Chapter 10

Conditional Probability and Expectation

Conditioning is a very important tool in statistics. One application is to eliminate the dependence of the distribution of a test statistic on unknown parameters. For instance, Fisher's exact test, which compares two groups with respect to the proportions with events, conditions on the total number of people with events. Its resulting null hypergeometric distribution does not depend on any unknown parameters. Similarly, clinical trials comparing a vaccine to placebo often use what is called the "conditional binomial procedure," which entails conditioning on the total number of infections across both arms and the amount of follow-up time in each arm. Under the assumption that the numbers of infections in the two arms follow Poisson distributions with rates proportional to the amount of follow-up, the conditional distribution of the number of vaccine infections follows a binomial distribution with probability parameter free of any unknown parameters. Permutation tests condition on even more, namely all data other than the treatment labels. The null distribution of the permutation test statistic depends only on those data, not on parameters.

Another reason for conditioning is to create independence. For example, in the comparison of several treatments to a control with respect to a continuous outcome Y like log viral load or blood pressure, we examine differences of means, $\hat{\delta}_1 = \bar{Y}_1 - \bar{Y}_0, \ldots, \hat{\delta}_k = \bar{Y}_k - \bar{Y}_0$, where $\bar{Y}_i$ is the mean in treatment arm i and $\bar{Y}_0$ is the mean of the control arm. The $\hat{\delta}_i$ are dependent because they share the same control mean $\bar{Y}_0$; once we condition on $\bar{Y}_0$, $\hat{\delta}_1, \ldots, \hat{\delta}_k$ are independent. Likewise, a mixed model might assume person-specific intercepts that make different observations on the same person correlated. Once we condition on the random effect for a given person, these observations become independent.

Conditioning also allows different entities to interpret data from their own perspectives. For instance, ideally, a medical diagnostic test should declare the disease present if the patient truly has it, and absent if the patient does not have it. The probabilities of these conclusions are known as *sensitivity* and *specificity*, respectively. These are conditional probabilities of correct diagnoses given that the patient does or does not truly have the disease. The doctor wants to ensure a small proportion of incorrect diagnoses, hence high sensitivity and specificity. The patient is concerned only about the accuracy of his or her own diagnosis. "Given that the test was positive, what is the probability that I really have the disease," or "Given that the test was negative, what is the probability that I am really disease free?" These are different conditional probabilities, known as *positive predictive value*

and *negative predictive value*, respectively.

The importance of conditioning may be matched only by the care required to avoid mistakes while carrying it out. Many paradoxes, including the two envelope paradox of Example 1.4, involve errors in conditional probability or expectation. These generally involve conditioning on sets of probability 0; conditioning on sets of positive probability does not cause problems. If $\mathrm{E}(|Y|) < \infty$ and B is any Borel set with $P(X \in B) > 0$, the expected value of Y given that $X \in B$ is unambiguously defined by

$$\mathrm{E}(Y \,|\, X \in B) = \frac{\mathrm{E}\{Y I(X \in B)\}}{P(X \in B)}. \tag{10.1}$$

But suppose we try to condition on the actual value of X by replacing $X \in B$ with $X = x$. If X is continuous, then $P(X = x) = 0$ for each x, so we would be dividing by 0 in (10.1). The solution is to think more generally: instead of considering the **specific** value x such as $x = 5$, think about the information you gain about ω, and therefore about $Y(\omega)$, from knowledge of $X(\omega)$. This more general thinking leads us to define the random variable $\mathrm{E}(Y \,|\, X)$, the conditional expectation of Y given X. Once we understand this concept, we generalize even more. Recall that Section 4.1.1 encouraged us to think in terms of the sigma-field generated by a random variable. Knowing X tells us, for each Borel set B, whether $X \in B$ occurred, and therefore whether ω lies in $X^{-1}(B)$. That is, we know whether $\omega \in A$ for each A in $\sigma(X)$, the sigma-field generated by X. Therefore, conditioning on a random variable is really conditioning on the sigma-field generated by that random variable. More generally, we could condition on an arbitrary sigma-field $\mathcal{C} \subset \mathcal{F}$ by knowing, for each $C \in \mathcal{C}$, whether $\omega \in C$. We begin with an elementary setting.

10.1 When There is a Density or Mass Function

This section reviews conditional distribution functions and expectation as usually presented in more elementary probability and statistics courses, under the assumption that the random variables have a joint density or probability mass function. The intent is to motivate the more rigorous definition of conditional expectation and distribution given in subsequent sections.

Let (X, Y) have joint density function or probability mass function $f(x, y)$. The conditional density (or mass function) of Y given $X = x$ is defined as $h(y \,|\, x) = f(x, y)/g(x)$ if $g(x) \neq 0$. It does not matter how we define $h(y \,|\, x)$ when $g(x) = 0$. For each x such that $g(x) > 0$, $h(y \,|\, x)$ is a density function (or mass function) in y, to which there corresponds a conditional distribution function $H(y \,|\, x) = \int_{-\infty}^{y} h(u \,|\, x) du$ or $H(y \,|\, x) = \sum_{u \leq y} h(u \,|\, x)$; $H(y \,|\, x)$ has all of the properties of an ordinary distribution function in y.

If $\mathrm{E}(|Y|) < \infty$, we define the conditional expected value of Y given $X = x$ by

$$\mathrm{E}(Y \,|\, X = x) = \begin{cases} 0 & \text{if } g(x) = 0 \\ \int \frac{y f(x,y)}{g(x)} dy \;\; \textit{or} \;\; \sum_y \frac{y f(x,y)}{g(x)} & \text{if } g(x) \neq 0. \end{cases} \tag{10.2}$$

Note that the definition of $\mathrm{E}(Y \,|\, X = x)$ when $g(x) = 0$ is arbitrary. We could define it to be any fixed number.

Example 10.1. Conditional binomial In a vaccine clinical trial, let N_P and N_V be the numbers of disease events in the placebo and vaccine arms, respectively. A common assumption is that N_P and N_V are independent Poissons with parameters $\lambda_P = \mu_P \sum_{i \in P} T_i$

and $\lambda_V = \mu_V \sum_{i \in V} T_i$, where P and V denote the set of indices for the placebo and vaccine arms, $\sum_{i \in P} T_i$ and $\sum_{i \in V} T_i$ are the total amounts of follow-up time in the two arms, and μ_P and μ_V are the placebo and vaccine rates per unit time. Then $N = N_P + N_V$ is Poisson with parameter $\lambda_P + \lambda_V$. The joint probability mass function of (N, N_P) is

$$\begin{aligned} P(N = n, N_P = n_P) &= P(N_P = n_P \cap N_V = n - n_P) \\ &= \left\{ \frac{\exp(-\lambda_P)\lambda_P^{n_P}}{n_P!} \right\} \left\{ \frac{\exp(-\lambda_V)\lambda_V^{n-n_P}}{(n-n_P)!} \right\} \\ &= \frac{\exp\{-(\lambda_P + \lambda_V)\}\lambda_P^{n_P}\lambda_V^{n-n_P}}{n_P!(n-n_P)!} \end{aligned} \tag{10.3}$$

whenever n_P and n are nonnegative integers and $n_P \le n$. The conditional probability mass function for N_P given $N = n$, namely $f(n, n_P)/g(n)$ is

$$\frac{\exp\{-(\lambda_P + \lambda_V)\}\lambda_P^{n_P}\lambda_V^{n-n_P} n!}{n_P!(n-n_P)!\exp\{-(\lambda_P + \lambda_V)\}(\lambda_P + \lambda_V)^n} = \binom{n}{n_P} \pi^{n_P}(1-\pi)^{n-n_P}, \tag{10.4}$$

where $\pi = \lambda_P/(\lambda_P + \lambda_V)$ and $n \in \{0, 1, 2, \ldots\}$. When $n = 0$, the conditional probability mass function of N_P given $N = n$ is a point mass at 0. We recognize the right side of Equation (10.4) as the binomial probability mass function with n trials and probability parameter π. We have shown that the conditional distribution of the number of disease events in the placebo arm, given the total number of disease events across both arms and the follow-up times in each arm, is binomial (n, π).

The expected value of N_P given $N = n$ is

$$\sum_{i=0}^{n} i \binom{n}{i} \pi^i (1-\pi)^{n-i} = n\pi = n\left(\frac{\lambda_P}{\lambda_P + \lambda_V}\right). \tag{10.5}$$

We can substitute $n = 3$ or $n = 10$ or any other value of n into Equation (10.5) to get the expected value of N_P given $N = n$. But remember that the ultimate goal is to summarize the information contained not just in a particular value of N, but in the random variable N. We do this by substituting N for n into Equation (10.5). This tells us that the expected value of N_P given the random value N is $\mathrm{E}(N_P \mid N) = N\lambda_P/(\lambda_P + \lambda_V)$. Notice that the expected value of N_P given N is a random variable, namely a linear function of N in this example. □

More generally, if we substitute the random variable $X(\omega)$ for the value x in Equation (10.2), we get the random variable

$$Z(\omega) = \begin{cases} 0 & \text{if } g(X(\omega)) = 0 \\ \int \frac{yf(X(\omega),y)}{g(X(\omega))} dy \quad or \quad \sum_y \frac{yf(X(\omega),y)}{g(X(\omega))} & \text{if } g(X(\omega)) \neq 0. \end{cases} \tag{10.6}$$

The random variable Z is a Borel function of X by Fubini's theorem (Theorem 5.28).

We have taken the first big step toward a more rigorous definition of conditional expectation when there is a probability density or mass function, namely conditioning on a random variable rather than on a value of the random variable. The key property of Z defined by (10.6) is that it has the same conditional expectation given $X \in B$ (defined by equation (10.1)) as Y does for all Borel sets B such that $P(X \in B) > 0$. For instance, when (X, Y) are continuous with joint density function $f(x, y)$,

$$\begin{aligned} \mathrm{E}(Z \mid X \in B) &= \frac{1}{P(X \in B)} \int_B \left\{ \int \frac{yf(x,y)}{g(x)} dy \right\} g(x) dx \\ &= \frac{1}{P(X \in B)} \int \int_B yf(x,y) dx dy = \frac{\mathrm{E}\{YI(X \in B)\}}{P(X \in B)} \\ &= \mathrm{E}(Y \mid X \in B). \end{aligned} \tag{10.7}$$

The interchange of order of integration leading to Equation (10.7) is justified by Fubini's theorem because $\mathrm{E}(|Y|) < \infty$ by assumption. Write Equation (10.7) as

$$\frac{\mathrm{E}\{ZI(X \in B)\}}{P(X \in B)} = \frac{\mathrm{E}\{YI(X \in B)\}}{P(X \in B)}$$

and multiply both sides by $P(X \in B) > 0$ to deduce the equivalent condition,

$$\mathrm{E}\{ZI(X \in B)\} = \mathrm{E}\{YI(X \in B)\}. \tag{10.8}$$

In fact, Equation (10.8) holds even if $P(X \in B) = 0$, because in that case both sides are 0.

We have shown that if (X, Y) has joint density function $f(x, y)$, then Equation (10.8) holds. We can also start with Equation (10.8) as the definition of conditional expectation and reproduce Equation (10.2) and therefore (10.6). Equation (10.8) motivates the more rigorous definition of $\mathrm{E}(Y \mid X)$ given in the next section.

Exercises

1. Let X and Y be independent Bernoulli (p) random variables, and let $S = X+Y$. What is the conditional probability mass function of Y given $S = s$ for each of $s = 0, 1, 2$? What is the conditional expected value of Y given the random variable S?

2. Verify directly that in the previous problem, $Z = \mathrm{E}(Y \mid S)$ satisfies Equation (10.8) with X in this expression replaced by S.

3. If X and Y are independent with respective densities $f(x)$ and $g(y)$ and $\mathrm{E}(|Y|) < \infty$, what is $\mathrm{E}(Y \mid X = x)$? What about $Z = \mathrm{E}(Y \mid X)$? Verify directly that Z satisfies Equation (10.8).

4. Let U_1 and U_2 be independent observations from a uniform distribution on $[0, 1]$, and let $X = \min(U_1, U_2)$ and $Y = \max(U_1, U_2)$. What is the joint density function for (X, Y)? Using this density, find $Z = \mathrm{E}(Y \mid X)$. Verify directly that Z satisfies Equation (10.8).

5. Let Y have a discrete uniform distribution on $\{1, -1, 2, -2, \ldots, n, -n\}$. I.e., $P(Y = y) = 1/(2n)$ for $y = \pm i$, $i = 1, \ldots, n$. Define $X = |Y|$. What is $\mathrm{E}(Y \mid X = x)$? What about $Z = \mathrm{E}(Y \mid X)$? Verify directly that Z satisfies Equation (10.8).

6. Notice that Expression (10.2) assumes that (X, Y) has a density with respect to two-dimensional Lebesgue measure or counting measure. Generalize Expression (10.2) to allow (X, Y) to have a density with respect to an arbitrary product measure $\mu_X \times \mu_Y$.

7. $\uparrow$ A mixed Bernoulli distribution results from first observing the value p from a random variable P with density $f(p)$, and then observing a random variable Y from a Bernoulli (p) distribution.

(a) Determine the density function $g(p, y)$ of the pair (P, Y) with respect to the product measure $\mu_L \times \mu_C$, where μ_L and μ_C are Lebesgue measure on $[0, 1]$ and counting measure on $\{0, 1\}$, respectively.

(b) Use your result from the preceding problem to prove that $\mathrm{E}(Y \mid P) = P$ a.s.

10.2 More General Definition of Conditional Expectation

There is only one problem with defining conditional expectations with respect to densities or probability mass functions, as done in Equation (10.2): (X, Y) need not have a density function or probability mass function with respect to a two-dimensional product measure. The development at the end of the preceding section showed us a better way to define conditional expectation. Note that the collection of sets $\{X \in B, B \in \mathcal{B}\}$ is $\sigma(X)$, the sigma-field generated by X, so condition (10.8) can be rephrased as in the following definition.

Definition 10.2. Conditional expectation given a random variable *Let X and Y be random variables on $(\Omega, \mathcal{F}, P)$, with $\mathrm{E}(|Y|) < \infty$. A conditional expected value of Y given X is a random variable $Z(\omega)$ on $(\Omega, \mathcal{F}, P)$ that is measurable with respect to $\sigma(X)$ and satisfies $\mathrm{E}\{ZI(A)\} = \mathrm{E}\{YI(A)\}$ for all $A \in \sigma(X)$. Z is said to be a version of $\mathrm{E}(Y \mid X)$.*

Notice that Definition 10.2 says "**A** conditional expected value" and not "**The** conditional expected value." The definition allows more than one conditional expectation. We can change the value of $\mathrm{E}(Y \mid X)$ on a set $N \in \sigma(X)$ with $P(N) = 0$, and it will still satisfy the definition of conditional expectation.

We have proven the following result when there is a probability density function. The proof for probability mass functions is similar and left as an exercise.

Proposition 10.3. Conditional expectation when there is a density *If (X, Y) has joint density or mass function $f(x, y)$ and X has marginal density or mass function $g(x)$, then one version of $\mathrm{E}(Y \mid X)$ is given by Equation (10.6).*

One of the conditions of Definition 10.2 is that $Z = \mathrm{E}(Y \mid X)$ is $\sigma(X)$-measurable. This means that it is an extended Borel function of X, as we see in the next result.

