Lecture 15: Order Statistics; Conditional Expectation

Relevant textbook passages:
Pitman [10]: Section 4.6; Chapter 6
Larsen–Marx [7]: Section 3.11

15.1 Order statistics

Pitman [10]: § 4.6

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

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

The random 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. The results are standard and the exposition follows Casella and Berger [2, Section 5.4, pp. 226–232].

15.2 Marginal Distribution of Order Statistics

For the remainder of the discussion of order statistics, we shall assume that the original random variables $X_1, \dots, X_n$ are independent and identically distributed with an absolutely continuous distribution.

Let $X_1, \dots, X_n$ be independent and identically distributed random variables with common cumulative distribution function F and density $f = F'$, and let $(X_{(1)}, \dots, X_{(n)})$ denote the vector of order statistics.

The marginal cumulative distribution function of the k^{th} order statistic is the probability of the event $(X_k \leqslant x)$. Let U be the random variable that counts how many of the X_j's fall in the interval $(-\infty, x]$. Then U has a binomial$\big(n, F(u)\big)$ distribution, and

$$\big(X_{(k)} \leqslant x\big) = (U \geqslant k).$$

Thus using the properties of the binomial distribution we see that

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

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

15.3 Marginal Density of Order Statistics

Pitman [10]: p. 326

15.3.1 Theorem *Let $\big(X_{(1)},\dots,X_{(n)}\big)$ be the vector of order statistics from the independent and identically distributed random sample $X_1,\dots,X_n$, where the common distribution has cumulative distribution function F and density f. Then the marginal density of the k^{th} order statistic from a sample of size n is:*

$$\begin{aligned} f_{(k,n)}(x) &= \frac{n!}{(k-1)!(n-k)!}\,F(x)^{k-1}\big(1-F(x)\big)^{n-k}f(x) \\ &= k\binom{n}{k}F(x)^{k-1}\big(1-F(x)\big)^{n-k}f(x) \\ &= n\binom{n-1}{k-1}F(x)^{k-1}\big(1-F(x)\big)^{(n-1)-(k-1)}f(x). \end{aligned}$$

Proof: To find the density we need to differentiate the cdf. So assume F has a density $f = F'$, then the marginal density of $X_{(k)}$ is $F'_{(k,n)}$:

$$\begin{aligned} &\frac{d}{dx}F_{(k,n)}(x) \\ &= \frac{d}{dx}\sum_{j=k}^{n}\binom{n}{j}F(x)^{j}\big(1-F(x)\big)^{n-j} \\ &= \sum_{j=k}^{n} j\binom{n}{j}F(x)^{j-1}\big(1-F(x)\big)^{n-j}f(x) \;-\; \sum_{j=k}^{n}(n-j)\binom{n}{j}F(x)^{j}\big(1-F(x)\big)^{n-j-1}f(x) \\ &= k\binom{n}{k}F(x)^{k-1}\big(1-F(x)\big)^{n-k}f(x) \\ &\quad+\sum_{j=k+1}^{n} j\binom{n}{j}F(x)^{j-1}\big(1-F(x)\big)^{n-j}f(x) \;-\; \sum_{j=k}^{n-1}(n-j)\binom{n}{j}F(x)^{j}\big(1-F(x)\big)^{n-j-1}f(x) \end{aligned} \tag{2}$$

The last line above cancels out, since using the change of variables $\ell = j-1$,

$$\sum_{j=k+1}^{n} j\binom{n}{j}F^{j-1}(1-F)^{n-j} = \sum_{\ell=k}^{n-1}(n-\ell)\binom{n}{\ell}F^{\ell}(1-F)^{n-\ell-1},$$

since $j\binom{n}{j} = \frac{n!}{(j-1)!(n-j)!} = \frac{n!}{\ell!(n-\ell-1)!} = (n-\ell)\binom{n}{\ell}$. ∎

15.4 Joint Density of Order Statistics

15.4.1 Proposition *If F has a density $f = F'$, the joint density of the vector of all order statistics from a sample of size n is given by*

$$f_{X_{(1)},\dots,X_{(n)}}(x_1,\dots,x_n) = \begin{cases} n!\,f(x_1)\cdots f(x_n) & x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n \\ 0 & \textit{otherwise.} \end{cases}$$

Where does the $n!$ come from? The vector of order statistics depends only on the set of values $\{X_1,\dots,X_n\}$ and not their order. There are $n!$ ways to rearrange the vector $\boldsymbol{X}$ to get the same vector of order statistics, and the density is the product of the individual densities by independence. The order of the values of the vector of order statistics will satisfy $x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$, and the inequalities will be strict a.s.

15.5 Joint distribution of pairs of order statistics

The following may be found in Casella and Berger [2, Theorem 5.4.6, p. 230].

