \sum_y y\frac{p(x, y)}{p(x)}.$$

Note that we have used the common convention not to subscript the probability mass functions, but to use the names of the arguments, x or y or (x, y), to indicate whether we are talking about the marginal distribution X or Y, or the joint distribution of the vector (X, Y).

14.10.1 Example Continuing with the previous example, Example 14.9.1: Let X and Y be independent Poisson random variables with parameters μ and λ. Then we saw that the distribution of X given $X + Y = n$ was a Binomial$\big(n, \mu/(\mu + \lambda)\big)$ distribution:

$$P\left(X = k \,|\, X + Y = n\right) = \binom{n}{k} \left(\frac{\mu}{\mu + \lambda}\right)^k \left(\frac{\lambda}{\mu + \lambda}\right)^{n-k}.$$

Now a Binomial$\big(n, \mu/(\mu + \lambda)\big)$ has expectation $n\mu/(\mu + \lambda)$, so

$$\boldsymbol{E}\left(X \,|\, X + Y = n\right) = \frac{n\mu}{\mu + \lambda}$$

Similarly

$$\boldsymbol{E}\left(Y \,|\, X + Y = n\right) = \frac{n\lambda}{\mu + \lambda}$$

which implies the comforting conclusion that

$$\boldsymbol{E}\left(X \,|\, X + Y = n\right) + \boldsymbol{E}\left(Y \,|\, X + Y = n\right) = n.$$

□

14.11 Conditional Expectation, Part 2

So far we have defined $\boldsymbol{E}(Y \,|\, X = x)$ for discrete random variables. This quantity depends on x, so we can write it as a function of x. Let's use the name υ for this function, because y is the Latin equivalent of the Greek υ (ypsilon).

Pitman [5]: p. 402

$$\upsilon(x) = \boldsymbol{E}\left(Y \,|\, X = x\right).$$

The random variable $v(X)$ is known as the **conditional expectation of Y given X**, which is written

$$\boldsymbol{E}\left(Y \,\middle|\, X\right) = v(X).$$

The thing to see is that this is a random variable since it is a function of the random variable X. By Theorem 14.8.1 $\boldsymbol{E}\left(Y \,\middle|\, X\right)$ is $\sigma(X)$-measurable. Another way to say this is that

$\boldsymbol{E}\left(Y \,\middle|\, X\right)$ is a random variable that equals $\boldsymbol{E}\left(Y \,\middle|\, X = x\right)$ when $X = x$.

In terms of the probability space $(S, \mathcal{E}, P)$ on which X is defined, we have

$$X(s) = x \implies \boldsymbol{E}\left(Y \,\middle|\, X\right)(s) = v\big(X(s)\big) = \boldsymbol{E}\left(Y \,\middle|\, X = x\right).$$

14.11.1 Example Let S be the six-point sample space consisting of (x, y) pairs

$$S = \{(0,0), (0,1), (0,2), (1,0), (1,1), (1,2)\}$$

with probability measure given in the following diagram where the numbers in each box represent the probability of the sample point:

$y = 2$	$\frac{3}{24}$	$\frac{1}{24}$
$y = 1$	$\frac{5}{24}$	$\frac{3}{24}$
$y = 0$	$\frac{4}{24}$	$\frac{8}{24}$
	$x = 0$	$x = 1$

Let X and Y be random variables on S defined by $X(x, y) = x$ and $Y(x, y) = y$. Then

$$p_X(0) = p_X(1) = \frac{1}{2},$$

so

$$\boldsymbol{E}\, X = 1p_X(1) + 0p_X(0) = \frac{1}{2},$$

and

$$p_Y(0) = \frac{1}{2},\ p_Y(1) = \frac{1}{3},\ p_Y(2) = \frac{1}{6}$$

so

$$\boldsymbol{E}\, Y = 2p_Y(2) + 1p_Y(1) + 0p_Y(0) = \frac{2}{3}.$$

The event $X = 0$ is highlighted below.