Proposition 10.4. Conditional expectation as an extended Borel function *Let X be a random variable on $(\Omega, \mathcal{F}, P)$, and suppose that Y is $\sigma(X)$-measurable. Then $Y = \phi(X)$ for some extended Borel function $\phi : R \longmapsto \bar{R}$. Therefore, one version of $\mathrm{E}(Y \mid X)$ is $\phi(X)$ for some extended Borel function ϕ.*

Proof. Assume first that Y is a nonnegative simple random variable on $\sigma(X)$. Then $Y = \sum_{i=1}^{k} a_i I(F_i)$, where each $F_i \in \sigma(X)$. Each F_i is of the form $X^{-1}(B_i)$ for some Borel set $B_i \subset \mathcal{B}$. Then $Y = \sum_{i=1}^{k} a_i I(X \in B_i) = \phi(X)$, where $\phi(x) = \sum_{i=1}^{k} a_i I(x \in B_i)$ is clearly a Borel function.

Now suppose that Y is any nonnegative random variable. Then $Y = \lim_{n\to\infty} Y_n$, where each Y_n is a simple random variable on $\sigma(X)$. By what we just proved, $Y_n = \phi_n(X)$ for Borel functions ϕ_n. This means that $\phi_n\{X(\omega)\} \to Y(\omega)$ for each ω, so $\phi_n(x)$ must converge

to some function $\phi(x)$ for each $x \in \chi$, the range of $X(\omega)$. The problem is that χ need not be an extended Borel set, which means that the function

$$\begin{cases} \lim_{n\to\infty} \phi_n(x)\} & x \in \chi \\ 0 & x \in \chi^C \end{cases} \tag{10.9}$$

need not be an extended Borel function. For example, $\phi(x) = \lim_{n\to\infty} \phi_n(x)$ could be 1 for all $x \in \chi$, in which case Expression (10.9) is not an extended Borel function because $\phi^{-1}(1)$ is not a Borel set. We can avoid this quandary by defining $\phi(x)$ by $\overline{\lim}\, \phi_n(x)$ for all $x \in R$. Proposition 4.9 implies that ϕ is an extended Borel function, and $Y = \phi(X)$.

Now suppose that Y is any random variable on $\sigma(X)$. Then $Y = Y^+ - Y^-$, where Y^+ and Y^- are nonnegative $\sigma(X)$-measurable random variables and $Y^+(\omega) - Y^-(\omega)$ is not of the form $\infty - \infty$. By what we have just proven, $Y^+ = \phi_1(X)$ and $Y^- = \phi_2(X)$ for some extended Borel functions ϕ_1 and ϕ_2 such that $\phi_1(x) - \phi_2(x)$ is not of the form $\infty - \infty$. Then $Y = \phi_1(X) - \phi_2(X)$, and $\phi_1 - \phi_2$ is an extended Borel function. □

Notation 10.5. *If $\phi(X)$ is a conditional expected value $\mathrm{E}(Y \mid X)$ of Y given X, then we write $\mathrm{E}(Y \mid X = x)$ for $\phi(x)$.*

Defining conditional expectation by Definition 10.2 makes clear that it depends on X only through the sigma-field generated by X. This makes sense. If X is a binary random variable, it should not and does not matter whether we condition on X or $1 - X$ because they both give the same information, and Section 4.1.1 taught us to think of information in terms of sigma-fields. The following example further illustrates this point.

Example 10.6. Conditional expectation depends only on the sigma-field generated by a random variable Let (X, Y) have probability mass function $f(x, y)$, and let $g(x)$ be the probability mass function of X, where $g(0) = 0$. Define $U = X^3$. The joint probability mass function of (U, Y) is $P(U = u, Y = y) = P(\{X = u^{1/3}\} \cap \{Y = y\}) = f(u^{1/3}, y)$, and the marginal probability mass function of U is $P(U = u) = P(X = u^{1/3}) = g(u^{1/3})$. It follows that $\sum_y f(U^{1/3}, y)/g(U^{1/3})$ is a version of $\mathrm{E}(Y \mid U)$. But of course $U^{1/3} = X$, so $\sum_y f(X, y)/g(X)$ is a version of $\mathrm{E}(Y \mid U)$. In other words, $\mathrm{E}(Y \mid X)$ is a version of $\mathrm{E}(Y \mid U)$. We are **not** saying that $\mathrm{E}(Y \mid U = u) = \mathrm{E}(Y \mid X = u)$; that is not true. However, as random variables, $\mathrm{E}(Y \mid X^3(\omega))$ and $\mathrm{E}(Y \mid X(\omega))$ are the same function of ω. We get the same information whether we condition on X or on X^3 because X and X^3 generate the same sigma-field. □

The observation that $\mathrm{E}(Y \mid X)$ depends only on the sigma-field generated by X leads us to the following generalization of Definition 10.2.

Definition 10.7. Conditional expectation with respect to a sigma-field *Let Y be a random variable on $(\Omega, \mathcal{F}, P)$ with $\mathrm{E}(|Y|) < \infty$, and let $\mathcal{A} \subset \mathcal{F}$ be a sigma-field. A conditional expectation of Y given $\mathcal{A}$ is an $\mathcal{A}$-measurable random variable $Z(\omega)$ satisfying $\mathrm{E}\{ZI(A)\} = \mathrm{E}\{YI(A)\}$ for all $A \in \mathcal{A}$. If $Y = I(B)$, then $\mathrm{E}(Y \mid \mathcal{A})$ is said to be a conditional probability of B given $\mathcal{A}$.*

Conditioning on a sigma-field $\mathcal{A}$ is very general because $\mathcal{A}$ could be generated by a random variable, a random vector, or even an uncountable collection of random variables. Thus, if $X_1, \ldots, X_n$ and Y are random variables, we could denote the expected value of Y given $X_1, \ldots, X_n$ by either $\mathrm{E}(Y \mid X_1, \ldots, X_n)$ or $\mathrm{E}(Y \mid \mathcal{A})$, where $\mathcal{A} = \sigma(X_1, \ldots, X_n)$ is the sigma-field generated by $X_1, \ldots, X_n$. Likewise, if Y depends on uncountably many variables $X_s,\ s \le t$, then we could denote the expectation of Y given $X_s,\ s \le t$ by either

$\mathrm{E}(Y \mid X_s,\ s \le t)$ or $\mathrm{E}(Y \mid \mathcal{A})$, where $\mathcal{A} = \sigma(X_s,\ s \le t)$ is the sigma-field generated by $X_s,\ s \le t$. In many applications, we condition on a random variable or random vector. Nonetheless, it is just as easy mathematically to treat the more general case of conditioning on an arbitrary sigma-field.

We often surmise the conditional expectation $Z = \mathrm{E}(Y \mid \mathcal{A})$ from general principles, and then prove that Z is $\mathrm{E}(Y \mid \mathcal{A})$ using Definition 10.7. That is, we establish that Z is $\mathcal{A}$-measurable and satisfies $\mathrm{E}\{ZI(A)\} = \mathrm{E}\{YI(A)\}$ for all $A \in \mathcal{A}$. We illustrate this technique with the following example that is the "flip side" of Example 10.6. It shows that two random variables that are "almost the same" can generate very different sigma-fields, and therefore very different conditional expectations.

Example 10.8. Conditioning is not necessarily continuous Let X and Y be independent Bernoulli (1/2) random variables. It is intuitively clear from the independence of X and Y that one version of the conditional expectation $Z = \mathrm{E}(Y \mid X)$ is the unconditional expectation of Y, namely 1/2. We can verify this fact using Definition 10.7 as follows. The first condition is satisfied because the sigma-field generated by the constant 1/2 is $\{\emptyset, \Omega\}$. To verify the second condition, note that each $A \in \sigma(X)$ is either $X^{-1}(0)$, $X^{-1}(1)$, $\emptyset$, or Ω. Consider $A = X^{-1}(0)$. Then $\mathrm{E}\{ZI(A)\} = \mathrm{E}\{(1/2)I(A)\} = (1/2)P(A) = (1/2)P(X = 0)$. By the independence of X and Y, $\mathrm{E}\{YI(A)\} = \mathrm{E}\{YI(X = 0)\} = \mathrm{E}(Y)\mathrm{E}\{I(X = 0)\} = (1/2)P(X = 0)$. Therefore, for $A = X^{-1}(0)$, $\mathrm{E}\{(1/2)I(A)\} = \mathrm{E}\{YI(A)\}$. A similar argument can be used for the other sets in $\sigma(X)$ to show that 1/2 is a version of $\mathrm{E}(Y \mid X)$.

On the other hand, suppose we condition on something "very close to" X, namely $X+(1/1000)Y$. Intuitively, once we know $X+(1/1000)Y$, we know Y because $X+(1/1000)Y$ is an integer if and only if $Y = 0$. Once we know Y and $X + (1/1000)Y$, we also know X. Therefore, conditioning on $X + (1/1000)Y$ fixes the values of X and Y, so $\mathrm{E}\{Y \mid X + (1/1000)Y\}$ must be Y a.s. We can verify this intuition using sigma-fields. Even though $X + (1/1000)Y$ is close to X, the sigma-field generated by $X + (1/1000)Y$ is completely different from that generated by X. The sigma-field generated by X is $\{X^{-1}(0), X^{-1}(1), \emptyset, \Omega\}$. On the other hand, the four values of $X + (1/1000)Y$ corresponding to $(X = 0, Y = 0)$, $(X = 0, Y = 1)$, $(X = 1, Y = 0)$, and $(X = 1, Y = 1)$ are all distinct. Therefore, the sigma-field generated by $X + (1/1000)Y$ is the smallest sigma-field containing the sets $X^{-1}(0) \cap Y^{-1}(0)$, $X^{-1}(0) \cap Y^{-1}(1)$, $X^{-1}(1) \cap Y^{-1}(0)$, and $X^{-1}(1) \cap Y^{-1}(1)$. This sigma-field is $\{X^{-1}(0) \cap Y^{-1}(0), X^{-1}(0) \cap Y^{-1}(1), X^{-1}(1) \cap Y^{-1}(0), X^{-1}(1) \cap Y^{-1}(1), X^{-1}(0), X^{-1}(1), Y^{-1}(0), Y^{-1}(1), \emptyset, \Omega\}$. Notice that this is also the sigma-field generated by (X, Y). Therefore, conditioning on $X + (1/1000)Y$ is the same as conditioning on (X, Y). To verify that $\mathrm{E}\{Y \mid (X, Y)\} = Y$ a.s., note first that Y is clearly measurable with respect to $\sigma(X, Y)$. Also, the second condition of Definition 10.7 is satisfied trivially because $Z = Y$.

We have demonstrated that $\mathrm{E}(Y \mid X) = 1/2$ a.s., yet $\mathrm{E}\{Y \mid X + (1/1000)Y\} = \mathrm{E}\{Y \mid (X, Y)\} = Y$ a.s. In other words, even though X and $X + (1/1000)Y$ are very close to each other, conditioning on X yields a dramatically different answer than conditioning on $X + (1/1000)Y$. The same result obtains if we replace 1/1000 by $1/10^{10}$ or $1/10^{100}$, etc. □

The following result shows that two versions of $\mathrm{E}(Y \mid \mathcal{A})$ can differ only on a set of probability 0.

Proposition 10.9. Existence and almost sure uniqueness of conditional expectation *If* $\mathrm{E}(|Y|) < \infty$*, there is always at least one version of* $\mathrm{E}(Y \mid \mathcal{A})$*. Two versions,* Z_1 *and* Z_2*, of* $\mathrm{E}(Y \mid \mathcal{A})$ *are equal with probability* 1*.*

Proof. The existence part follows from a deep result in analysis called the Radon-Nikodym theorem. See Section 10.9 for details. To prove uniqueness, let Z_1 and Z_2 be two versions of $\mathrm{E}(Y \mid \mathcal{A})$. Because Z_1 and Z_2 are both $\mathcal{A}$-measurable, $A = I(Z_1 - Z_2 > 0) \in \mathcal{A}$. Also, Z_1 and Z_2 are integrable. Therefore,

$$\begin{aligned} \mathrm{E}\{(Z_1 - Z_2)I(A)\} &= \mathrm{E}\{Z_1 I(A)\} - \mathrm{E}\{Z_2 I(A)\} \\ &= \mathrm{E}\{Y I(A)\} - \mathrm{E}\{Y I(A)\} = 0. \end{aligned} \tag{10.10}$$

But $Z_1 - Z_2$ is strictly positive on A, so $\mathrm{E}\{(Z_1 - Z_2)I(A) = 0\}$ implies that $P(A) = 0$. That is, $P(Z_1 > Z_2) = 0$. The same argument with Z_1 and Z_2 reversed shows that $P(Z_2 > Z_1) = 0$. Thus, $P(Z_1 = Z_2) = 1$. □

Certain results for conditional expectation follow almost immediately from the definition. For instance, if we take $A = \Omega$, then $\mathrm{E}(Z) = \mathrm{E}\{ZI(\Omega)\} = \mathrm{E}\{YI(\Omega)\} = \mathrm{E}(Y)$. We have proven the following.

Proposition 10.10. Computing expectations by first computing conditional expectations and then "unconditioning" *If* $\mathrm{E}(|Y|) < \infty$*, then* $\mathrm{E}\{\mathrm{E}(Y \mid \mathcal{A})\} = \mathrm{E}(Y)$.

Example 10.11. Assume that whether a patient in a clinical trial experiences a drug-related adverse event is a Bernoulli random variable, but that patient i has his or her own Bernoulli parameter p_i. We can imagine the Bernoulli parameters for different patients as random draws from a distribution $F(p)$. If Y is the indicator that a randomly selected patient experiences an adverse event, then $P(Y = 1 \mid P = p)$ is Bernoulli (p), so $\mathrm{E}(Y \mid P) = P$. By Proposition 10.10, the probability that a randomly selected patient has an adverse event is $\mathrm{E}(Y) = \mathrm{E}\{E(Y \mid P)\} = \mathrm{E}(P) = \int p dF(p)$. □

Example 10.12. In imaging studies of the lungs, we may express the burden of a disease by the total volume of diseased lesions. Let N be the number of lesions for a given patient, and Y_i be the volume of lesion i. Assuming that the number of lesions is small, it may be reasonable to assume that N is bounded and independent of the Ys (this probably would not be reasonable if N is large, in which case the larger the number of lesions, the smaller their volumes must be). Let μ_Y and μ_N be the (finite) means of Y and N, respectively. The disease burden is $S_N = \sum_{i=1}^{N} Y_i$. Also, because N is bounded by some integer B, $|S_N| \leq \sum_{i=1}^{B} |Y_i|$. Therefore, $\mathrm{E}(|S_N|) < \infty$. To find $\mathrm{E}(S_N)$, first condition on $N = n$. Then $S_N = S_n$, the sum of n independent observations, each with mean μ_Y. The conditional mean $\mathrm{E}(S_N \mid N = n)$ is $n\mu_Y$, so $\mathrm{E}(S_N) = \mathrm{E}\{\mathrm{E}(S_N \mid N)\} = \mathrm{E}(N\mu_Y) = \mu_N \mu_Y$. □

Proposition 10.13. Elementary properties of conditional expectation *Let Y and Y_n, $n = 1, 2, \ldots$ be integrable random variables on $(\Omega, \mathcal{F}, P)$, and let $\mathcal{A} \subset \mathcal{F}$ be a sigma-field. Then*

1. $\mathrm{E}(c_1 Y_1 + c_2 Y_2 \mid \mathcal{A}) = c_1 \mathrm{E}(Y_1 \mid \mathcal{A}) + c_2 \mathrm{E}(Y_2 \mid \mathcal{A})$ *a.s.*
2. *If* $P(Y_1 \leq Y_2) = 1$*, then* $\mathrm{E}(Y_1 \mid \mathcal{A}) \leq \mathrm{E}(Y_2 \mid \mathcal{A})$ *a.s.*
3. *If* $Y_n \uparrow Y$ *a.s., then* $\mathrm{E}(Y_n \mid \mathcal{A}) \uparrow \mathrm{E}(Y \mid \mathcal{A})$ *a.s.*
4. *If* $Y_n \downarrow Y$ *a.s., then* $\mathrm{E}(Y_n \mid \mathcal{A}) \downarrow \mathrm{E}(Y \mid \mathcal{A})$ *a.s.*
5. **DCT for conditional expectation** *If* $Y_n \to Y$ *a.s. and* $|Y_n| \leq U$ *a.s., where* $\mathrm{E}(U) < \infty$*, then* $\mathrm{E}(Y_n \mid \mathcal{A}) \to \mathrm{E}(Y \mid \mathcal{A})$ *a.s.*

Proof. The general method of proof for conditional expectation of Y given $\mathcal{A}$ is to show that the candidate random variable Z is $\mathcal{A}$-measurable and has the same expectation as Z over sets $A \in \mathcal{A}$.