15.5.1 Theorem *Let $\big(X_{(1)},\dots,X_{(n)}\big)$ be the vector of order statistics from the independent and identically distributed random sample $X_1,\dots,X_n$, where the common distribution has cumulative distribution function F and density f. Then the joint density of the pair $(X_{(k)},X_{(\ell)})$, where $k<\ell$, is given by*

$$f_{X_{(k)},X_{(\ell)}}(u,v) = \frac{n!}{(k-1)!\,(\ell-k-1)!\,(n-\ell)!}F(u)^{k-1}\big(F(v)-F(u)\big)^{\ell-k-1}\big(1-F(v)\big)^{n-\ell}f(u)f(v), \quad (3)$$

for $u<v$ (and $=0$ otherwise).

Let's spend some time developing some intuition. Suppose some X_i is equal to u and another is equal to v. This accounts for the $f(u)f(v)$ term. In order for these to be the k^{th} and ℓ^{th} order statistics, of the remaining $n-2$ values, there would have to be $k-1$ of the X_i's $\leqslant u$ (the $F(u)^{k-1}$ term), and $n-\ell$ of the X_i's $>v$ (the $\big(1-F(v)\big)^{n-\ell}$ term), and there would have to be $\ell-k-1$ terms $>u$ and $\leqslant v$ (the $\big(F(v)-F(u)\big)^{\ell-k-1}$ term). There are $\frac{n!}{(k-1)!\,(\ell-k-1)!\,(n-\ell)!}$ ways we can partition n values in this way.

If that makes you happy, then I suggest you stop here. Otherwise prepare yourself for some unpleasant arithmetic.[1]

Proof of Theorem 15.5.1: (Cf. [2, Exercise 5.26, p. 260].) We start, as usual, by finding the joint cdf of the pair. Fix $u<v$. Partition the line into three bins

$$I_1=(-\infty,u], \quad I_2=(u,v], \quad I_3=(v,\infty).$$

Let

$$\alpha(u,v)=F(u), \quad \beta(u,v)=F(v)-F(u), \quad \gamma(u,v)=1-F(v). \quad (4)$$

Define

$$U=\#\{i : X_i\in I_1\}, \quad V=\#\{i : X_i\in I_2\}.$$

Then

$$U+V=\#\{i : X_i\leqslant v\} \quad \text{and} \quad \#\{i : X_i\in I_3\}=n-U-V.$$

Thus the vector $(U,V,n-U-V)$ has a multinomial$(n,\boldsymbol{p})$ distribution with bin probabilities $\boldsymbol{p}=(\alpha,\beta,\gamma)$. (Recall Section 3.9.)

The joint cumulative distribution function is defined by

$$F_{X_{(k)},X_{(\ell)}}(u,v)=P\big(X_{(k)}\leqslant u \;\&\; X_{(\ell)}\leqslant v\big).$$

Now the event

$$\big(X_{(k)}\leqslant u \;\&\; X_{(\ell)}\leqslant v\big)=(U\geqslant k \;\&\; U+V\geqslant \ell).$$

That is, in order for the k^{th} smallest value to be less than or equal to u, at least k values of the X's must fall into interval I_1, in other words, $U\geqslant k$. Likewise, in order for the ℓ^{th} smallest value to be less than or equal to v, we must have $U+V\geqslant\ell$.

If $U\geqslant\ell$, then a fortiori $U+V\geqslant\ell$. When $\ell>U\geqslant k$, which values can V assume? In order to have $U+V\geqslant\ell$, we must have $V\geqslant\ell-U$, but V cannot exceed $n-U$ because $U+V\leqslant n$, so $\ell-U\leqslant V\leqslant n-U$ for $\ell>U\geqslant k$. Thus

$$(U\geqslant k \;\&\; U+V\geqslant\ell)=(U\geqslant\ell)\bigcup\bigcup_{i=k}^{\ell-1}\bigcup_{j=\ell-i}^{n-i}(U=i \;\&\; V=j),$$

[1] "Unpleasant arithmetic" is a phrase I borrowed from my old professors, Tom Sargent and Neil Wallace [12], not to be confused with the late Caltech astronomer Wallace Sargent.

which is a union of pairwise disjoint events, so

$$\begin{aligned}
F_{X_{(k)},X_{(\ell)}}(u,v) &= P\big(X_{(k)} \leqslant u \;\&\; X_{(\ell)} \leqslant v\big) \\
&= P(U \geqslant k \;\&\; U+V \geqslant \ell) \\
&= P(U \geqslant \ell) + \sum_{i=k}^{\ell-1} \sum_{j=\ell-i}^{n-i} P(U=i \;\&\; V=j) \\
&= P(U \geqslant \ell) + \\
&\quad \sum_{i=k}^{\ell-1} \sum_{j=\ell-i}^{n-i} \frac{n!}{i!\,j!\,(n-i-j)!} \alpha^i \beta^j \gamma^{n-i-j}, \qquad (5)
\end{aligned}$$

where the last equality uses the multinomial probabilities for $P(U=i \;\&\; V=j)$.