2	$\frac{3}{24}$	$\frac{1}{24}$
1	$\frac{5}{24}$	$\frac{3}{24}$
0	$\frac{4}{24}$	$\frac{8}{24}$
	0	1

So

$$\begin{aligned}\boldsymbol{E}(Y \,\big|\, X=0) &= 2P(Y=2|X=0) + 1P(Y=1|X=0) + 0P(Y=0|X=0)\\ &= 2\frac{P(Y=2,X=0)}{P(X=0)} + 1\frac{P(Y=1,X=0)}{P(X=0)} + 0\frac{P(Y=0,X=0)}{P(X=0)}\\ &= 2\frac{\frac{3}{24}}{\frac{1}{2}} + 1\frac{\frac{5}{24}}{\frac{1}{2}} + 0\frac{\frac{4}{24}}{\frac{1}{2}}\\ &= \frac{11}{12}.\end{aligned}$$

Similarly

$$\boldsymbol{E}(Y \,\big|\, X=1) = \frac{5}{12}.$$

Thus the random variable $\boldsymbol{E}(Y \,\big|\, X)$ is defined on S by

$$\boldsymbol{E}(Y \,\big|\, X)(x,y) = \begin{cases} \frac{11}{12} & x=0,\\ \frac{5}{12} & x=1,\end{cases}$$

which can be represented in the diagram, where now the numbers in each box represent the value of the random variable $\boldsymbol{E}(Y \,\big|\, X)$:

2	$\frac{11}{12}$	$\frac{5}{12}$
1	$\frac{11}{12}$	$\frac{5}{12}$
0	$\frac{11}{12}$	$\frac{5}{12}$
	0	1

The expectation of the random variable $\boldsymbol{E}(Y \,\big|\, X)$ is

$$\boldsymbol{E}\Big(\boldsymbol{E}(Y \,\big|\, X)\Big) = \frac{11}{12}\cdot\frac{1}{2} + \frac{5}{12}\cdot\frac{1}{2} = \frac{2}{3}.$$

From the calculation above,

$$\boldsymbol{E}\Big(\boldsymbol{E}(Y \,\big|\, X)\Big) = \frac{2}{3} = \boldsymbol{E}\,Y.$$

□

14.12 Conditional Expectation is a Positive Linear Operator Too

Ordinary expectation is a positive linear operator that assigns random variables a real number. Conditional expectation assigns random variables another random variable, but it is also linear and positive:

Pitman [5]: p. 402

$$\boldsymbol{E}\left(aY + bZ \,\big|\, X\right) = a\,\boldsymbol{E}\left(Y \,\big|\, X\right) + b\,\boldsymbol{E}\left(Z \,\big|\, X\right)$$
$$Y \geqslant 0 \implies \boldsymbol{E}\left(Y \,\big|\, X\right) \geqslant 0.$$

14.13 Iterated Conditional Expectation

Since $\boldsymbol{E}\left(Y \,\middle|\, X\right)$ is a random variable, we can take its expectation.

Pitman [5]: p. 403

$$\boldsymbol{E}\big(\boldsymbol{E}\left(Y \,\middle|\, X\right)\big) = \boldsymbol{E}\,Y.$$

More remarkable is the following generalization.

14.13.1 Theorem *For a (Borel) function φ,*

$$\boldsymbol{E}\big(\varphi(X)\,\boldsymbol{E}\left(Y \,\middle|\, X\right)\big) = \boldsymbol{E}\big(\varphi(X)Y\big).$$

Proof: Let $\upsilon(x) = \boldsymbol{E}\left(Y \,\middle|\, X = x\right)$. Then

$$\upsilon(x) = \sum_y \frac{yp(x,y)}{p(x)},$$

and so

$$\begin{aligned}
\boldsymbol{E}\big(\varphi(X)\,\boldsymbol{E}\left(Y \,\middle|\, X\right)\big) &= \boldsymbol{E}\big(\varphi(X)\upsilon(X)\big) \\
&= \sum_x \varphi(x)\upsilon(x)p(x) \\
&= \sum_x \varphi(x)\left(\sum_y \frac{yp(x,y)}{p(x)}\right)p(x) \\
&= \sum_x \varphi(x)\left(\sum_y yp(x,y)\right) \\
&= \sum_{(x,y)} \varphi(x)yp(x,y) \\
&= \boldsymbol{E}\big(\varphi(X)Y\big).
\end{aligned}$$

∎

In light of Theorem 14.8.1 we have the following corollary.