For part 1, note that $c_1\mathrm{E}(Y_1 \mid \mathcal{A}) + c_2\mathrm{E}(Y_2 \mid \mathcal{A})$ is $\mathcal{A}$-measurable because $\mathrm{E}(Y_1 \mid \mathcal{A})$ and $\mathrm{E}(Y_2 \mid \mathcal{A})$ are $\mathcal{A}$-measurable. Moreover,

$$\begin{aligned} & \int_A \{c_1\mathrm{E}(Y_1 \mid \mathcal{A}) + c_2\mathrm{E}(Y_2 \mid \mathcal{A})\}dP(\omega) = c_1 \int_A \mathrm{E}(Y_1 \mid \mathcal{A})dP(\omega) + c_2 \int_A \mathrm{E}(Y_2 \mid \mathcal{A})dP(\omega) \\ = \quad & c_1 \int_A Y_1 dP(\omega) + c_2 \int_A Y_2 dP(\omega) = \int_A (c_1 Y_1 + c_2 Y_2)dP(\omega) \end{aligned} \tag{10.11}$$

for each $A \in \mathcal{A}$, proving part 1.

For parts 3 and 4, we prove the results first for nonnegative random variables. For example, for part 3, $Z_n = \mathrm{E}(Y_n \mid \mathcal{A})$ are $\mathcal{A}$-measurable random variables and are increasing by part 2. Therefore, the limit $Z_\infty = \lim_{n\to\infty} Z_n$ exists (Proposition A.33) and is $\mathcal{A}$-measurable (Proposition 4.9). We will demonstrate that $\int Z_\infty I(A)dP(\omega) = \int Y(\omega)I(A)dP(\omega)$ for each $A \in \mathcal{A}$, which will show that Z_∞ satisfies the definition of $\mathrm{E}(Y \mid \mathcal{A})$. Note that $Z_n I(A) \uparrow Z_\infty I(A)$ a.s. and is nonnegative, so the MCT implies that $\mathrm{E}\{Z_n I(A)\} \to \mathrm{E}\{Z_\infty I(A)\}$. The MCT also implies that $\mathrm{E}\{Y_n I(A)\} \to \mathrm{E}\{Y I(A)\}$. Therefore, for each $A \in \mathcal{A}$,

$$\mathrm{E}\{Z_\infty I(A)\} = \lim_{n\to\infty} \mathrm{E}\{Z_n I(A)\} = \lim_{n\to\infty} \mathrm{E}\{Y_n I(A)\} = \mathrm{E}\{Y I(A)\}. \tag{10.12}$$

A similar argument shows that part 4 holds for nonnegative random variables. To prove parts 3 and 4 for arbitrary Y_n, write Y_n as $Y_n^+ - Y_n^-$ and use the fact that $Y_n^+ \uparrow Y^+$ and $Y_n^- \downarrow Y^-$.

Proofs of the remaining parts are left as exercises. □

When we condition on information that fixes the value of a random variable, we can essentially treat that random variable as a constant. For instance, when we condition on Y_1, then Y_1Y_2 behaves as if Y_1 were a constant: $\mathrm{E}(Y_1Y_2 \mid Y_1) = Y_1\mathrm{E}(Y_2 \mid Y_1)$ when the expectations exist. Specifically:

Proposition 10.14. Treating known random variables as constants *If* $\mathrm{E}(|Y_2|) < \infty$ *and* $\mathrm{E}(|Y_1Y_2|) < \infty$ *and* Y_1 *is* $\mathcal{A}$*-measurable, then* $\mathrm{E}(Y_1Y_2 \mid \mathcal{A}) = Y_1\mathrm{E}(Y_2 \mid \mathcal{A})$ *almost surely.*

Proof. First, $Y_1\mathrm{E}(Y_2 \mid \mathcal{A})$ is $\mathcal{A}$-measurable because Y_1 is $\mathcal{A}$-measurable by assumption, and $\mathrm{E}(Y_2 \mid \mathcal{A})$ is $\mathcal{A}$-measurable by definition of conditional expectation. It remains to prove that if $A \in \mathcal{A}$, $\int Y_1\mathrm{E}(Y_2 \mid \mathcal{A})I(A)dP(\omega) = \int Y_1Y_2I(A)dP(\omega)$. We show first that this holds if Y_1 is a simple random variable.

If $Y_1 = \sum_{i=1}^n a_i I(A_i)$, $A_i \in \mathcal{A}$, then $Y_1\mathrm{E}(Y_2 \mid \mathcal{A})I(A) = \sum_{i=1}^n a_i I(A_i \cap A)\mathrm{E}(Y_2 \mid \mathcal{A})$, and $A_i \cap A \in \mathcal{A}$. Moreover, each $I(A_i \cap A)\mathrm{E}(Y_2 \mid \mathcal{A})$ is integrable by definition of $\mathrm{E}(Y_2 \mid \mathcal{A})$. It follows that for each $A \in \mathcal{A}$,

$$\begin{aligned} \int Y_1\mathrm{E}(Y_2 \mid \mathcal{A})I(A)dP(\omega) &= \int \sum_{i=1}^n a_i I(A_i \cap A)\mathrm{E}(Y_2 \mid \mathcal{A})dP(\omega) \\ &= \sum_{i=1}^n a_i \int I(A_i \cap A)\mathrm{E}(Y_2 \mid \mathcal{A})dP(\omega) \\ &= \sum_{i=1}^n a_i \int I(A_i \cap A)Y_2\, dP(\omega) \end{aligned}$$

$$\begin{aligned} &= \int \left\{ \sum_{i=1}^{n} a_i I(A_i) \right\} Y_2 \, I(A) dP(\omega) \\ &= \int Y_1 Y_2 \, I(A) dP(\omega), \end{aligned} \tag{10.13}$$

proving the result when Y_1 is a nonnegative simple random variable.

If Y_1 is any nonnegative random variable, then $Y_1 = \lim_{n\to\infty} U_n$, where U_n are simple, $\mathcal{A}$-measurable random variables increasing to Y_1 (see Section 5.2.1). Then for each $A \in \mathcal{A}$, $\int Y_1 \mathrm{E}(Y_2 \,|\, \mathcal{A}) I(A) dP(\omega)$ is

$$\begin{aligned} &= \int Y_1 \mathrm{E}(Y_2^+ \,|\, \mathcal{A}) I(A) dP(\omega) - \int Y_1 \mathrm{E}(Y_2^- \,|\, \mathcal{A}) I(A) dP(\omega) \text{ (if finite)} \\ &= \int \lim_{n\to\infty} \{U_n \mathrm{E}(Y_2^+ \,|\, \mathcal{A}) I(A)\} dP(\omega) - \int \lim_{n\to\infty} \{U_n \mathrm{E}(Y_2^- \,|\, \mathcal{A}) I(A)\} dP(\omega) \\ &= \lim_{n\to\infty} \int \{U_n \mathrm{E}(Y_2^+ \,|\, \mathcal{A}) I(A)\} dP(\omega) - \lim_{n\to\infty} \int \{U_n \mathrm{E}(Y_2^- \,|\, \mathcal{A}) I(A)\} dP(\omega) \quad \text{(MCT)} \\ &= \lim_{n\to\infty} \int \{U_n Y_2^+ I(A)\} dP(\omega) - \lim_{n\to\infty} \int \{U_n Y_2^- I(A)\} dP(\omega) \quad \text{(result for simple r.v.s)} \\ &= \int \lim_{n\to\infty} \{U_n Y_2^+ I(A)\} dP(\omega) - \int \lim_{n\to\infty} \{U_n Y_2^- I(A)\} dP(\omega) \quad \text{(MCT)} \\ &= \int \{Y_1 Y_2^+ I(A)\} dP(\omega) - \int \{Y_1 Y_2^- I(A)\} dP(\omega) \\ &= \int \{Y_1 Y_2 I(A)\} dP(\omega). \end{aligned} \tag{10.14}$$

The first line is finite if and only if the last line is finite, and the last line is finite because $\mathrm{E}(|Y_1 Y_2|)$ is assumed finite. Thus, the result holds when Y_1 is any nonnegative $\mathcal{A}$-measurable random variable.

If Y_1 is any $\mathcal{A}$-measurable random variable, then $Y_1 = Y_1^+ - Y_1^-$, and Y_1^+ and Y_1^- are nonnegative and $\mathcal{A}$-measurable. By what we have just proven, $\int_A Y_1^+ \mathrm{E}(Y_2 \,|\, \mathcal{A}) dP(\omega) = \int_A Y_1^+ Y_2 dP(\omega)$. Similarly, $\int_A Y_1^- \mathrm{E}(Y_2 \,|\, \mathcal{A}) dP(\omega) = \int_A Y_1^- Y_2 dP(\omega)$ for each $A \in \mathcal{A}$. Therefore,

$$\begin{aligned} \int_A Y_1 \mathrm{E}(Y_2 \,|\, \mathcal{A}) dP(\omega) &= \int_A Y_1^+ \mathrm{E}(Y_2 \,|\, \mathcal{A}) dP(\omega) - \int_A Y_1^- \mathrm{E}(Y_2 \,|\, \mathcal{A}) dP(\omega) \\ &= \int_A Y_1^+ Y_2 dP(\omega) - \int_A Y_1^- Y_2 dP(\omega) \\ &= \int_A Y_1 Y_2 dP(\omega) \end{aligned} \tag{10.15}$$

for each $A \in \mathcal{A}$, completing the proof. □

Exercises

1. Let Y have a discrete uniform distribution on $\{\pm 1, \pm 2, \ldots, \pm n\}$, and let $X = Y^2$. Find $\mathrm{E}(Y \,|\, X = x)$. Does $\mathrm{E}(Y \,|\, X = x)$ match what you got for $\mathrm{E}(Y \,|\, X = x)$ for $X = |Y|$ in Problem 5 in the preceding section? Now compute $Z = \mathrm{E}(Y \,|\, X)$ and compare it with your answer for $\mathrm{E}(Y \,|\, X)$ in Problem 5 in the preceding section.

2. Let Y be as defined in the preceding problem, but let $X = Y^3$ instead of Y^2. Find $\mathrm{E}(Y \,|\, X = x)$ and $Z = \mathrm{E}(Y \,|\, X)$. Does Z match your answer in the preceding problem?

3. Tell whether the following is true or false. If it is true, prove it. If it is false, give a counterexample. If $\mathrm{E}(Y \mid X_1) = \mathrm{E}(Y \mid X_2)$, then $X_1 = X_2$ almost surely.

4. Let Y be a random variable defined on $(\Omega, \mathcal{F}, P)$ with $\mathrm{E}(|Y|) < \infty$. Verify the following using Definition 10.2.

   (a) If $\mathcal{A} = \{\Omega, \emptyset\}$, then $\mathrm{E}(Y \mid \mathcal{A}) = \mathrm{E}(Y)$ a.s.

   (b) If $\mathcal{A} = \sigma(Y)$, then $\mathrm{E}(Y \mid \mathcal{A}) = Y$ a.s.

5. Let $X_1, \ldots, X_n$ be iid with $\mathrm{E}(|X_i|) < \infty$. Prove that $\mathrm{E}(X_1 \mid \bar{X}) = \bar{X}$ a.s. Hint: it is clear that $\mathrm{E}\{(1/n)\sum_{i=1}^n X_i \mid \bar{X}\} = \bar{X}$ a.s.

6. If Y is a random variable with $\mathrm{E}(|Y|) < \infty$, and g is a Borel function, then $\mathrm{E}\{Y \mid X, g(X)\} = \mathrm{E}(Y \mid X)$ a.s.

7. Suppose that X is a random variable with mean 0 and variance $\sigma^2 < \infty$, and assume that $\mathrm{E}(Y \mid X) = X$ a.s. Find $\mathrm{E}(XY)$.

8. Prove Proposition 10.3 when (X, Y) has a joint probability mass function.

9. Prove part 2 of Proposition 10.13.

10.3 Regular Conditional Distribution Functions

We defined the conditional probability of an event B given the sigma-field $\mathcal{A}$ as $\mathrm{E}\{I(B) \mid \mathcal{A}\}$. Therefore, for any random variable Y and value y, the conditional probability that $Y \le y$ given $\mathcal{A}$ is $P(Y \le y \mid \mathcal{A}) = \mathrm{E}\{I(Y \le y) \mid \mathcal{A}\}$. The question is: will this result in a distribution function in y for fixed ω? That is, will an arbitrary version of $\mathrm{E}\{I(Y \le y) \mid \mathcal{A}\}$ be a distribution function in y? To see that the answer is no even in a simple setting, let Y have a standard normal distribution and $X = 0$ with probability 1. One version of $\mathrm{E}\{I(Y \le y) \mid X\}$ is $\Phi(y)$, which is, of course, a distribution function in y. On the other hand, another version of $\mathrm{E}\{I(Y \le y \mid X\}$ is

$$\begin{cases} \Phi(y) & \text{if } X \neq 0 \\ 1 & \text{if } X = 0 \text{ and } y = 0, \\ \Phi(y) & \text{if } X = 0 \text{ and } y \neq 0, \end{cases}$$

and this is not monotone in y if $X = 0$. Therefore, using just any version of $\mathrm{E}\{I(Y \le y) \mid \mathcal{A}\}$ does not guarantee that it will be a distribution function in y for each ω. Nonetheless, we can always find a version of $\mathrm{E}\{I(Y \le y) \mid \mathcal{A})$ that is a distribution function in y. Such a function is called a *regular conditional distribution function of Y given $\mathcal{A}$.*

Theorem 10.15. Distribution function of Y given $\mathcal{A}$ *Let Y be a random variable on $(\Omega, \mathcal{F}, P)$ and $\mathcal{A} \subset \mathcal{F}$ be a sigma-field. There exists a version $F(y, \omega)$ of $\mathrm{E}\{I(Y \le y) \mid \mathcal{A})$ that is a regular conditional distribution function of Y given $\mathcal{A}$.*

Proof. Except on a null set N_1, the following conditions hold for rational r.

$$F(r, \omega) = \mathrm{E}\{I(Y \le r) \mid \mathcal{A}\} \uparrow \text{ in } r, \; F(r, \omega) \to 0 \text{ or } 1 \text{ as } r \to -\infty \text{ or } \infty. \tag{10.16}$$

To see the monotonicity property, note that for each pair $r_1 < r_2$ of rationals, the set $B(r_1, r_2) = \{\omega : \; F(r_1, \omega) > F(r_2, \omega)$ has probability 0 by part 2 of Proposition 10.13. The set of pairs (r_1, r_2) of rational numbers such that $F(r_1, \omega) > F(r_2, \omega)$ is the countable union, $\cup_{r_1, r_2} B(r_1, r_2)$, of sets of probability 0, so $P\{\cup_{r_1, r_2} B(r_1, r_2)\} \le \sum_{r_1, r_2} P\{B(r_1, r_2)\} = 0$.

This shows that except on a null set, $F(r,\omega)$ is monotone increasing in r. The limits as $r \to \pm\infty$ follow from properties 3 and 4 of Proposition 10.13 because $I(Y \le r)$ converges almost surely to 0 or 1 as $r \to -\infty$ or ∞, respectively. Thus, except on a null set N_1, conditions (10.16) are satisfied.