We now use Theorem 9.4.1 to compute the density by differentiating the cdf, $f(u,v) = \frac{\partial^2 F(u,v)}{\partial u \partial v}$. First observe that since $P(U \geqslant \ell)$ does not depend on v, its mixed partial is zero. Thus the joint density is given by

$$f_{X_{(k)},X_{(\ell)}}(u,v) = \frac{\partial^2}{\partial u \partial v} \sum_{i=k}^{\ell-1} \sum_{j=\ell-i}^{n-i} \frac{n!}{i!\,j!\,(n-i-j)!} \alpha^i \beta^j \gamma^{n-i-j} \qquad (6)$$

This is the point where all the textbooks (e.g., Casella and Berger [2, Exercise 5.26, p. 260]) bail and leave the rest of the argument as an exercise, saying only that arguments similar to that in equation (2) are used.

To simplify notation, recall (4) and let

$$A(u,v) = \alpha^i(u,v), \quad B(u,v) = \beta^j(u,v), \quad C(u,v) = \gamma^{n-i-j}(u,v).$$

Then the partial derivatives of (6) involve the partials of ABC. Letting subscripts denotes partial derivatives ($A_u = \partial A/\partial u$, $A_{uv} = \partial^2 A/\partial u \partial v$, etc.), we have the following general formula for the derivatives of a product of three functions:

$$\begin{aligned}
&\frac{\partial^2}{\partial u\,\partial v} ABC = \\
&A_{uv}BC + A_uB_vC + A_uBC_v + A_vB_uC + AB_{uv}C + AB_uC_v + A_vBC_u + AB_vC_u + ABC_{uv}.
\end{aligned}$$

In our particular case, by equation (4), $A_v = A_{uv} = C_u = C_{uv} = 0$, so this reduces to

$$A_uB_vC + AB_{uv}C + A_uBC_v + AB_uC_v, \qquad (7)$$

where

$$\begin{aligned}
A_u = i\alpha^{i-1}f(u), \quad B_u = -j\beta^{j-1}f(u), \quad B_v = j\beta^{j-1}f(v), \\
B_{uv} = -j(j-1)\beta^{j-2}f(u)f(v), \quad C_v = -(n-i-j)\gamma^{n-i-j-1}f(v).
\end{aligned}$$

Using this and (7), we can rewrite (6) as $f(u)f(v)$ times the following quantity

$$\begin{aligned}
&\sum_{i=k}^{\ell-1} \frac{i\alpha^{i-1}}{i!} \underbrace{\sum_{j=\ell-i}^{n-i} \frac{n!}{j!\,(n-i-j)!} j\beta^{j-1}\gamma^{n-i-j}}_{=a} \\
&-\sum_{i=k}^{\ell-1} \frac{\alpha^{i}}{i!} \underbrace{\sum_{j=\ell-i}^{n-i} \frac{n!}{j!\,(n-i-j)!} j(j-1)\beta^{j-2}\gamma^{n-i-j}}_{=b} \\
&-\sum_{i=k}^{\ell-1} \frac{i\alpha^{i-1}}{i!} \underbrace{\sum_{j=\ell-i}^{n-i} \frac{n!}{j!\,(n-i-j)!} (n-i-j)\beta^{j}\gamma^{n-i-j-1}}_{=c} \\
&+\sum_{i=k}^{\ell-1} \frac{\alpha^{i}}{i!} \underbrace{\sum_{j=\ell-i}^{n-i} \frac{n!}{j!\,(n-i-j)!} j(n-i-j)\beta^{j-1}\gamma^{n-i-j-1}}_{=d}.
\end{aligned} \tag{8}$$

We start by looking at the inside sums labeled above and analyzing them as differences of pairs, $a-c$ and $b-d$.

Start with the difference $a-c$. Here, in the sum c, the upper bound for j is $n-i$, but $j=n-i$ implies $n-i-j=0$, so the upper bound can be replaced by $n-i-1$. Then $a-c$ becomes

$$\underbrace{\sum_{j=\ell-i}^{n-i} \frac{n!}{j!\,(n-i-j)!} j\beta^{j-1}\gamma^{n-i-j}}_{a} - \underbrace{\sum_{j=\ell-i}^{n-i-1} \frac{n!}{j!\,(n-i-j-1)!} \beta^{j}\gamma^{n-i-j-1}}_{c}$$

Separating out the $j=\ell-i$ term from the sum a gives

$$a = \boxed{\frac{n!}{(\ell-i)!\,(n-\ell)!}(\ell-i)\beta^{\ell-i-1}\gamma^{n-\ell}} + \underbrace{\sum_{j=\ell-i+1}^{n-i} \frac{n!}{(j-1)!\,(n-i-j)!} \beta^{j-1}\gamma^{n-i-j}}_{=a'}.$$