14.13.2 Corollary *If Z is $\sigma(X)$-measurable, that is, if we can write $Z = \varphi(X)$ for some (Borel) function φ, then*

$$\boldsymbol{E}\big(Z\,\boldsymbol{E}\left(Y \,\middle|\, X\right)\big) = \boldsymbol{E}(ZY).$$

14.14 ⋆ Conditional Expectation is an Orthogonal Projection

Recall Corollary 14.8.2, which states that the space of $\sigma(X)$-measurable random variables is a vector subspace. If we further restrict attention to L_2, the space of square-integrable random variables, we have the inner product defined by

$$(X,Y) = \boldsymbol{E}(XY).$$

See Section 9.11 ⋆.

Recall from your linear algebra class that in an inner product space, the orthogonal projection of a vector y on a subspace M is the unique vector y_M such that $y_M \in M$, and $(y - y_M) \perp M$. It turns out that conditional expectation with respect to X is the orthogonal projection on to the subspace of $\sigma(X)$-measurable random variables.

14.14.1 Theorem *Let X, Y, and Z have finite variances, and assume that Z is $\sigma(X)$-measurable. Then*

$$\boldsymbol{E}\big(\big(Y - \boldsymbol{E}\left(Y \,\big|\, X\right)\big)Z\big) = 0.$$

Proof: Expand $\big(Y - \boldsymbol{E}\left(Y \,\big|\, X\right)\big)Z$ to get

$$\boldsymbol{E}\big(\big(YZ - \boldsymbol{E}\left(Y \,\big|\, X\right)\big)Z\big) = \boldsymbol{E}(YZ) - \boldsymbol{E}\big(\boldsymbol{E}\left(Y \,\big|\, X\right) Z\big),$$

since expectation is a linear operator. By Corollary 14.13.2,

$$\boldsymbol{E}\big(\boldsymbol{E}\left(Y \,\big|\, X\right) Z\big) = \boldsymbol{E}(YZ).$$

Substituting this in the previous equation proves the theorem. ∎

What this says is that the random variable $Y - \boldsymbol{E}\left(Y \,\big|\, X\right)$ is orthogonal to Z for every $\sigma(X)$-measurable random variable Z. Since $\boldsymbol{E}\left(Y \,\big|\, X\right)$ is itself $\sigma(X)$-measurable, we have

> $\boldsymbol{E}\left(Y \,\big|\, X\right)$ is the orthogonal projection of Y onto the vector space of $\sigma(X)$-measurable random variables, where the inner product (X, Y) is given by $\boldsymbol{E}(XY)$.

14.15 Conditional Expectation and Densities

When X and Y have a joint density the definition of $P\left(Y = y \,\big|\, X = x\right)$ seems ill-defined: Since when X has a density, $P\left(X = x\right) = 0$ for every x, we cannot define $P\left(Y = y \,\big|\, X = x\right)$ as $P\left(Y = y,\ X = x\right) / P\left(X = x\right)$, since that would entail division by zero.

It is beyond the scope of this course to prove it, but the following approach works. Given an interval B, a real number x, and $\varepsilon > 0$, consider

$$P\left(Y \in B \,\big|\, X \in (x - \varepsilon, x + \varepsilon)\right) = \frac{P\left(Y \in B,\ X \in (x - \varepsilon, x + \varepsilon)\right)}{P\left(X \in (x - \varepsilon, x + \varepsilon)\right)}.$$

If the marginal density f_X of X is positive and continuous at x, then the denominator is positive, so we are no longer dividing by zero. What we want is for this to tend to a limit as ε tends to zero, and in fact it does, a result known as the **Radon–Nikodym Theorem**.