Except on another null set N_2,

$$F(r + 1/n, \omega) \to F(r, \omega) \text{ as } n \to \infty \tag{10.17}$$

for all rational numbers r. To see this, note that for a particular r, part 4 of Proposition 10.13 implies that condition (10.17) holds except on a null set $N_2(r)$. The set of ω for which condition (10.17) fails to hold for at least one rational r is the countable union $N_2 = \cup_r N_2(r)$ of null sets, so $P(N_2) \le \sum_r P(N_2(r)) = 0$. Therefore, outside the null set N_2, condition (10.17) holds.

We have defined the candidate distribution function $F(r,\omega)$ on the rational numbers in a way that, except for ω in the null set $N = N_1 \cup N_2$, conditions (10.16) and (10.17) hold. We now define $F(y,\omega)$ for $\omega \in N^C$ and y irrational: $F(y,\omega) = \underline{\lim}_{r>y} F(r,\omega)$. Then $F(y,\omega)$, being a liminf of $\mathcal{A}$-measurable random variables, is also $\mathcal{A}$-measurable. For each y, whether rational or irrational, there is a sequence of rational numbers r_n decreasing to y such that $F(r_n,\omega) \to F(r,\omega)$ as $n \to \infty$. This fact can be used to show that $F(y,\omega)$ is right-continuous on N^C for each y. To see this, let $y_n \downarrow y$. We can find a rational number $r_n \ge y_n$ such that

$$0 \le F(r_n,\omega) - F(y_n,\omega) < 1/n. \tag{10.18}$$

Therefore, $F(r_n,\omega) - 1/n < F(y_n,\omega) \le F(r_n,\omega)$ for all n. The limits of the left and right sides as $n \to \infty$ are both $F(y,\omega)$, so $F(y_n,\omega) \to F(y,\omega)$ as $n \to \infty$. We have established that $F(y,\omega)$ is right continuous. It is also easy to show that $\lim_{y\to-\infty} F(y,\omega) = 0$, $\lim_{y\to\infty} F(y,\omega) = 1$.

For $\omega \in N$, take $F(y,\omega)$ to be any fixed distribution function, such as $\mathrm{N}(0,1)$. □

Notation 10.16. *When $\mathcal{A}$ is the sigma-field generated by a random variable X, the conditional distribution function $F(y,\omega)$ of Y given $\mathcal{A} = \sigma(X)$ is a Borel function of X. This follows from Proposition 10.4 and the fact that $F(y,\omega)$ is a version of $\mathrm{E}\{I(Y \le y \,|\, X\}$. Therefore, we sometimes denote the conditional distribution of Y given $X = x$ by $F(y \,|\, x)$.*

It seems like we have attained all we need. After all, the distribution function for a random variable Y determines the probability that $Y \in B$ for every Borel set B (see Proposition 4.22). There is only one glitch: we have shown only that the conditional distribution function $F(y,\omega)$ is $\mathcal{A}$-measurable. How can we be sure that $P(Y \in B \,|\, \mathcal{A})$ is $\mathcal{A}$-measurable for an arbitrary Borel set B? Fortunately, it is.

Theorem 10.17. Probability measure of Y given $\mathcal{A}$ *Let Y be a random variable on $(\Omega, \mathcal{F}, P)$ and $\mathcal{A} \subset \mathcal{F}$ be a sigma-field. There exists a probability measure $\mu(B,\omega)$ defined on Borel subsets B such that $\mu(B,\omega)$ is a version of $\mathrm{E}\{I(Y \in B) \,|\, \mathcal{A}\}$.*

Proof. We use the notation $F_\omega(y)$ for $F(y,\omega)$ to emphasize that we are fixing ω and regarding F as a function of y. Define $\mu(B,\omega)$ by $\int_B dF_\omega(y)$. It is an exercise to show that the set $\mathcal{B}'$ of Borel sets B such that $\mu(B,\omega)$ is $\mathcal{A}$-measurable is a monotone class containing the field in Proposition 3.8. By the monotone class theorem (Theorem 3.32), $\mathcal{B}'$ contains all Borel sets. □

The reason for defining a conditional probability measure for Y given $\mathcal{A}$ is to facilitate the calculation of conditional expected values of functions of Y given $\mathcal{A}$.

Proposition 10.18. Using conditional distribution functions to compute conditional expectations *If $F(y,\omega)$ is a regular conditional distribution function of Y given $\mathcal{A}$ and $g(Y)$ is a Borel function with $\mathrm{E}\{|g(Y)|\} < \infty$, then $\int g(y)dF(y,\omega)$ is a version of $\mathrm{E}\{g(Y)\,|\,\mathcal{A}\}$.*

The proof follows the familiar pattern of beginning with nonnegative simple g, then extending to all nonnegative Borel functions, then to all Borel functions (exercise).

Example 10.19. Simon and Simon, 2011: rerandomization tests protect type I error rate There are many different ways to randomize patients to treatment (T) or control (C) in a clinical trial, including: (1) simple randomization, akin to flipping a fair coin for each new patient, (2) permuted block randomization, whereby k patients in each block of size $2k$ are assigned to T, the other k to C, (3) Efron's biased coin design, whereby a fair coin is used whenever the numbers of Ts and Cs are balanced, and an unfair coin with probability, say 2/3, favoring the under-represented treatment when the numbers of Ts and Cs are unbalanced, and (4) various covariate-adaptive schemes making it more likely that the treatment assigned to the next patient balances the covariate distributions across the arms.

A very general principle is to "analyze as you randomize." To do this, treat all data $\mathcal{D}$ in the clinical trial other than the treatment labels as fixed constants and re-generate the randomization sequence using whatever method was used to generate the original labels. For this rerandomized dataset, compute the value of the test statistic T. Repeat this process of rerandomizing the patients and computing the test statistic until all possible rerandomizations have been included. This generates the rerandomization distribution $F(t)$ of T. For a one-tailed test rejecting the null hypothesis for large values of T, determine $c_* = \inf\{c : 1 - F(c) \leq \alpha\}$. By the right-continuity of distribution functions, $1 - F(c_*) \leq \alpha$. Reject the null hypothesis if T_{orig} corresponding to the original randomization exceeds c_*. This test is called a rerandomization test. This is a generalization of a permutation test that accommodates any randomization scheme.

Let $\mathcal{D}$ be the sigma-field generated by all data other than the treatment labels. The rerandomization distribution is the conditional distribution function of T given $\mathcal{D}$. By construction, $1 - F(c_*) = P(T > c_* \,|\, \mathcal{D}) \leq \alpha$. In other words, the conditional type I error rate given the data $\mathcal{D}$ is no greater than α. By Proposition 10.10, the unconditional type I error rate, $P(T > c_*)$, is simply $\mathrm{E}[\mathrm{E}\{I(T > c_*) \,|\, \mathcal{D})] \leq \alpha$. Therefore, any rerandomization test controls the type I error rate both conditional on the observed data and unconditionally. □

Example 10.20. Asymptotic equivalence of rerandomization test and stratified t-test under permuted block randomization Consider a clinical trial using random permuted blocks of size 4 to assign patients to treatment (T) or control (C). Let X_{i1}, X_{i2}, X_{i3}, and X_{i4} be the observations on a continuous outcome for patients in block i. The treatment less control difference in block i is $D_i = \sum_{j=1}^{4} X_{ij} Z_{ij}$, where Z_{ij} is $+1$ if patient j of block i is assigned to treatment, and -1 if assigned to control. Let $\mathcal{A} = \sigma(X_{ij},\ i = 1, 2, \ldots,\ j = 1, \ldots, 4)$ be the infinite set of data from which the first $4n$ patients constitute the clinical trial. Let $T_n = \sum_{i=1}^{n} D_i / (\sum_{i=1}^{n} D_i^2)^{1/2}$ be the test statistic, and suppose that we reject the null hypothesis for large values of T_n.

The rerandomization distribution is the conditional distribution of T_n given $\mathcal{A}$. We claim that under the null hypothesis, this conditional distribution converges to $\mathrm{N}(0,1)$ as $n \to \infty$. To see this, first compute the conditional distribution $F_n(t \,|\, \mathcal{C})$, where $\mathcal{C}$ is the sigma-field $\sigma(X_{ij},\ i = 1, 2, \ldots,\ j = 1, \ldots, 4,\ |D_1|, |D_2|, \ldots)$. This conditional distribution is the distribution of $\sum_{i=1}^n \delta_i d_i / (\sum_{i=1}^n d_i^2)^{1/2}$, where the d_i are fixed constants and δ_i are iid random variables taking values ± 1 with probability $1/2$. We have already seen in Example 8.20 that the conditional distribution $F_n(t \,|\, \mathcal{C})$ of T_n given $\mathcal{C}$ converges to $\mathrm{N}(0,1)$ almost surely. Denoting the conditional distribution function of T_n given $\mathcal{A}$ by $F_n(t \,|\, \mathcal{A})$, we have

$$\begin{aligned} F_n(t \,|\, \mathcal{A}) &= \mathrm{E}\{I(T_n \le t \,|\, \mathcal{A})\} = \mathrm{E}[\mathrm{E}\{I(T_n \le t) \,|\, \mathcal{C}\} \,|\, \mathcal{A}] \\ &\to \mathrm{E}\{\Phi(t) \,|\, \mathcal{A}\} = \Phi(t) \text{ a.s.} \end{aligned} \tag{10.19}$$

The second line follows from the DCT for conditional expectation because $\mathrm{E}\{I(T_n \le t) \,|\, \mathcal{C}\}$ is dominated by 1. We have thus shown that the permutation test is asymptotically equivalent to rejecting the null hypothesis when $\sum_{i=1}^n D_i / (\sum_{i=1}^n D_i^2)^{1/2} > z_\alpha$, where z_α is the $(1-\alpha)$th quantile of the standard normal distribution. Under the null hypothesis, $(1/n)\sum_{i=1}^n D_i^2 \to \sigma^2 = \mathrm{var}(X_i)$. It follows that the rerandomization test is asymptotically equivalent to rejecting the null hypothesis if $\sum_{i=1}^n D_i / (n\sigma^2)^{1/2} > z_\alpha$. Also, the difference between the usual t-statistic and $\sum_{i=1}^n D_i / (n\sigma^2)^{1/2}$ converges almost surely to 0, so the rerandomization test is asymptotically equivalent to the t-test under random permuted block randomization.

Proposition 10.21. Inequalities for conditional expectation *Let X and Y be random variables defined on $(\Omega, \mathcal{F}, P)$, and let $\mathcal{A} \subset \mathcal{F}$ be a sigma-field.*

1. *Jensen's inequality: If $\phi(\cdot)$ is convex and Y and $\phi(Y)$ are integrable, then $\mathrm{E}\{\phi(Y) \,|\, \mathcal{A}\} \ge \phi\{\mathrm{E}(Y \,|\, \mathcal{A})\}$ a.s.*
2. *Markov's inequality: If C is any $\mathcal{A}$-measurable random variable, then $P(|Y| \ge C \,|\, \mathcal{A}) \le (1/C)\mathrm{E}(|Y| \,\big|\, \mathcal{A})$ a.s.*
3. *Chebychev's inequality: If $\mathrm{E}(Y^2) < \infty$, and C is $\mathcal{A}$-measurable, then $P\{|Y - \mathrm{E}(Y \,|\, \mathcal{A})| \ge C \,|\, \mathcal{A}\} \le (1/C^2)\mathrm{var}(Y \,|\, \mathcal{A})$ a.s.*
4. *Hölder's inequality: if $p > 0$, $q > 0$, $1/p + 1/q = 1$, and $\mathrm{E}(|X|^p) < \infty$, $\mathrm{E}(|Y|^q) < \infty$, then $\mathrm{E}(|XY| \,\big|\, \mathcal{A}) \le \{\mathrm{E}(|X|^p \,|\, \mathcal{A})\}^{1/p}\{\mathrm{E}(|Y|^q \,|\, \mathcal{A})\}^{1/q}$ a.s.*
5. *Schwarz's inequality: If X and Y are random variables with $\mathrm{E}(X^2) < \infty$, $\mathrm{E}(Y^2) < \infty$, then $\mathrm{E}(|XY| \,\big|\, \mathcal{A}) \le \sqrt{\mathrm{E}(X^2 \,|\, \mathcal{A})\mathrm{E}(Y^2 \,|\, \mathcal{A})}$ a.s.*
6. *Minkowski's inequality: If $p \ge 1$ and $\mathrm{E}(|X|^p) < \infty$ and $\mathrm{E}(|Y|^p) < \infty$, then $\{\mathrm{E}(|X + Y|^p \,|\, \mathcal{A})\}^{1/p} \le \{\mathrm{E}(|X|^p \,|\, \mathcal{A})\}^{1/p} + \{\mathrm{E}(|Y|^p \,|\, \mathcal{A})\}^{1/p}$ a.s.*

Proof. We prove only the first result. Proofs of the others are similar and left as exercises. Let $F_\omega(y)$ be a regular conditional distribution function of Y given $\mathcal{A}$. Because $F_\omega(y)$ is a distribution function in y for fixed ω, the usual Jensen's inequality implies that $\int \phi(y) dF_\omega(y) \ge \phi\{\int y dF_\omega(y)\}$ for each ω. But $\int \phi(y) dF_\omega(y)$ and $\int y dF_\omega(y)$ are versions of $\mathrm{E}\{\Phi(Y) \,|\, \mathcal{A}\}$ and $\mathrm{E}(Y \,|\, \mathcal{A})$, respectively, from which the result follows. □

Notice that with the Markov and Chebychev inequality for conditional expectation, C is allowed to be any $\mathcal{C}$-measurable random variable, whereas with the usual Markov and Chebychev inequalities, C is a constant. Remember that once we condition on $\mathcal{A}$, any $\mathcal{A}$-measurable random variable becomes constant.

Just as we defined the conditional mean of a random variable given a sigma-field $\mathcal{A}$, we can define conditional variances and covariances of random variables given $\mathcal{A}$.

Definition 10.22. Conditional variances and covariances *Let $\mathcal{A} \subset \mathcal{F}$ be a sigma-field.*

1. *If Y is a random variable with $\mathrm{E}(Y^2) < \infty$, the conditional variance of Y given $\mathcal{A}$ is $\mathrm{var}(Y \mid \mathcal{A}) = \mathrm{E}\{(Y - Z)^2 \mid \mathcal{A}\}$, where Z is a version of $\mathrm{E}(Y \mid \mathcal{A})$.*
2. *If Y_1 and Y_2 are random variables with $\mathrm{E}(Y_1^2) < \infty$ and $\mathrm{E}(Y_2^2) < \infty$, the conditional covariance of Y_1 and Y_2 given $\mathcal{A}$ is $\mathrm{cov}(Y_1, Y_2 \mid \mathcal{A}) = \mathrm{E}\{(Y_1 - Z_1)(Y_2 - Z_2) \mid \mathcal{A}\}$, where Z_i is a version of $\mathrm{E}(Y_i \mid \mathcal{A})$, $i = 1, 2$.*

The following are some elementary properties of conditional variances and covariances. Other important identities are presented in the next section.