Using the change of variable $m=j-1$ (so $j=m+1$) in a' we have

$$a' = \sum_{m=\ell-i}^{n-i-1} \frac{n!}{m!\,(n-i-m-1)!} \beta^{m}\gamma^{n-i-m-1}$$

so, swapping the dummy variable j for the dummy variable m, we see that $a'-c=0$. So $a-c$ leaves only the boxed term

$$a-c = \frac{n!}{(\ell-i)!\,(n-\ell)!}(\ell-i)\beta^{\ell-i-1}\gamma^{n-\ell}. \tag{9}$$

Take a breath, and consider $b-d$. Note that when $j=n-i$, then $n-i-j=0$, so we may replace the upper limit in the sum d by $n-i-1$. We want to use the same sort of cancellation that occurred in (2), so write the difference $b-d$ as

$$\underbrace{\sum_{j=\ell-i}^{n-i} \frac{n!}{j!\,(n-i-j)!} j(j-1)\beta^{j-2}\gamma^{n-i-j}}_{b} - \underbrace{\sum_{j=\ell-i}^{n-i-1} \frac{n!}{j!\,(n-i-j)!} j(n-i-j)\beta^{j-1}\gamma^{n-i-j-1}}_{d}.$$

Separate out the first term ($j = \ell - i$) from sum b, and write it as

$$b = \boxed{\frac{n!}{(\ell-i)!\,(n-\ell)!}(\ell-i)(\ell-i-1)\beta^{\ell-i-2}\gamma^{n-\ell}} + \underbrace{\sum_{j=\ell-i+1}^{n-i} \frac{n!}{j!\,(n-i-j)!} j(j-1)\beta^{j-2}\gamma^{n-i-j}}_{=b'}.$$

Now use the change of variable $m = j + 1$, so $j = m - 1$, to rewrite (d) as

$$\begin{aligned} d &= \sum_{m=\ell-i+1}^{n-i} \frac{n!}{(m-1)!\,(n-i-m+1)!}(m-1)(n-i-m+1)\beta^{m-2}\gamma^{n-i-m} \\ &= \sum_{m=\ell-i+1}^{n-i} \frac{n!}{(m-1)!\,(n-i-m)!}(m-1)\beta^{m-2}\gamma^{n-i-m} \end{aligned}$$

and replacing the dummy variable m with the dummy variable j we see that $d = b'$. (If it looks like b' has an extra j term, use it convert $j!$ to $(j-1)!$.) So $b - d$ becomes the boxed term,

$$b - d = \frac{n!}{(\ell-i-1)!\,(n-\ell)!}(\ell-i-1)\beta^{\ell-i-2}\gamma^{n-\ell}. \tag{10}$$

Now (8) is just

$$\sum_{i=k}^{\ell-1} \frac{i\alpha^{i-1}}{i!}(a-c) - \sum_{i=k}^{\ell-1} \frac{\alpha^i}{i!}(b-d),$$

which using (9) and (10) has just been reduced to

$$\sum_{i=k}^{\ell-1} \frac{n!}{(i-1)!\,(\ell-i-1)!\,(n-\ell)!}\alpha^{i-1}\beta^{\ell-i-1}\gamma^{n-\ell} - \sum_{i=k}^{\ell-1} \frac{n!}{i!\,(\ell-i-1)!\,(n-\ell)!}(\ell-i-1)\alpha^i\beta^{\ell-i-2}\gamma^{n-\ell}.$$

Using the now familiar tricks of replacing the upper limit in the second sum by $\ell - 2$ (since when $i = \ell - 1$, we have $\ell - i - 1 = 0$) and separating the $i = k$ term of the first sum we get

$$\boxed{\frac{n!}{(k-1)!\,(\ell-k-1)!\,(n-\ell)!}\alpha^{k-1}\beta^{\ell-k-1}\gamma^{n-\ell}} + \underbrace{\sum_{i=k+1}^{\ell-1} \frac{n!}{(i-1)!\,(\ell-i-1)!\,(n-\ell)!}\alpha^{i-1}\beta^{\ell-i-1}\gamma^{n-\ell}}_{=e} - \underbrace{\sum_{i=k}^{\ell-2} \frac{n!}{i!\,(\ell-i-2)!\,(n-\ell)!}\alpha^i\beta^{\ell-i-2}\gamma^{n-\ell}}_{=f}.$$

But you should know by now that $e - f = 0$ (use the change of variable $m = i + 1$ in f), so we are left with just the boxed term. Recalling that all of this is to be multiplied by $f(u)f(v)$, we are left with (3) and this whole nasty mess is now in the rear-view mirror. ∎

And then there is this generalization, which is given as Exercise 10.6 in Rao [11, p. 215].

15.5.2 Theorem *The joint density of the m-tuple $(X_{(k_1)}, \dots, X_{(k_m)})$, where $k_1 < \cdots < k_m$, is given by*

$$f_{X_{(k_1)},\dots,X_{(k_m)}}(u_1, \dots, u_m) =$$

$$\frac{n!}{(k_1-1)!(k_2-k_1-1)!\cdots(n-k_m-1)!}$$

$$\times F(u_1)^{k_1-1}\big(F(u_2)-F(u_1)\big)^{\ell-k-1}\cdots\big(1-F(u_m)\big)^{n-k_m} f(u_1)\cdots f(u_m),$$

for $u_1 < \cdots < u_m$ (and $= 0$ otherwise).