> Define
>
> $$f_Y\left(y \,\big|\, X = x\right) = \frac{f(x, y)}{f_X(x)}$$
>
> so
>
> $$f(x, y) = f_Y\left(y \,\big|\, X = x\right) f_X(x).$$

> For a function $h\colon \boldsymbol{R} \to \boldsymbol{R}$,
>
> $$\begin{aligned}\boldsymbol{E}\left(h(Y) \,\big|\, X = x\right) &= \int h(y) f_Y\left(y \,\big|\, X = x\right)\, dy \\ &= \int h(y) \frac{f(x, y)}{f_X(x)}\, dy\end{aligned}$$

14.16 Conditioning with Several Variables

Let Y, X_1, …, X_n have joint density $f(y, x_1, \dots, x_n)$. The conditional density of Y given $X_1 = X_1, \dots, X_n = x_n$ is then

$$f_Y(y \,|\, x_1, \dots, x_n) = \frac{f(y, x_1, \dots, x_n)}{f_{X_1,\dots,X_n}(x_1, \dots, x_n)},$$

and we may speak of $\boldsymbol{E}\left(Y \,|\, X_1, \dots, X_n\right)$, etc.

Similarly,

$$f_{X_1,\dots,X_n}(X_1, \dots, X_n \,|\, Y = y) = \frac{f(y, x_1, \dots, x_n)}{f_Y(y)},$$

and we may speak of $\boldsymbol{E}\left(X_1, \dots, X_n \,|\, Y\right)$, etc.

14.17 Conditional Independence

If

$$f_{(X,Y)}(x, y \,|\, Z = z) = f_X(x \,|\, Z = z) f_Y(y \,|\, Z = z),$$

we say that X and Y are **conditionally independent given $\boldsymbol{Z = z}$**. If this is true for all z, we say that X and Y are **conditionally independent given $\boldsymbol{Z}$**.

Here is a transparent but artificial example.

14.17.1 Example The outcome space has eight points, each equally probable. The values of the random variables X, Y, and Z are given in the following table.

	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
Z	0	0	0	0	1	1	1	1
X	2	1	2	1	0	1	0	1
Y	2	2	0	0	1	1	0	0

You can see that X and Y are not independent. For example,

$$0 = P\left(X = 2, Y = 1\right) \neq P\left(X = 2\right) P\left(Y = 1\right) = \frac{1}{16}.$$

But X and Y are conditionally independent given Z. Here is the distribution of X and Y conditional on values of z:

	$Z=0$	
	$X=1$	$X=2$
$Y=2$	$\frac{1}{4}$	$\frac{1}{4}$
$Y=0$	$\frac{1}{4}$	$\frac{1}{4}$

	$Z=1$	
	$X=0$	$X=1$
$Y=1$	$\frac{1}{4}$	$\frac{1}{4}$
$Y=0$	$\frac{1}{4}$	$\frac{1}{4}$

It is easy to verify conditional independence. □

A common way that dependent, but conditionally independent random variables can arise is like this. Let U, V, Z be independent random variables, and let $X = Z + U$ and $Y = Z + V$. Then usually X and Y are not independent, but they are conditionally independent given Z.

Bibliography

[1] C. D. Aliprantis and K. C. Border. 2006. *Infinite dimensional analysis: A hitchhiker's guide*, 3d. ed. Berlin: Springer–Verlag.

[2] R. J. Larsen and M. L. Marx. 2012. *An introduction to mathematical statistics and its applications*, fifth ed. Boston: Prentice Hall.

[3] J. Maynard Smith. 1974. The theory of games and the evolution of animal conflicts. *Journal of Theoretical Biology* 47(1):209–221. DOI: 10.1016/0022-5193(74)90110-6

[4] G. A. Parker and E. A. Thompson. 1980. Dung fly struggles: a test of the war of attrition. *Behavioral Ecology and Sociobiology* 7(1):37–44. DOI: 10.1007/BF00302516

[5] J. Pitman. 1993. *Probability*. Springer Texts in Statistics. New York, Berlin, and Heidelberg: Springer.

[6] M. M. Rao. 1981. *Foundations of stochastic analysis*. Mineola, NY: Dover. Reprint of the 1981 Academic Press edition.