Proposition 10.23. Elementary properties of conditional variances and covariances *Suppose that Y, Y_1, and Y_2 are random variables with finite second moments, and $\mathcal{A} \subset \mathcal{F}$ is a sigma-field.*

1. $\mathrm{var}(Y \mid \mathcal{A}) = \mathrm{E}(Y^2 \mid \mathcal{A}) - \{\mathrm{E}(Y \mid \mathcal{A})\}^2$ *a.s.*
2. $\mathrm{cov}(Y_1, Y_2 \mid \mathcal{A}) = \mathrm{E}(Y_1 Y_2 \mid \mathcal{A}) - \{\mathrm{E}(Y_1 \mid \mathcal{A})\}\{\mathrm{E}(Y_2 \mid \mathcal{A})\}$ *a.s.*
3. *If C is an $\mathcal{A}$-measurable random variable with $\mathrm{E}(C^2) < \infty$, then $\mathrm{var}(Y + C \mid \mathcal{A}) = \mathrm{var}(Y \mid \mathcal{A})$ a.s.*
4. *If C_1 and C_2 are $\mathcal{A}$-measurable random variables with $\mathrm{E}(C_1^2) < \infty$ and $\mathrm{E}(C_2^2) < \infty$, then $\mathrm{cov}(Y_1 + C_1, Y_2 + C_2 \mid \mathcal{A}) = \mathrm{cov}(Y_1, Y_2 \mid \mathcal{A})$ a.s.*

We close this section with extensions of results for conditional distribution functions of random variables to conditional distribution functions of random vectors.

Proposition 10.24. Conditional multivariate distribution function *Let $\mathbf{Y} = (Y_1, \ldots, Y_k)$ be a random vector and $\mathcal{A} \subset \mathcal{F}$ be a sigma-field. There exists a version $F(\mathbf{y}, \omega)$ of $\mathrm{E}\{I(Y_1 \le y_1, \ldots, Y_k \le y_k) \mid \mathcal{A}\}$ that is, for each ω, a distribution function in $(y_1, \ldots, y_k)$.*

Proposition 10.25. Using conditional multivariate distribution functions to compute conditional expectations *If $F(\mathbf{y}, \omega)$ is a conditional distribution function of $\mathbf{Y}$ given $\mathcal{A}$, and $g(\mathbf{Y})$ is an integrable, Borel function of $\mathbf{Y}$, then $\int g(\mathbf{y}) dF(\mathbf{y}, \omega)$ is a version of $\mathrm{E}\{g(\mathbf{Y}) \mid \mathcal{A}\}$.*

Exercises

1. Let (X, Y) take the values $(0, 0)$, $(0, 1)$, $(1, 0)$, and $(1, 1)$ with probabilities p_{00}, p_{01}, p_{10}, and p_{11}, respectively, where $p_{00} + p_{01} + p_{10} + p_{11} = 1$ and $p_{00} + p_{01} > 0$, $p_{10} + p_{11} > 0$.
   (a) What is the conditional distribution of Y given $X = 0$?
   (b) Show that $\mathrm{E}(Y \mid X)$ is linear in X, and determine the slope and intercept.
2. Roll a die and let X denote the number of dots showing. Then independently generate $Y \sim \mathrm{U}(0, 1)$, and set $Z = X + Y$.
   (a) Find a conditional distribution function of Z given $X = x$ and a conditional distribution function of Z given $Y = y$.
   (b) Find a conditional distribution function of X given $Z = z$ and a conditional distribution function of Y given $Z = z$.

3. Dunnett's one-tailed test for the comparison of k treatment means $\mu_1, \ldots, \mu_k$ to a control mean μ_0 with common known variance σ^2 and common sample size n rejects the null hypothesis if $\max_i Z_{i0} > c$, where

$$Z_{i0} = \frac{\bar{Y}_i - \bar{Y}_0}{\sqrt{2\sigma^2/n}}$$

and $\bar{Y}_i$ is the sample mean in arm i. Under the null hypothesis, $\mu_i = \mu_0$, $i = 1, \ldots, k$, and without loss of generality, assume that $\mu_i = 0$, $i = 0, 1, \ldots, k$. Therefore, assume that $\bar{Y}_i \sim \mathrm{N}(0, \sigma^2/n)$.

   (a) Find the conditional distribution of $\max_i Z_{i0}$ given $\bar{Y}_0 = y_0$.

   (b) Find the conditional distribution of $\max_i Z_{i0}$ given $\bar{Y}_0 = z_0\sigma/n^{1/2}$.

   (c) Find the unconditional distribution of $\max_i Z_{i0}$.

4. Let X, Y be iid $\mathrm{N}(0, \sigma^2)$. Find the conditional distribution of $X^2 - Y^2$ given that $X + Y = s$.

5. Fisher's least significant difference (LSD) procedure for testing whether means $\mu_1, \ldots, \mu_k$ are equal declares $\mu_1 < \mu_2$ if both the t-statistic comparing μ_1 and μ_2 and F-statistic comparing all means are both significant at level α. When the common variance σ^2 is known, this is equivalent to rejecting the null hypothesis if $Z_{12}^2 > c_{1,\alpha}$ and $R^2 > c_{k-1,\alpha}$, where

$$Z_{12} = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{2\sigma^2/n}}, \quad R^2 = \frac{n}{(k-1)\sigma^2} \sum_{i=1}^{k} (\bar{Y}_i - \bar{Y})^2$$

and $c_{i,\alpha}$ is the upper α point of a chi-squared distribution with i degrees of freedom. Use the result of Problem 11 of Section 8.6 to find the conditional distribution of $Z_{12}^2 + R^2$ given Z_{12}^2. Use this to find an expression for $P(Z_{12}^2 > c_1 \cap R^2 > c_{k-1,\alpha})$.

6. Let Y be a random variable with finite mean, and suppose that $\mathrm{E}\{\exp(Y)\} < \infty$. What is the probability that $\mathrm{E}\{\exp(Y) \,|\, \mathcal{A}\} < \exp\{\mathrm{E}(Y \,|\, \mathcal{A})\}$?

7. Prove Markov's inequality for conditional expectation (part 2 of Proposition 10.21).

8. Prove Chebychev's inequality for conditional expectation (part 3 of Proposition 10.21).

9. Prove Hölder's inequality for conditional expectation (part 4 of Proposition 10.21).

10. Prove Schwarz's inequality for conditional expectation (part 5 of Proposition 10.21).

11. Prove Minkowski's inequality for conditional expectation (part 6 of Proposition 10.21).

12. Prove parts 1 and 2 of Proposition 10.23.

13. Prove parts 3 and 4 of Proposition 10.23.

14. Complete the proof of Proposition 10.17 by showing that the set $\mathcal{B}'$ of Borel sets B such that $\int_B dF_\omega(y)$ is $\mathcal{A}$-measurable is a monotone class containing the field in Proposition 3.8.

15. Consider the Z-statistic comparing two means with known finite variance σ^2,

$$Z = \frac{\bar{Y} - \bar{X}}{\sigma\sqrt{1/n_X + 1/n_Y}}.$$

Suppose that n_X remains fixed, and let F be the distribution function for $\bar{X}$. Assume that $Y_1, Y_2, \ldots$ are iid with mean μ_Y. Show that the asymptotic (as $n_Y \to \infty$ and n_X remains fixed) conditional distribution function for Z given $\bar{X} = x$ is normal and determine its asymptotic mean and variance. What is the asymptotic unconditional distribution of Z as $n_Y \to \infty$ and n_X remains fixed?

16. Prove Proposition 10.18, first when g is simple, then when g is nonnegative, then when g is an arbitrary Borel function.

10.4 Conditional Expectation As a Projection

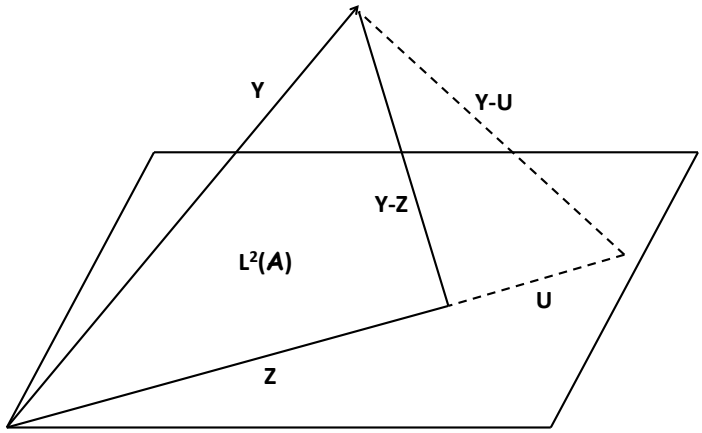

Figure 10.1: Conditional expectation as a projection. The conditional expected value $Z = \mathrm{E}(Y \,|\, \mathcal{A})$ is such that $Y - Z$ is orthogonal to every random variable in $L^2(\mathcal{A})$, the set of $\mathcal{A}$-measurable random variables in L^2.

We can view conditional expectation of Y given $\mathcal{A}$ in a geometric way. Assume throughout this subsection that $Y \in L^2$; i.e., $\mathrm{E}(Y^2) < \infty$. The set of L^2 random variables is a vector space with inner product $< U_1, U_2 >= \mathrm{E}(U_1 U_2)$. More generally, the set of L^2 functions on a measure space $(\Omega, \mathcal{F}, \mu)$ is a vector space with inner product $< f_1, f_2 >= \int f_1(\omega) f_2(\omega) d\mu$. This is a generalization of the dot product $\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^n x_i y_i$ of two vectors $\mathbf{x}$ and $\mathbf{y}$ because we can write this dot product as $\int_\Omega f_1(\omega) f_2(\omega) d\mu$, where μ is counting measure on $\Omega = \{\omega_1 = (x_1, y_1), \ldots, \omega_n = (x_n, y_n)\}$ and $f_1(\omega_i) = x_i$, $f_2(\omega_i) = y_i$. It suffices for our purposes to consider probability spaces. There is a close analogy between n-dimensional vectors equipped with dot products and random variables in L^2 equipped with inner products. Just as vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ are orthogonal if and only if their dot product $\mathbf{x}_1 \cdot \mathbf{x}_2$ is 0, random variables U_1 and U_2 in L^2 are orthogonal if and only if their inner product $< U_1, U_2 >= \mathrm{E}(U_1 U_2)$ is 0. Just as the length of a vector $\mathbf{x}$ is $(\mathbf{x} \cdot \mathbf{x})^{1/2}$, the length of a random variable U in L^2 is $< U, U >^{1/2} = \{\mathrm{E}(U^2)\}^{1/2}$.

Now consider the conditional expectation $\mathrm{E}(Y \,|\, \mathcal{A})$ of a random variable $Y \in L^2$. $\mathrm{E}(Y \,|\, \mathcal{A})$ is an $\mathcal{A}$-measurable random variable Z such that $\mathrm{E}\{(Y - Z) I(A)\} = 0$ for each

$A \in \mathcal{A}$. That is, $Y - Z$ is orthogonal to $I(A)$ for every $A \in \mathcal{A}$. But this means that $Y - Z$ is orthogonal to any nonnegative simple, $\mathcal{A}$-measurable random variable $\sum_{i=1}^{n} a_i I(A_i)$.

We show next that $Y - Z$ is orthogonal to U for every nonnegative $\mathcal{A}$-measurable random variable U with $\mathrm{E}(U^2) < \infty$. Any such U is a limit of simple, increasing $\mathcal{A}$-measurable random variables U_n. Therefore, $(Y - Z)^+ U_n \uparrow (Y - Z)^+ U$ and $(Y - Z)^- U_n \uparrow (Y - Z)U$. By the MCT, $\mathrm{E}\{(Y - Z)^+ U_n\} \to \mathrm{E}\{(Y - Z)^+ U\}$ and $\mathrm{E}\{(Y - Z)^- U_n\} \to \mathrm{E}\{(Y - Z)^- U\}$. Therefore,

$$\begin{aligned} 0 = \mathrm{E}\{(Y - Z)U_n\} &= \mathrm{E}\{(Y - Z)^+ U_n\} - \mathrm{E}\{(Y - Z)^- U_n\} \\ &\to \mathrm{E}\{(Y - Z)^+ U\} - \mathrm{E}\{(Y - Z)^- U\} \\ &= \mathrm{E}\{(Y - Z)U\}; \end{aligned} \tag{10.20}$$

i.e., $Y - Z$ is orthogonal to any nonnegative, $\mathcal{A}$-measurable random variable $U \in L^2$.

Similarly, we can show that $Y - Z$ is orthogonal to any $\mathcal{A}$-measurable random variable $U \in L^2$. We have proven the following result.

Proposition 10.26. Conditional expectation as a projection *If $Y \in L^2$, then $\mathrm{E}(Y \mid \mathcal{A})$ is a projection of Y onto $\mathcal{A}$-measurable random variables $U \in L^2$. That is, $Y - Z$ is orthogonal to every $\mathcal{A}$-measurable random variable $U \in L^2$.*

Example 10.27. Now suppose that we are attempting to predict the value of a random variable Y that might be difficult to measure. For instance, Y might require invasive medical imaging. But suppose that with less invasive imaging techniques, we have a set of variables $X_1, \ldots, X_k$ with which we might be able to predict Y accurately. Let $\mathcal{A}$ be $\sigma(X_1, \ldots, X_k)$, the sigma-field generated by $X_1, \ldots, X_k$. We will estimate Y using some function of $X_1, \ldots, X_k$, hence an $\mathcal{A}$-measurable random variable U. We want to minimize the mean-squared error $\mathrm{E}\{(Y - U)^2\}$ when using U to estimate Y.

In Figure 10.1, the plane represents the vector space of $\mathcal{A}$-measurable random variables in L^2, while the vector above the plane is the random variable Y. It is clear from the figure that among all vectors U in the plane, the one minimizing the squared residual $(Y - U)^2$ is the projection of Y onto the plane. That is, $Z = \mathrm{E}(Y \mid \mathcal{A})$ is the estimator of Y that minimizes the mean-squared error $\mathrm{E}\{(Y - U)^2\}$. □

In Example 10.27, we gave a geometric argument for the fact that $\mathrm{E}(Y \mid \mathcal{A})$ mimimizes $\mathrm{E}\{(Y - U)^2\}$ among $\mathcal{A}$-measurable functions $U \in L^2$. We now give a more formal proof of this fact.