Try your hand at proving it.

15.6 Some special order statistics

The 1st 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.

15.7 The range of a sample

This section is based on Guttman, Wilks, and Hunter [6, Section 13.5, pp. 314–316]. Why go to the bother of computing the joint distribution of pairs of order statistics? One reason is to find the distribution of the range of a sample. Given a sample $X_1, \dots, X_n$ from the distribution, the **range** R of the sample is

$$R = \max_i X_i - \min_i X_i = X_{(n)} - X_{(1)}.$$

By Theorem 15.5.1 the joint density of $(X_{(1)}, X_{(n)})$ is

$$f_{X_{(1)},X_{(n)}}(u,v) = n(n-1)\big(F(v) - F(u)\big)^{n-2} f(u) f(v).$$

The cdf of R is given by

$$\begin{aligned} P(R \leqslant x) &= P\big(X_{(n)} - X_{(1)} \leqslant x\big) \\ &= P\big(X_{(n)} \leqslant X_{(1)} + x\big) \\ &= \int_{-\infty}^{\infty} \int_{u}^{u+x} n(n-1)\big(F(v) - F(u)\big)^{n-2} f(u) f(v)\, dv\, du \\ &= n(n-1) \int_{-\infty}^{\infty} \left(\int_{u}^{u+x} \big(F(v) - F(u)\big)^{n-2} f(v)\, dv \right) f(u)\, du. \end{aligned} \tag{11}$$

To evaluate the inner integral, use the change of variable $v \mapsto t = F(v) - F(u)$ and note that

$$\begin{aligned} \int_{u}^{u+x} \big(F(v) - F(u)\big)^{n-2} f(v)\, dv &= \int_{0}^{F(u+x)-F(u)} t^{n-2}\, dt \\ &= \frac{1}{n-1} \int_{0}^{F(u+x)-F(u)} \frac{d}{dt} t^{n-1}\, dt \\ &= \frac{1}{n-1} \big(F(u+x) - F(u)\big)^{n-1}. \end{aligned}$$

So (11) becomes

$$F_R(x) = P(R \leqslant x) = n \int_{-\infty}^{\infty} \big(F(u+x) - F(u)\big)^{n-1} f(u)\, du$$

To find the density, we differentiate with respect to x:

$$f_R(x) = n(n-1) \int_{-\infty}^{\infty} \big(F(u+x) - F(u)\big)^{n-2} f(u) f(u+x)\, du.$$

15.8 Uniform order statistics and the Beta function

Pitman [10]: pp. 327–328

For a Uniform[0,1] distribution, $F(t) = t$ and $f(t) = 1$ on $[0,1]$. In this case, Theorem 15.3.1 tells us:

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

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

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

or

$$\int_0^1 (1-t)^{n-m}\,t^{m-1}\,dt = \frac{1}{n\binom{n-1}{m-1}} = \frac{(m-1)!(n-m)!}{n!}. \tag{12}$$

Now change variables by setting

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

Then rewrite (12) as

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

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 14.13.1.

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

15.8.1 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, the Beta function is related to the binomial coefficients as follows:

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

15.8.2 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 values of 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 (m,n) order statistic has a beta$(m, n-m+1)$ distribution and so has mean

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

[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 $1/n$.]

15.9 The war of attrition

In the 1970s, the ethologist John Maynard Smith [8] 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 $1 - e^{-\lambda t}$, he wins 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

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

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 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 it can't be that both 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 content 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

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

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 [9] where they find qualified support for this distribution, but argue that an asymmetric model would fit better.

15.10 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 auctions 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 a random variable V, and each company i gets an estimate X_i of the value of V. Let's suppose that each estimate is

$$X_i = V + U_i,$$

where V is the common value and each U_i is an independent draw from the same distribution F.

If the company bids X_i, their expected payoff is

$$\boldsymbol{E}(V - X_i) = 0,$$

and they will subject to the **winner's curse**. Namely, the winner i^* will be the company with the largest U_i. But the distribution of the largest value of U_i is the n^{th} order statistic of $(U_1, \dots, U_n)$, the maximum, which has cdf F^n, not F. The cumulative distribution function F^n stochastically dominates F, so

$$\boldsymbol{E}(V - X_{(n)}) < 0.$$

In order to avoid the winner's curse you have to take into account the fact that you win only when your $X_i = X_{(n)}$, and

$$\boldsymbol{E}\left(V - X_i \;\middle|\; X_i = X_{(n)}\right) < 0.$$

15.11 ⋆ Extreme value distributions

[**As of 2021, this section is provisional, and has not been fully vetted.**]

Flesh this out.