Proposition 10.28. Conditional expectation minimizes MSE *If $Y \in L^2$, then $Z = \mathrm{E}(Y \mid \mathcal{A})$ minimizes $\mathrm{E}\{(Y - U)^2\}$ among all $\mathcal{A}$-measurable functions $U \in L^2$.*

Proof.

$$\begin{aligned} \mathrm{E}\{(Y - U)^2\} &= \mathrm{E}\{(Y - Z + Z - U)^2\} \\ &= \mathrm{E}\{(Y - Z)^2\} + \mathrm{E}\{(Z - U)^2\} + 2\mathrm{E}\{(Y - Z)(Z - U)\} \\ &= \mathrm{E}\{(Y - Z)^2\} + \mathrm{E}\{(Z - U)^2\} + 2\mathrm{E}[\mathrm{E}\{(Y - Z)(Z - U) \mid \mathcal{A}\}] \\ &= \mathrm{E}\{(Y - Z)^2\} + \mathrm{E}\{(Z - U)^2\} + 2\mathrm{E}[(Z - U)\mathrm{E}\{(Y - Z) \mid \mathcal{A}\}] \text{ (Prop 10.14)} \\ &= \mathrm{E}\{(Y - Z)^2\} + \mathrm{E}\{(Z - U)^2\} \\ &\geq \mathrm{E}\{(Y - Z)^2\}. \end{aligned} \tag{10.21}$$

This shows that $Z = \mathrm{E}(Y \mid \mathcal{A})$ minimizes $\mathrm{E}(Y - U)^2$ among all $\mathcal{A}$-measurable functions $U \in L^2$. □

Another consequence of viewing $Z = \mathrm{E}(Y \mid \mathcal{A})$ as a projection is the identity: $\mathrm{E}(Y^2) = \mathrm{E}(Z^2) + \mathrm{E}(Y - Z)^2$. This is just the Pythagorean theorem in a different vector space. A less visual approach can be used to derive this identity analogously to the proof of Proposition 10.21 (exercise). Apply this identity when Y has been centered to have mean 0, so that $Z = \mathrm{E}(Y \mid \mathcal{A})$ has mean 0 as well. Then $\mathrm{E}(Z^2) = \mathrm{var}(Z) = \mathrm{var}\{\mathrm{E}(Y \mid \mathcal{A})\}$. Also,

$$\begin{aligned} \mathrm{E}(Y - Z)^2 &= \mathrm{E}[\mathrm{E}\{(Y - Z)^2 \mid \mathcal{A}\}] \\ &= \mathrm{E}\{\mathrm{var}(Y \mid \mathcal{A})\}. \end{aligned} \tag{10.22}$$

We have deduced that $\mathrm{var}(Y) = \mathrm{var}\{\mathrm{E}(Y \mid \mathcal{A})\} + \mathrm{E}\{\mathrm{var}(Y \mid \mathcal{A})\}$. Apply this identity to $Y_1 + Y_2$: $\mathrm{var}(Y_1) + \mathrm{var}(Y_2) + 2\,\mathrm{cov}(Y_1, Y_2) = \mathrm{var}(Y_1 + Y_2)$

$$\begin{aligned} &= \mathrm{var}\{\mathrm{E}(Y_1 + Y_2 \mid \mathcal{A})\} + \mathrm{E}\{\mathrm{var}(Y_1 + Y_2 \mid \mathcal{A})\} \\ &= \mathrm{var}\{\mathrm{E}(Y_1 \mid \mathcal{A})\} + \mathrm{var}\{\mathrm{E}(Y_2 \mid \mathcal{A})\} + 2\,\mathrm{cov}\{\mathrm{E}(Y_1 \mid \mathcal{A}), \mathrm{E}(Y_2 \mid \mathcal{A})\} \\ &+ \mathrm{E}\{\mathrm{var}(Y_1 \mid \mathcal{A}) + \mathrm{var}(Y_2 \mid \mathcal{A}) + 2\,\mathrm{cov}(Y_1, Y_2 \mid \mathcal{A})\} \\ &= \mathrm{var}(Y_1) + \mathrm{var}(Y_2) + 2\,\mathrm{cov}\{\mathrm{E}(Y_1 \mid \mathcal{A}), \mathrm{E}(Y_2 \mid \mathcal{A})\} + 2\,\mathrm{E}\{\mathrm{cov}(Y_1, Y_2 \mid \mathcal{A})\}. \end{aligned}$$

Therefore, $\mathrm{cov}(Y_1, Y_2) = \mathrm{cov}\{\mathrm{E}(Y_1 \mid \mathcal{A}), \mathrm{E}(Y_2 \mid \mathcal{A})\} + \mathrm{E}\{\mathrm{cov}(Y_1, Y_2 \mid \mathcal{A})\}$.

We have proven the following useful result.

Proposition 10.29. Decomposition formulas for variances and covariances *Let Y, Y_1, Y_2 be random variables on $(\Omega, \mathcal{F}, P)$ with finite variance and $\mathcal{A} \subset \mathcal{F}$ be a sigma-field. The following decomposition formulas for variances and covariances hold.*

$$\mathrm{var}(Y) = \mathrm{var}\{\mathrm{E}(Y \mid \mathcal{A})\} + \mathrm{E}\{\mathrm{var}(Y \mid \mathcal{A})\}. \tag{10.23}$$

and

$$\mathrm{cov}(Y_1, Y_2) = \mathrm{cov}\{\mathrm{E}(Y_1 \mid \mathcal{A}), \mathrm{E}(Y_2 \mid \mathcal{A})\} + \mathrm{E}\{\mathrm{cov}(Y_1, Y_2 \mid \mathcal{A})\}. \tag{10.24}$$

Helpful mnemonic devices for (10.23) and (10.24) are $V = VE + EV$ and $C = CE + EC$.

Other results are also apparent from viewing conditional expectation as a projection. For instance, if Y is already $\mathcal{A}$-measurable, then the projection of Y onto $L^2(\mathcal{A})$ is Y itself. That is $\mathrm{E}(Y \mid \mathcal{A}) = Y$ almost surely if Y is $\mathcal{A}$-measurable. Also, suppose that $\mathcal{A} \subset \mathcal{C}$. Projecting Y first onto the larger sigma-field $\mathcal{C}$, and then onto the sub-sigma-field $\mathcal{A}$, is equivalent to projecting Y directly onto $\mathcal{A}$. That is $\mathrm{E}\{\mathrm{E}(Y \mid C) \mid \mathcal{A}\} = \mathrm{E}(Y \mid \mathcal{A})$ almost surely.

Proposition 10.30. Projecting first onto a larger space and then onto a subspace is equivalent to projecting directly onto the subspace *Let Y be a random variable with $\mathrm{E}(|Y|) < \infty$. If $\mathcal{A}$ and $\mathcal{C}$ are sigma-fields with $\mathcal{A} \subset \mathcal{C}$, then $\mathrm{E}\{\mathrm{E}(Y \mid C) \mid \mathcal{A}\} = \mathrm{E}(Y \mid \mathcal{A})$ a.s.*

Exercises

1. Show that Proposition 10.10 is a special case of Proposition 10.30.
2. Let Y be a random variable with $\mathrm{E}(|Y|) < \infty$, and suppose that Z is a random variable such that $Y - Z$ is orthogonal to X (i.e., $\mathrm{E}\{(Y - Z)X\} = 0$) for each $\mathcal{A}$-measurable random variable X . Prove that $Z = \mathrm{E}(Y \mid \mathcal{A})$ a.s.
3. Suppose that $\mathrm{E}(Y^2) < \infty$, and let $Z = \mathrm{E}(Y \mid \mathcal{A})$. Prove the identity $\mathrm{E}(Y^2) = \mathrm{E}(Z^2) + \mathrm{E}(Y - Z)^2$.

10.5 Conditioning and Independence

Recall from Section 4.5.1 that we defined random variables X and Y to be independent if $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$ for all Borel sets A and B, and we saw that this was equivalent to $P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$ for each x and y. The same is true if we replace random variables X and Y by random vectors $\mathbf{X}$ and $\mathbf{Y}$, A and B by k-dimensional Borel sets, and $X \leq x$ and $Y \leq y$ by $\mathbf{X} \leq \mathbf{x}$ (meaning $X_i \leq x_i$ for each i) and $\mathbf{Y} \leq \mathbf{y}$. We can also formulate independence in terms of conditional distribution functions as follows.

Proposition 10.31. ***X* and *Y* are independent if and only if conditional distribution given $\boldsymbol{X = x}$ des not depend on $\boldsymbol{x}$** *$(\mathbf{X}, \mathbf{Y})$ are independent if and only if there is a conditional distribution function $\psi(\mathbf{y} \,|\, \mathbf{x})$ of $(\mathbf{Y} \,|\, \mathbf{X} = \mathbf{x})$ that does not depend on $\mathbf{x}$.*

Proof. Suppose that $\mathbf{X}$ and $\mathbf{Y}$ are independent. We claim that the distribution function $G(\mathbf{y})$ of $\mathbf{Y}$ is a conditional distribution function of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ that does not depend on $\mathbf{x}$. It is clearly a distribution function in $\mathbf{y}$ and does not depend on $\mathbf{x}$. We need only show that $G(\mathbf{y})$ satisfies the definition of a conditional expected value of $\{I(\mathbf{Y} \leq \mathbf{y}) \,|\, \mathbf{X}\}$. Let B be any k-dimensional Borel set, where k is the dimension of $\mathbf{X}$. We must show that

$$\mathrm{E}\{G(\mathbf{y})I(\mathbf{X} \in B)\} = \mathrm{E}\{I(\mathbf{Y} \leq \mathbf{y})I(\mathbf{X} \in B)\}. \tag{10.25}$$

The left side of Equation (10.25) is clearly $P(\mathbf{X} \in B)G(\mathbf{y})$ because $G(\mathbf{y})$ is a constant. By Proposition 5.30, the right side of Equation (10.25) is $\mathrm{E}\{I(\mathbf{Y} \leq \mathbf{y})\}\mathrm{E}\{I(\mathbf{X} \in B)\} = P(\mathbf{X} \in B)G(\mathbf{y})$. Thus, Equation (10.25) holds. This completes the proof that if $(\mathbf{X}, \mathbf{Y})$ are independent, there exists a conditional distribution of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ that does not depend on $\mathbf{x}$. The proof of the reverse direction is left as an exercise. □

Another important concept is that of conditional independence given another random vector.

Definition 10.32. Conditional independence given a random vector *Random vectors $\mathbf{X}$ and $\mathbf{Y}$ are said to be conditionally independent given $\mathbf{Z}$ if there are conditional distribution functions $H(\mathbf{x}, \mathbf{y} \,|\, \mathbf{z})$, $F(\mathbf{x} \,|\, \mathbf{z})$, and $G(\mathbf{y} \,|\, \mathbf{z})$ of $(\mathbf{X}, \mathbf{Y} \,|\, \mathbf{Z} = \mathbf{z})$, $(\mathbf{X} \,|\, \mathbf{Z} = \mathbf{z})$, and $(\mathbf{Y} \,|\, \mathbf{Z} = \mathbf{z})$ such that $H(\mathbf{x}, \mathbf{y} \,|\, \mathbf{z}) = F(\mathbf{x} \,|\, \mathbf{z})G(\mathbf{y} \,|\, \mathbf{z})$.*

Proposition 10.33. *$\mathbf{X}$ and $\mathbf{Y}$ are conditionally independent given $\mathbf{Z}$ if and only if there is a conditional distribution function $H(\mathbf{y} \,|\, \mathbf{x}, \mathbf{z})$ of $(\mathbf{Y} \,|\, \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z})$ that does not depend on $\mathbf{x}$.*

Example 10.34. Conditionally independent but unconditionally positively associated Many statistical applications involve random variables $Y_1, \ldots, Y_n$ that are iid conditional on other information. Examples include the following.

1. The simple mixed model $Y_{ij} = \mu + b_i + \epsilon_{ij}$ for observation j on participant i, where the person-specific effects b_i are iid from some distribution and the ϵ_{ij} are iid mean 0 random errors. Once we condition on b_i, the observations are iid with mean b_i and variance σ_ϵ^2.

2. The Bayesian statistical paradigm under which, conditional on the parameter θ, the data Y_i are iid from density $f(y, \theta)$, and θ is randomly drawn from a "prior" distribution $\pi(\theta)$.

3. Comparisons of several sample means to a control mean with common sample size n: $Y_i = \bar{X}_i - \bar{X}_0$, $i = 1, \ldots, m$. Once we condition on the control sample mean $\bar{X}_0$, the Y_i are independent with mean $\mu_i - \bar{X}_0$ and variance σ_i^2/n. Under the null hypothesis that $\mu_i = \mu_0$, $i = 1, \ldots, m$, the Y_i are conditionally iid given $\bar{X}_0$.

Whenever $Y_1, Y_2, \ldots, Y_n$ are iid non-degenerate random variables conditional on other information, they are unconditionally positively associated in a certain sense. To see this, let the sigma-field $\mathcal{A} \subset \mathcal{F}$ represent the other information, and suppose that $Y_1, \ldots, Y_n$ are conditionally iid and non-degenerate given $\mathcal{A}$. Then

$$\begin{aligned} \operatorname{cov}(Y_1, Y_2) &= \operatorname{cov}\{\mathrm{E}(Y_1 \,|\, \mathcal{A}), \mathrm{E}(Y_2 \,|\, \mathcal{A})\} + \mathrm{E}\{\operatorname{cov}(Y_1, Y_2 \,|\, \mathcal{A})\}. \\ &= \operatorname{cov}\{\mathrm{E}(Y_1 \,|\, \mathcal{A}), \mathrm{E}(Y_1 \,|\, \mathcal{A})\} + 0 \\ &= \operatorname{var}\{\mathrm{E}(Y_1 \,|\, \mathcal{A})\} > 0. \end{aligned} \tag{10.26}$$

Of course if the Y_i are iid given $\mathcal{A}$, so are $Z_i = I(Y_i \leq y)$, $i = 1, \ldots, n$. It follows from Equation (10.26) applied to the Z_i that $\operatorname{cov}(Z_i, Z_j) > 0$. But

$$\begin{aligned} \operatorname{cov}(Z_i, Z_j) &= \mathrm{E}\{I(Y_i \leq y) I(Y_j \leq y)\} - \mathrm{E}\{I(Y_i \leq y)\}\mathrm{E}\{I(Y_j \leq y)\} \\ &= P(Y_i \leq y, Y_j \leq y) - P(Y_i \leq y) P(Y_j \leq y). \end{aligned} \tag{10.27}$$

Therefore, $P(Y_1 \leq y, Y_2 \leq y) > P(Y_1 \leq y)P(Y_2 \leq y)$, and similarly, $P(Y_1 > y, Y_2 > y) > P(Y_1 > y)P(Y_2 > y)$. That is, Y_1 and Y_2 tend to track together more frequently than they would if they were independent. □

Example 10.35. Unconditionally independent but conditionally dependent Example 10.34 shows that random variables can be conditionally independent, but very highly dependent unconditionally. The opposite is also true: two random variables X and Y can be unconditionally independent but conditionally highly correlated. For example, let Z_1 and Z_2 be independent normal random variables with variance σ^2, and let $X = Z_1 + Z_2$ and $Y = Z_1 - Z_2$. Then (X, Y) is bivariate normal with correlation 0, so X and Y are independent. On the other hand, conditional on Z_2, X is Z_1 plus the constant Z_2, and Y is Z_1 minus the constant Z_2. Therefore, $\operatorname{cor}(X, Y \,|\, Z_2) = 1$. That is, although X and Y are unconditionally independent, they are perfectly positively correlated given Z_2. Similarly, X and Y are conditionally perfectly negatively correlated given Z_1.

Independence of $X = Z_1 + Z_2$ and $Y = Z_1 - Z_2$ holds whenever Z_1 and Z_2 are bivariate normal with the same variance, even if they are not independent, and this has many important applications in statistics. For example, Z_1 and Z_2 may be measurements of the same quantity using two different devices, e.g., two different blood pressure machines, and we want to determine whether the results agree. Suppose that Z_1 is the measurement from a new, untested machine, whereas Z_2 is the measurement from the well-tested "gold standard" machine on the same patient. We regard the difference $Y = Z_1 - Z_2$ as the error of the new machine. To see if the magnitude of errors depends on the patient's blood pressure, we are tempted to plot Y against Z_2, but this is problematic. For instance, suppose that the two measurements have the same variance σ^2. Then

$$\begin{aligned} \operatorname{cov}(Y, Z_2) &= \operatorname{cov}(Z_1 - Z_2, Z_2) = \operatorname{cov}(Z_1, Z_2) - \operatorname{cov}(Z_2, Z_2) \\ &= \operatorname{cov}(Z_1, Z_2) - \sigma^2 \\ &= \sigma^2\{\operatorname{cor}(Z_1, Z_2) - 1\} < 0 \end{aligned} \tag{10.28}$$

unless Z_1 and Z_2 are perfectly negatively correlated. Thus, it will look like the accuracy of the new machine depends on the patient's blood pressure. We should instead plot $Y = Z_1 - Z_2$ against $X/2$, the average of Z_1 and Z_2 (Bland and Altman, 1986).