Extreme value theory has to do with the distribution of extreme values. For instance, the distribution of the hottest day of the year, the distribution of the maximum daily rainfall, the maximum storm surge, etc. Now it is clear that if $X_1, \dots, X_n$ are independent and identically distributed, then the cumulative distribution function of the n^{th}order statistic, $F_{(n)} = F^n$, has the property that $F^n(x) \to 0$ for every x that is not at the maximum of the support. So the limiting distribution, if one exists, is degenerate. This is not very useful for answering questions, such as how long between extreme events?

A useful approach has been to change the scale of the units with n, so that $X_{(n)}$ converges in distribution to something interesting. There is a remarkable theorem, known as he Extremal Types Theorem, which categorizes such rescaled maxima. We don't have the time in this course to go into this in any depth, but if you are interested in issues such as climate change, you will want to brush up on the subject. A good starting point is the book by Stuart Coles [3], especially Chapter 3, pp. 45–73.

So let $X_1, \dots, X\_n, \dots$ be independent and identically distributed with cumulative distribution function F, and set

$$M_n = \max\{X_1, \dots, X_n\}.$$

We have already argued that M_n converges in distribution to a degenerate random variable, or else it may not converge at all. (Its cumulative distribution function could converge to zero, which is not the cumulative distribution function of any probability distribution.) But it may be possible to rescale M_n so that

$$M_n^* = \frac{M_n - b_n}{a_n} \xrightarrow{\mathcal{D}} G,$$

where G is a nondegenerate cumulative distribution function. The next result may be found in Coles [3, Theorem 3.1, p. 46].

15.11.1 Extremal Types Theorem *If there exist sequences $a_n > 0$, b_n, and a nondegenerate cumulative distribution function G such that*

$$\frac{M_n - b_n}{a_n} \xrightarrow{\mathcal{D}} G,$$

then G takes one of just three forms.

Type I *G is of the form*

$$G(x) = e^{-\left(-e^{-\left(\frac{x-b}{a}\right)}\right)}, \qquad (-\infty < x < \infty)$$

where $a > 0$.

Type II *G is of the form*

$$G(x) = \begin{cases} e^{-\left(\frac{x-b}{a}\right)^{-\alpha}} & x > b \\ 0 & x \leqslant b \end{cases}$$

where $a, \alpha > 0$.

Type III *G is of the form*

$$G(x) = \begin{cases} e^{-\left(-\left(\frac{x-b}{a}\right)^{\alpha}\right)} & x < b \\ 1 & x \geqslant b \end{cases}$$

where $a, \alpha > 0$.

*The distributions are referred to as **extreme value distributions**. The parameters a and b are scale and location parameters.*

These distributions are referred to as the Type I, Type II, and Type III extreme value distributions, but they also have other names.

- A Type I extremal distribution is also known as a **Gumbel distribution**.
- A Type II extremal distribution is also known as a **Fréchet distribution**.
- A Type III extremal distribution is also known as a **reversed Weibull distribution**, or sometimes, as in Coles [3, pp. 46–47], simply a **Weibull distribution**.

This latter terminology is unfortunate, but if we restrict our attention to the case $b = 0$ and replace x by $-x$, the expression for the Type III cumulative distribution function above becomes

$$G(x) = \begin{cases} 1 - e^{-\left(\frac{x}{a}\right)^{\alpha}} & x > 0 \\ 0 & x \leqslant 0 \end{cases}$$

which is our Weibull cumulative distribution function. This definition of the Weibull distribution is also the one used by Pitman [10, p. 301] as well as Forbes, et al. [4, p. 193]. The term "reversed Weibull" may be found for instance on wikipedia.

It is possible to combine the three types into a single family of cumulative distribution functions, the **generalized extreme value distribution**, or **GEV**, cumulative distribution function, which is given by

$$G(x) = e^{-\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-1/\xi}}$$

for x in the set

$$\{x : 1 + \xi\frac{x-\mu}{\sigma} > 0\}.$$

Elaborate on these results.

The parameter μ is a location parameter, σ is a scale parameter, and ξ is a shape parameter. If X is a GEV random variable with $\xi > 0$, let $\xi = \alpha^{-1} > 0$. Then $Y = 1 + \xi(X - \mu)/\sigma$ has a Fréchet distribution. When $\xi < 0$, let $\xi = -\alpha^{-1} < 0$. Then $Y = -\big(1+\xi(X-\mu)/\sigma\big)$ has reversed Weibull distribution. The expression for the GEV cumulative distribution function is undefined for $\xi = 0$, but it has a limit as $\xi \to 0$, and that limit is a Gumbel cumulative distribution function.

An important property of the GEV family, is that it is closed under change of units. That is, if X has a GEV distribution, then $aX + b$ also has a GEV distribution (with different parameters). Moreover if G is a GEV cumulative distribution function, then so is G^n.

As a result, if M_n is a maximum (n^{th} order statistic) such that $(M_n - b_n)/a_n \xrightarrow{\mathcal{D}} G$, then there is GEV distribution $\hat{G}$ such that $M_n \approx \hat{G}$. See Coles [3, pp. 48–49].

15.12 Conditioning on the value of a random variable: The discrete case

Pitman [10]: 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

$$P\big(Y = y \bigm| X = x\big) = \frac{P(Y = y,\ X = x)}{P(X = x)} = \frac{p(x, y)}{\sum_{y'} p(x, y')} = \frac{p(x, y)}{p_X(x)}. \tag{13}$$

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 $X = x$**

15.12.1 Example (Pitman [10, Exercise 6.1.5, p. 399]) Let X and Y be independent Poisson random variables with parameters μ and λ. What is the conditional 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

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

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

15.12.1 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

$$P(N_s = k \mid N_t = n)?$$

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 15.12.1 that we just worked out, $P(N_s = k \mid N_t = n)$ is the Binomial probability

$$\begin{aligned} P(N_s = k \mid N_t = n) &= \binom{n}{k} \left(\frac{\lambda s}{\lambda s + \lambda(t-s)}\right)^k \left(\frac{\lambda(t-s)}{\lambda s + \lambda(t-s)}\right)^{n-k} \\ &= \binom{n}{k} \left(\frac{s}{t}\right)^k \left(\frac{t-s}{t}\right)^{n-k}. \end{aligned}$$

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.

15.13 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,

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

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).

15.13.1 Example Continuing with the previous example, Example 15.12.1: 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:

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

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

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

Similarly

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

which implies the comforting conclusion that

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

□

15.14 Conditional Expectation, Part 2

So far we have defined $\boldsymbol{E}(Y \mid X=x)$ 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 [10]: p. 402

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

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

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

The thing to see is that this is a random variable since it is a function of the random variable X. By Theorem 2.11.2 $\boldsymbol{E}\left(Y \;\middle|\; X\right)$ 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 $(\Omega,\mathcal{F},P)$ on which X is defined, we have

$$X(\omega)=x \quad\Longrightarrow\quad \boldsymbol{E}\left(Y \;\middle|\; X\right)(\omega)=\upsilon\big(X(\omega)\big)=\boldsymbol{E}\left(Y \;\middle|\; X=x\right).$$

15.14.1 Example Consider the following random experiment. A die is rolled and a coins tossed (independently). If X is the value shown on the die, and Y is the indicator of Tails, let

$$W=X+Y.$$

The natural sample space S for this experiment has twelve equally likely points:

$$S=\{(x,y):x=1,\dots,6;\ y=0,1\}.$$

The random variable W can be represented by the following table:

$y=1$	2	3	4	5	6	7
$y=0$	1	2	3	4	5	6
	$x=1$	$x=2$	$x=3$	$x=4$	$x=5$	$x=6$

It is easy to see that $\boldsymbol{E}\,X=\frac{1}{2}$, $\boldsymbol{E}\,Y=3\frac{1}{2}$, and $\boldsymbol{E}\,W=4$,

The event $Y=0$ is highlighted below.

1	2	3	4	5	6	7
0	1	2	3	4	5	6
	1	2	3	4	5	6

So

$$\boldsymbol{E}(W \mid Y=0)=\sum_{w=1}^{6} wP(W=w \mid Y=0)=\sum_{w=1}^{6} w\frac{P(W=w\ \&\ Y=0)}{P(Y=0)}=\sum_{w=1}^{6} w\frac{\frac{1}{12}}{\frac{1}{2}}=3\frac{1}{2}.$$

Similarly

$$\boldsymbol{E}(W \mid Y = 1) = 4\frac{1}{2}.$$

Thus the random variable $\boldsymbol{E}(W \mid Y)$ is defined on S by

$$\boldsymbol{E}(W \mid Y)(x, y) = \begin{cases} 3\frac{1}{2} & x = 0, \\ 4\frac{1}{2} & x = 1, \end{cases}$$

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

1	$4\frac{1}{2}$	$4\frac{1}{2}$	$4\frac{1}{2}$	$4\frac{1}{2}$	$4\frac{1}{2}$	$4\frac{1}{2}$
0	$3\frac{1}{2}$	$3\frac{1}{2}$	$3\frac{1}{2}$	$3\frac{1}{2}$	$3\frac{1}{2}$	$3\frac{1}{2}$
	1	2	3	4	5	6

The expectation of the random variable $\boldsymbol{E}(W \mid Y)$ is

$$\boldsymbol{E}\Big(\boldsymbol{E}(W \mid Y)\Big) = 3\frac{1}{2} \cdot \frac{1}{2} + 4\frac{1}{2} \cdot \frac{1}{2} = 4,$$

so,

$$\boldsymbol{E}\Big(\boldsymbol{E}(W \mid Y)\Big) = 4 = \boldsymbol{E}\, W.$$

□

15.15 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 to random variables another random variable, but it is also linear and positive:

Pitman [10]: p. 402

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

15.16 Iterated Conditional Expectation

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

Pitman [10]: p. 403

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

More remarkable is the following generalization.