If $\text{var}(Z_1) = \text{var}(Z_2)$, X and Y are uncorrelated. The plot of Y against X should show no discernible pattern. If the plot shows a linear relationship with positive slope, that is an indication that $\text{var}(Z_1) > \text{var}(Z_2)$, whereas a negative slope indicates that $\text{var}(Z_1) < \text{var}(Z_2)$. This observation provides the basis for Pitman's test of equality of variances for paired observations: use the sample correlation coefficient between $X = Z_1 + Z_2$ and $Y = Z_1 - Z_2$ to test whether the population correlation coefficient is 0 (Pitman, 1939). This is equivalent to testing whether the slope of the regression of Y on X is 0. □

Example 10.36. Independence/dependence can sometimes depend on point of view We have seen that random variables can be conditionally, but not unconditionally, independent and vice versa. Whether we think conditionally or unconditionally depends on what we want to make inferences about. For instance, consider the simple mixed model

$$Y_{ij} = \mu + b_i + \epsilon_{ij}, \tag{10.29}$$

where Y_{ij} is the jth measurement on person i, μ is the mean over all people, and b_i is a random effect for person i. Usually, we want to make inferences about the mean μ over people. For instance, a pharmaceutical company or regulatory agency is interested in the mean effect of a drug over people. In this case, multiple measurements on each person are highly correlated. A one-patient study would never suffice to prove that the drug was effective in the general population. An individual patient has a different focus: "Does the drug help me." In this case it makes sense to condition on b_i. Under the simple mixed model (10.29), the multiple observations on person i are conditionally independent given b_i. Of course (10.29) is just a model, and may or may not be accurate. This example shows that one might view multiple measurements on the same person as dependent or (conditionally) independent, depending on whether the focus is on a population average or a single person. □

Example 10.37. Missing data There is an important distinction between conditional independence of X and Y given Z and conditional independence of X and Y given **particular** (but not all) values of z. For example, consider Rubin's (1976) seminal paper on missing data. Let $\mathbf{Y}$ be a random vector of data, and let $\mathbf{M}$ be the indicators that Y_i is nonmissing (a common, but unfortunate choice of notation, since the letter M sounds like it should indicate missing rather than nonmissing), $i = 1, \ldots, n$. Let $\mathbf{y}_{\text{obs}}$ and $\mathbf{y}_{\text{mis}}$ denote the values of the data that are observed and missing, respectively. Rubin defines the data to be *missing at random (MAR)* if

$$P(\mathbf{M} = \mathbf{m} \,|\, \mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}) \text{ is the same for all possible values of } \mathbf{y}_{\text{mis}}. \tag{10.30}$$

This sounds like it means $\mathbf{M}$ and $\mathbf{Y}_{\text{mis}}$ are conditionally independent, given $\mathbf{Y}_{\text{obs}}$, but it does not because condition (10.30) might hold only for the value of $\mathbf{y}_{\text{obs}}$ actually observed. Rubin offers an example involving hospital survey data that includes blood pressure. Blood pressure is missing if and only if that blood pressure is lower than the mean blood pressure in the population, μ. Suppose all participants in the survey have blood pressure exceeding μ. Then $\mathbf{y}_{\text{obs}}$ will be the vector of all blood pressures (none are missing), and for that value of $\mathbf{y}_{\text{obs}}$, condition (10.30) is satisfied vacuously. However, if any of the participants in the survey had blood pressure readings below μ, then there would have been missing data and condition (10.30) would not have held. Therefore, the data are missing at random if and only if none are missing. □

Example 10.38. Regression to the mean Regression to the mean is a phenomenon whereby extreme measurements tend to be less extreme when they are repeated. This is most often associated with a continuous outcome like blood pressure under the assumption that

the initial and later measurements Y_0 and Y_1 follow a bivariate normal distribution. If the means and variances are the same and $\text{cor}(Y_0, Y_1) = \rho$, then $\text{E}(Y_1 \mid Y_0 = y_0) = \rho y_0 + (1-\rho)\mu$ is a weighted average of y_0 and the mean μ. That is, Y_1 tends to "regress toward the mean" μ.

Regression to the mean can occur with discrete random variables as well, and can lead the uninitiated observer to misinterpret changes from baseline to the end of a clinical trial whose entry criteria require patients to have the disease at baseline. For example, let Y_0 and Y_1 be the indicator that a patient is diseased at baseline and the end of follow-up, respectively. Imagine that there is a "frailty" parameter P indicating the probability that the patient is diseased at a given time. Assume that Y_0 and Y_1 are conditionally independent given P, with $\text{E}(Y_i \mid P) = P$, $i = 0, 1$. If the trial recruits only patients with disease at baseline, then we must condition on $Y_0 = 1$. Even if a patient receives no treatment, Y_1 will be smaller than Y_0, on average.

To determine how much smaller Y_1 tends to be than Y_0, calculate $\text{E}(Y_1 \mid Y_0)$ using

$$\begin{aligned} \text{E}(Y_1 \mid Y_0) &= \text{E}\{\text{E}(Y_1 \mid Y_0, P) \mid Y_0\} \quad \text{(Proposition 10.30)} \\ &= \text{E}(P \mid Y_0) \quad \text{(conditional independence of } Y_0, Y_1 \text{ given } P\text{)}. \end{aligned} \tag{10.31}$$

Assume that the frailty parameters vary from patient to patient according to a uniform distribution on $[0, 1]$. That is, the density for p is $\pi(p) = 1$ for $p \in [0, 1]$. Before conditioning on $Y_0 = 1$, the joint density of (Y_0, P) with respect to the product measure $\mu_C \times \mu_L$ of counting measure and Lebesgue measure, evaluated at $(1, p)$, is $f(1, p) = P(Y_0 = 1 \mid p)\pi(p) = p$. The marginal density of Y_0 with respect to counting measure (i.e., its marginal probability mass function) is obtained by integrating the joint density over p: $g(1) = \int_0^1 p dp = 1/2$. Therefore, the conditional expectation of P given $Y_0 = 1$ is

$$\int_0^1 p \frac{f(1,p)}{g(1)} dp = \int_0^1 p \frac{p}{1/2} dp = 2/3.$$

Therefore, on average, $\text{E}(Y_1 - Y_0 \mid Y_0 = 1) = 2/3 - 1 = -1/3$. In other words, patients tend to have less disease at follow-up even in the absence of treatment. McMahon et al. (1994) showed a similar result applying a Poisson-gamma model to data from the Asymptomatic Cardiac Ischemia Pilot (ACIP) study. □

We hope the examples in this section give the readers an appreciation for the wealth of applications of conditional independence. Section 11.9 covers a useful graphical tool for understanding conditional independence relationships between variables.

Exercises

Prove that if there is a regular conditional distribution function $F(\mathbf{y} \mid \mathbf{x})$ of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ that does not depend on $\mathbf{x}$, then $\mathbf{X}$ and $\mathbf{Y}$ are independent.

10.6 Sufficiency

10.6.1 Sufficient and Ancillary Statistics

Suppose we want to estimate the probability p that a patient in the treatment arm of a clinical trial has an adverse event (AE). We have a random sample of 50 patients, and let

Y_i be the indicator that patient i has the adverse event; Y_i are iid Bernoulli (p). If someone supplements our data with the weights of 100 randomly selected rocks, should we use these weights to help us estimate p? Of course not. The rock weights are *ancillary*, meaning that their distribution does not depend on p.

Now imagine that the AE data are obtained by first generating the sum $S = \sum_{i=1}^{50} Y_i$ from its distribution, namely binomial $(50, p)$, and then generating the individual Y_is from their conditional distribution given S. If $S = s$, each outcome $y_1, \ldots, y_{50}$ with sum s has the same conditional probability, namely $1/\binom{50}{s}$. Notice that this conditional probability does not depend on p. Once we observe $S = s$, no additional information can be gleaned from observations generated from a conditional distribution that does not depend on p. The situation is completely analogous to the above example of augmenting adverse event data with rock weights. This motivates the following definition.

Definition 10.39. Sufficient statistic *The vector statistic* $\mathbf{S}$ *is said to be sufficient if there is a regular conditional distribution function of of* $Y_1, \ldots, Y_n$ *given* $\mathbf{S}$ *that does not depend on* θ*.*

Again it is helpful to imagine generating the data $Y_1, \ldots, Y_n$ by first generating the sufficient statistic $\mathbf{S}$ from its distribution, and then drawing the individual observations from their conditional distribution given $\mathbf{S}$. As these latter draws are from a distribution that is free of θ, they cannot help us make inferences about θ. This makes clear the fact that inferences about θ should be based solely on the sufficient statistic.

10.6.2 Completeness and Minimum Variance Unbiased Estimation

An estimator $\tilde{\theta}$ of a parameter θ is said to be *unbiased* if $\mathrm{E}(\tilde{\theta}) = \theta$. Loosely speaking, an unbiased estimator is on target on average. If we restrict attention to unbiased estimators of θ, it is natural to seek one with smallest variance, called a *minimum variance unbiased estimator (MVUE)*. The heuristic discussion above suggests that we should not even consider estimators that are not functions of the sufficient statistic. The following makes this more explicit. Any unbiased estimator $\tilde{\theta}$ of θ can be improved by taking $\hat{\theta} = \mathrm{E}(\tilde{\theta} \,|\, \mathbf{S})$; then $\hat{\theta}$ is unbiased (Proposition 10.10), is an extended Borel function of the sufficient statistic (Proposition 10.4), and has variance no greater than that of $\tilde{\theta}$ (Equation 10.23 of Proposition 10.29). Without loss of generality then, we can restrict attention to functions of the sufficient statistic. If there is only one unbiased function of $\mathbf{S}$, then our search is over. We therefore seek a condition that ensures there is only one unbiased function of $\mathbf{S}$. This motivates the following definition.

Definition 10.40. Complete statistic *A statistic* $\mathbf{S}$ *with distribution function* $F(s_1, \ldots, s_k; \theta)$ *is said to be complete if* $\mathrm{E}\{f(\mathbf{S})\} \equiv 0$ *for all values of* θ *implies that* $f(\mathbf{S}) = 0$ *with probability* 1 *for each* θ*.*

Proposition 10.41. Only one unbiased function if S is complete *If* $\mathbf{S}$ *is complete and* $f(\mathbf{S})$ *and* $g(\mathbf{S})$ *are unbiased, Borel functions of* $\mathbf{S}$*, then* $f(\mathbf{S}) = g(\mathbf{S})$ *with probability* 1 *for each* θ*.*

Proof. $\mathrm{E}\{f(\mathbf{S}) - g(\mathbf{S})\} = \theta - \theta = 0$. By completeness, $f(\mathbf{S}) - g(\mathbf{S}) = 0$ with probability 1 for each θ, so $f(\mathbf{S}) = g(\mathbf{S})$ with probability 1 for each θ. □

Proposition 10.42. Unbiased functions of a complete sufficient statistic are UMVUE *If* $\mathbf{S}$ *is complete and sufficient and* $f(\mathbf{S})$ *is unbiased, then* $f(\mathbf{S})$ *is a UMVUE.*

Proof. If not, then there is an unbiased estimator $\tilde{\theta}$ with smaller variance than $f(\mathbf{S})$. But then $\mathrm{E}(\tilde{\theta} \,|\, \mathbf{S})$ is an unbiased, Borel function of $\mathbf{S}$ with smaller variance than $f(\mathbf{S})$. But this is a contradiction because, by Proposition 10.41, $\mathrm{E}(\tilde{\theta} \,|\, \mathbf{S}) = f(\mathbf{S})$ with probability 1. □

The importance of complete, sufficient statistics has led to a thorough investigation of settings admitting such statistics. Complete, sufficient statistics have been established for the very large class of exponential families.

10.6.3 Basu's Theorem and Applications

Ancillary and sufficient statistics $\mathbf{A}$ and $\mathbf{S}$ are at opposite ends of the spectrum: in a real sense, ancillary statistics tell us nothing and sufficient statistics tell us everything about θ. It makes sense, then, that $\mathbf{A}$ and $\mathbf{S}$ might be independent. A beautiful theorem by Basu (1955) asserts that this is true, provided that $\mathbf{S}$ is complete.

Theorem 10.43. Basu's theorem: Ancillary and complete sufficient statistics are independent *Let* $\mathbf{S}$ *be a complete, sufficient statistic and* $\mathbf{A}$ *be ancillary. Then* $\mathbf{A}$ *and* $\mathbf{S}$ *are independent.*

Proof. Let $F(a_1, \ldots, a_k) = P(A_1 \leq a_1, \ldots, A_k \leq a_k)$. Because $\mathbf{A}$ is ancillary, $F(a_1, \ldots, a_k)$ does not depend on θ. Let $G(a_1, \ldots, a_k \,|\, S)$ be a regular conditional distribution function of $\mathbf{A}$ given $\mathbf{S}$. Then $F(a_1, \ldots, a_k) = \mathrm{E}\{G(a_1, \ldots, a_k \,|\, \mathbf{S})\}$. Therefore,

$$\mathrm{E}\{G(a_1, \ldots, a_k \,|\, \mathbf{S}) - F(a_1, \ldots, a_k)\} = 0.$$

But $G(a_1, \ldots, a_k \,|\, \mathbf{S}) - F(a_1, \ldots, a_k)$ is a function of $\mathbf{S}$, and $\mathbf{S}$ is complete, so $G(a_1, \ldots, a_k \,|\, \mathbf{S}) - F(a_1, \ldots, a_k) = 0$ a,s. I.e., the conditional distribution function of $\mathbf{A}$ given $\mathbf{S}$ is the same as its unconditional distribution function. By result 10.31, $\mathbf{A}$ and $\mathbf{S}$ are independent. □

There are many applications of Basu's theorem. For example, if $Y_1, \ldots, Y_n$ are iid $\mathrm{N}(\mu, \sigma^2)$, with σ^2 known, the sample mean $\bar{Y}_n$ is sufficient and complete for μ. On the other hand, the sample variance s_n^2 is ancillary for μ. This follows from the fact that adding the same constant c to each observation does not change s_n^2, so does not change the distribution of s_n^2. Basu's theorem implies the well-known result that $\bar{Y}_n$ and s_n^2 are independent for iid normal observations with finite variance. In fact, we can say more. The set $(s_2^2, s_3^2, \ldots, s_n^2$ of consecutive sample variances with sample sizes $2, 3, \ldots, n$ is also ancillary for the same reason. Therefore, $\bar{Y}_n$ is independent of $(s_2^2, s_3^2, \ldots, s_n^2)$. This has ramifications for adaptive clinical trials in which design changes might be based on interim variances. Conditioned on those variances, the distribution of $\bar{Y}_n$ is the same as its unconditional distribution. The next example expands on the use of Basu's theorem in adaptive clinical trials.

Example 10.44. Application of Basu's theorem: adaptive sample size calculation in clinical trials Consider a clinical trial with paired data and a continuous outcome Y such as change in cholesterol from baseline to 1 year. Let D_i be the difference between treatment and control measurements on pair i, and assume that $D_i \sim N(\mu, \sigma^2)$. We are interested in testing the null hypothesis $H_0 : \mu = 0$ versus the alternative hypothesis $H_1 : \mu > 0$. Before the trial begins, we determine the approximate number of pairs required for a one-tailed t-test at $\alpha = 0.025$ and 90% power using the formula

$$n \approx \frac{(1.96 + 1.28)^2 \sigma^2}{\mu^2}. \tag{10.32}$$

The two quantities we need for this calculation are the size of the treatment effect, μ, and the variance σ^2. We can usually determine the treatment effect more easily than the variance because we argue that if the effect is not at least a certain magnitude, the treatment is not worth developing. The variance, on the other hand, must be estimated. Sometimes there is good previous data on which to base σ^2, other times not. It would be appealing if we could begin with a pre-trial estimate σ_0^2 of the variance, but then modify that estimate and the sample size after seeing data from the trial itself.