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

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

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

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

and so

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

∎

In light of Theorem 2.11.2 we have the following corollary.

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

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

15.17⋆ Conditional Expectation is an Orthogonal Projection

Recall Corollary 2.11.3, 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 9.13⋆.

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.

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

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

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

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

since **expectation is a positive linear operator**. By Corollary 15.16.2,

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

Substituting this in the previous equation proves the theorem. ∎

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

> $\boldsymbol{E}\left(Y \mid X\right)$ 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 $\boldsymbol{E}(XY)$.

15.18 Conditional Expectation and Densities

When X and Y have a joint density the definition of $P\big(Y = y \mid X = x\big)$ seems ill-defined: Since when X has a density, $P(X = x) = 0$ for every x, we cannot define $P\big(Y = y \mid X = x\big)$ as $P(Y = y,\ X = x)/P(X = x)$, 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

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

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. We want this to tend to a limit as ε tends to zero, and in fact it does. See Aliprantis and Burkinshaw [1, pp. 366–371], especially Theorem 39.4.

> Define the **conditional density** of Y given $X = x$ by
>
> $$f\left(y \mid x\right) = \frac{f(x, y)}{f(x)}$$
>
> so $f(x, y) = f\left(y \mid x\right) f(x)$.
>
> Then for a function $\varphi\colon \boldsymbol{R} \to \boldsymbol{R}$, there is a (Borel) function h_φ satisfying
>
> $$\boldsymbol{E}\left(\varphi(Y) \mid X\right) = h_\varphi(X),$$
>
> where
>
> $$\begin{aligned} h_\varphi(x) &= \int \varphi(y) f\left(y \mid x\right)\, dy \\ &= \int \varphi(y) \frac{f(x, y)}{f(x)}\, dy. \end{aligned}$$
>
> As a special case,
>
> $$\boldsymbol{E}\left(Y \mid X\right) = h(X),$$
>
> where
>
> $$\begin{aligned} h(x) &= \int y f\left(y \mid x\right)\, dy \\ &= \int y \frac{f(x, y)}{f(x)}\, dy. \end{aligned}$$

15.19 Conditioning with Several Variables

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

$$f(y \mid x_1, \dots, x_n) = \frac{f(y, x_1, \dots, x_n)}{f(x_1, \dots, x_n)},$$

and we may speak of $\boldsymbol{E}\left(Y \mid X_1, \dots, X_n\right)$, etc. Note that I am using the convention of not subscripting the density functions, but instead letting the arguments identify the density.

Similarly,

$$f(x_1, \dots, x_n \mid Y = y) = \frac{f(y, x_1, \dots, x_n)}{f(y)},$$

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

15.20 Conditional Independence

If

$$p_(x, y \mid Z = z) = p(x \mid z) \cdot p(y \mid z),$$

or

$$f(x, y \mid Z = z) = f(x \mid z) \cdot f(y \mid z),$$

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 $\boldsymbol{Z}$**.

15.20.1 Example 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

$$X = Z + U \quad \text{and} \quad Y = Z + V.$$

Then X and Y are not usually independent, but they are conditionally independent given Z:

$$\begin{aligned} P\big(Z + U = x,\ Z + V = y \mid Z = z\big) &= P(U = x - z,\ V = y - z) \\ &= P(U = x - z)\, P(V = y - z) \\ &= P\big(Z + U = x \mid Z = z\big)\, P\big(Z + V = y \mid Z = z\big). \end{aligned}$$

□

Bibliography

[1] C. D. Aliprantis and O. Burkinshaw. 1998. *Principles of real analysis*, 3d. ed. San Diego: Academic Press.

[2] G. Casella and R. L. Berger. 2002. *Statistical inference*, 2d. ed. Belmont, California: Brooks/Cole Cengage Learning.

[3] S. Coles. 2001. *An introduction to statistical modeling of extreme values*. London: Springer-Verlag.

[4] C. Forbes, M. Evans, N. Hastings, and B. Peacock. 2011. *Statistical distributions*, 4th. ed. Hoboken, New Jersey: John Wiley & Sons.

[5] E. J. Gumbel. 2004. *Statistics of extremes*. Mineola, New York: Dover. Reprint of the 1958 edition published by Columbia University Press.

[6] I. Guttman, S. S. Wilks, and J. S. Hunter. 1971. *Introductory engineering statistics*, second ed. New York: John Wiley & Sons.

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

[8] 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

[9] 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

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

[11] C. R. Rao. 1973. *Linear statistical inference and its applications*, 2d. ed. Wiley Series in Probability and Mathematical Statistics. New York: Wiley.

[12] T. J. Sargent and N. Wallace. 1981. Some unpleasant monetarist arithmetic. *Federal Reserve Bank of Minneapolis Quarterly Report* 5(3).
https://www.minneapolisfed.org/research/qr/qr531.pdf