Consider the following two-stage procedure. The first stage consists of half ($n_1 = n_0/2$) of the originally planned number of observations. From this first-stage data, we use $\hat{\sigma}^2 = (1/n)\sum D_i^2$ to estimate the variance σ^2. This is slightly different from the usual sample variance, which subtracts $\bar{D}$ from each observation and uses $n-1$ instead of n in the denominator. Nonetheless, $\hat{\sigma}^2$ is actually very accurate for typical clinical trials in which the treatment effect μ is not very large. We then substitute $\hat{\sigma}^2$ for σ^2 in (10.32) and compute the sample size $n = n(\hat{\sigma}^2)$. If $n \le n_0/2$, we collect no additional observations. Otherwise, the second stage consists of the number of additional observations required, $n_2 = n(\hat{\sigma}^2) - n_0/2$. It is tempting to pretend that the sample size had been fixed in advance, compute the usual t-statistic, and refer it to a t-distribution with $n-1$ degrees of freedom. This is actually a very good approximation, but the resulting test statistic is not exactly t_{n-1}.

We can construct an exact test as follows. Let T_1 be the t-statistic computed on the first-stage data. Under the null hypothesis that $\mu = 0$, the first-stage data are iid $\mathrm{N}(0, \sigma^2)$, and $\hat{\sigma}^2$ is a complete, sufficient statistic for σ^2. On the other hand, T_1 is ancillary for σ^2. This follows from the fact that dividing each observation by σ does not change T_1. By Basu's theorem, T_1 and $\hat{\sigma}^2$ are independent. This implies that, conditional on $\hat{\sigma}^2$, T_1 has a t-distribution with $n_1 - 1 = n_0/2 - 1$ degrees of freedom. If there is a second stage, then conditional on $\hat{\sigma}^2$, the t-statistic T_2 using only data from stage 2 has a t-distribution with $n_2 - 1 = n - n_0/2 - 1$ degrees of freedom and is independent of T_1. Let P_1 and P_2 denote the p-values corresponding to T_1 and T_2. Conditional on $\hat{\sigma}^2$, P_1 and P_2 are independent uniforms, so the inverse probability transformations $Z_1 = \Phi^{-1}(1-P_1)$ and $Z_2 = \Phi^{-1}(1-P_2)$ are independent standard normals. If there is no second stage, we can generate a superfluous standard normal deviate Z_2. Then, conditional on $\hat{\sigma}^2$,

$$Z = \frac{\sqrt{n_1}Z_1 + \sqrt{n_2}Z_2}{\sqrt{n_1 + n_2}}$$

has a standard normal distribution (when there is no second stage, $n_2 = 0$, and the superfluous random variable Z_2 receives zero weight). We reject the null hypothesis when $Z > z_\alpha$, the $(1-\alpha)$th quantile of a standard normal distribution. The conditional type I error rate given $\hat{\sigma}^2$ is exactly α. Therefore, the unconditional type I error rate is $\mathrm{E}\{P(Z > z_\alpha \,|\, \hat{\sigma}^2)\} = \mathrm{E}(\alpha) = \alpha$. That is, this two-stage procedure provides an exact level α test. □

10.6.4 Conditioning on Ancillary Statistics

It is good statistical practice to condition on the values of ancillary statistics. For instance, in a one-sample t-test, suppose that we flipped a coin to decide whether to use a sample size of 50 or 100. Thus, the estimated mean is $\bar{Y}_N$, where N is the random sample size. Suppose that as a result of the coin flip, $N = 50$. It would be silly to treat the sample size as random and compute the variance of $\bar{Y}_N$ as

$$\begin{aligned} \mathrm{var}(\bar{Y}_N) &= \mathrm{E}\{\mathrm{var}(\bar{Y}_N \mid N)\} + \mathrm{var}\{\mathrm{E}(\bar{Y}_N \mid N)\} \\ &= (1/2)\mathrm{var}(\bar{Y}_{50}) + (1/2)\mathrm{var}(\bar{Y}_{100}) + \mathrm{var}(\mu) \\ &= (1/2)(\sigma^2/50) + (1/2)(\sigma^2/100). \end{aligned} \tag{10.33}$$

This unconditional approach gives a misleading estimate of the variance for the sample size actually used, 50. We would instead use $\mathrm{var}(\bar{Y}_{50}) = \sigma^2/50$. That is, we would condition on $N = 50$ because N is an ancillary statistic.

The above example may seem artificial because we usually do not flip a coin to decide whether to double the sample size. But the same issue arises in a more realistic setting. In a clinical trial with n patients randomly assigned to treatment or control, analyses condition on the numbers actually assigned to treatment and control. For instance, suppose that the sample sizes in the treatment and control arms are 22 and 18, respectively. When we use a permutation test, we consider all $\binom{40}{22}$ different ways to assign 22 of the 40 patients to treatment; we do not consider all 2^{40} possibilities that would result if we did not fix the sample sizes. It makes sense to condition on the sample sizes actually used because they are ancillary. They give us no information about the treatment effect. However, the next example is a setting in which the sample sizes give a great deal of information about the treatment effect.

Example 10.45. Sample size is not always ancillary: ECMO The Extracorporeal Membrane Oxygenation (ECMO) trial (Bartlett et al., 1985) was in infants with primary pulmonary hypertension, a disease so serious that the mortality rate using the standard treatment, placing the baby on a ventilator, was expected to be 80%. The new treatment was extracorporeal membrane oxygenation (ECMO), an outside the body heart and lung machine used to allow the baby's lungs to rest and heal. Because of the very high mortality expected on the standard treatment, the trial used a nonstandard urn randomization technique that can be envisioned as follows. Place one standard (S) therapy ball and one ECMO (E) ball in an urn. For the first baby's assignment, randomly draw one of the two balls. If the ball is ECMO and the baby survives, or standard therapy and the baby dies, then "stack the deck" in favor of ECMO by replacing the ball and then adding another ECMO ball. On the other hand, if the first baby is assigned to ECMO and dies, or to the standard treatment and survives, replace the ball and add a standard therapy ball. That way, the second baby has probability 2/3 of being assigned to the therapy doing better so far. Likewise, after each new assignment, replace that ball and add a ball of the same or opposite treatment depending on whether that baby survives or dies. This is called a *response-adaptive* randomization scheme.

Table 10.1: Data from the ECMO trial; 0 and 1 denote survival and death, and E and S denote ECMO and standard treatment.

Outcome	0	1	0	0	0	0	0	0	0	0	0	0
Assignment	E	S	E	E	E	E	E	E	E	E	E	E
Probability	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{3}{4}$	$\frac{4}{5}$	$\frac{5}{6}$	$\frac{6}{7}$	$\frac{7}{8}$	$\frac{8}{9}$	$\frac{9}{10}$	$\frac{10}{11}$	$\frac{11}{12}$	$\frac{12}{13}$

The actual data in order are shown in Table 10.1, where 0 and 1 denote alive and dead, and E and S denote ECMO and standard therapy. The first baby was assigned to ECMO and survived. The next baby was assigned to the standard therapy and died. Then the next 10 babies were all assigned to ECMO and survived. At that point, randomization was

Table 10.2: Summary 2×2 table for the ECMO trial.

	Dead	Alive	
ECMO	0	11	11
Standard	1	0	1
	1	11	

discontinued. Table 10.2 summarizes the outcome data. If we use Fisher's exact test, which is equivalent to a permutation test on binary data, the one-tailed p-value is

$$\frac{\binom{11}{0}\binom{1}{1}}{\binom{12}{1}} = 1/12 = 0.083.$$

But the above calculation assumes that the 12 randomizations leading to 11 patients assigned to ECMO and 1 to standard therapy are equally likely. This is not true if we condition on the data and order of entry of patients shown in Table 10.1. To see this, consider the probability of the actual treatment assignments shown in Table 10.1. The probability that the first baby is assigned to E is 1/2. Because the first baby survived on E, there are 2 Es and 1 S in the urn when the second baby is randomized. Therefore, the conditional probability that the second baby is assigned to S is 1/3. Because that baby died on S, there are 3 Es and 1 S when the third baby is assigned. The conditional probability that the third baby is assigned to E is 3/4, etc. The probability of the observed assignment, given the outcome vector, is $(1/2)(2/3)(3/4)\ldots(12/13) = 1/26$. On the other hand, the randomization assignment (S,E,E,E,E,E,E,E,E,E,E,E) has probability 1/1716 given the observed outcome vector. The probability of each of the other 10 randomization sequences leading to the marginal totals of Table 10.2 is 1/429 (exercise). Therefore, the sum of probabilities of treatment sequences leading to these marginals is $1/26+1/1716+10(1/429) = 107/1716$. To obtain the p-value conditional on marginal totals of Table 10.2, we must sum the conditional probabilities of assignment vectors leading to tables at least as extreme as the observed one, and divide by 107/1716. But the actual assignment vector produces the most extreme table consistent with the given marginals, so the p-value conditional on marginal totals is $p = (1/26)/(107/1716) = 66/107 = 0.62$. This hardly seems to reflect the level of evidence in favor of ECMO treatment!

The problem with conditioning on the sample sizes is that they are not ancillary. They are quite informative about the treatment effect. Eleven of 12 babies were assigned to ECMO precisely because ECMO was working. Therefore, it makes no sense to condition on information that is informative about the treatment effect. It would be like arguing that a z-score of 3.1 is not at all unusual, conditioned on the fect that the z-score exceeded 3; conditioning on $Z > 3$ makes no sense because we are conditioning away the evidence of a treatment effect. That is why Wei (1988) did not condition on the marginals. He summed the probabilities of the actual assignments and the probability of (E,E,E,E,E,E,E,E,E,E,E,E), which yielded a p-value of 0.051. In the end, the trial generated substantial controversy (see Begg, 1990 and commentary, or Section 2 of Proschan and Nason, 2009) and did not convince the medical community. A subsequent larger trial showed that ECMO was superior to the standard treatment. There are many valuable lessons from the original ECMO trial, but the one we stress here is that the sample sizes in clinical trials are not always ancillary. When sample sizes are informative about the treatment effect, the analysis should not condition on them. □

Exercises

1. A common test statistic for the presence of an outlier among iid data from $N(\mu, \sigma^2)$ is the maximum normed residual

$$U = \max_{1 \le i \le n} \frac{|X_i - \bar{X}|}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}}.$$

   Using the fact that the sample mean and variance $(\bar{X}, s^2)$ is a complete, sufficient statistic, prove that U is independent of s^2.

2. Let $Y \sim \mathrm{N}(\mu, 1)$, and suppose that A is a set such that $P(Y \in A)$ is the same for all μ. Use Basu's theorem to prove that $P(Y \in A) = 0$ for all μ or $P(Y \in A) = 1$ for all μ.

3. In the ECMO example, consider the set of possible treatment assignment vectors that are consistent with the marginals of Table 10.2. Show that the probability of each of the 10 assignment vectors other than (E,S,E,E,E,E,E,E,E,E,E,E) and (S,E,E,E,E,E,E,E,E,E,E,E) is 1/429.

4. Let X be exponential (λ). It is known that X is a complete and sufficient statistic for λ. Use this fact to deduce the following result on the uniqueness of the *Laplace transform* $\psi(t) = \int_0^\infty f(x) \exp(tx) dx$ of a function $f(x)$ with domain $(0, \infty)$. If $f(x)$ and $g(x)$ are two functions whose Laplace transforms agree for all $t < 0$, then $f(x) = g(x)$ except on a set of Lebesgue measure 0.

10.7 Expect the Unexpected from Conditional Expectation

This section covers some common pitfalls and errors in reasoning with conditional expectation. In more elementary courses that assume there is an underlying density function, such reasoning works and is often encouraged. But our new, more general definition of conditional expectation applies whether or not there is a density function. With this added generality comes the opportunity for errors, as we shall see. In other cases, errors result from a simple failure to compute the correct conditional distribution, as in the two envelope paradox of Example 1.4.

Example 10.46. Return to the two envelopes: a simulation Recall that in the two envelope paradox of Example 1.4, one envelope has twice the amount of money as the other. The amounts in your and my envelopes are X and Y, respectively. Simulate this experiment as follows. Generate a random variable T_1 from a continuous distribution on $(0, \infty)$. For simplicity, let T_1 be exponential with parameter 1. Generate a Bernoulli (1/2) random variable Z_1 independent of T_1; if $Z_1 = 0$, set $T_2 = (1/2)T_1$, while if $Z_1 = 1$, set $T_2 = 2T_1$. Now generate another independent Bernoulli Z_2; if $Z_2 = 0$, set $(X, Y) = (T_1, T_2)$, while if $Z_2 = 1$, set $(X, Y) = (T_2, T_1)$. Repeat this experiment a thousand times, recording the (X, Y) pairs for each. See which of the following steps is the first not to hold.

1. All pairs have $Y = X/2$ or $Y = 2X$.

2. Approximately half the pairs have $Y = X/2$ and half have $Y = 2X$ (that is, $Y = X/2$ with probability 1/2 and $Y = 2X$ with probability 1/2).

3. Regardless of the value x of X, approximately half the pairs (x, Y) have $Y = x/2$, half have $Y = 2x$ (that is, given $X = x$, $Y = x/2$ or $2x$, with probability 1/2 each).

4. $\mathrm{E}(Y \mid X = x) = (x/2)(1/2) + (2x)(1/2) = (5/4)x$.

Figure 10.2 shows (X, Y) pairs from a thousand simulations. The points all lie on one of two lines, $Y = (1/2)X$ or $Y = 2X$. Thus, statement 1 is true. In this simulation, 491 of the 1000 pairs have $X < Y$, so statement 2 is true. However, among the 27 pairs with $X > 5$, none produced $X < Y$. It is clear that the conditional probability that $Y = x/2$ given that $X = x$ is not 1/2 for large values of x. The problem in the two envelopes paradox is that we incorrectly conditioned on $X = x$. The statement that Y is equally likely to be $X/2$ or $2X$ is true unconditionally, but not conditional on X. □

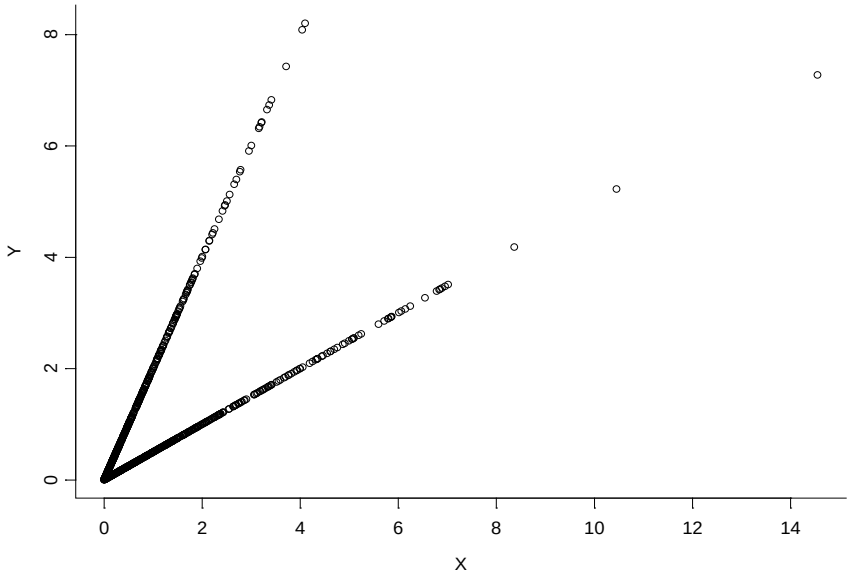

Figure 10.2: Plot of 1000 simulations of (X, Y) from the two envelopes paradox.

10.7.1 Conditioning on Sets of Probability 0

Most of the problems with conditional expectation stem from conditioning on sets of probability 0. The following is a case in point.

Example 10.47. Distribution of random draw from [0, 1] given it is rational Let $(\Omega, \mathcal{F}, P)$ be $([0,1], \mathcal{B}, \mu_L)$. Then ω corresponds to drawing a number randomly from the unit interval. What is the conditional distribution of ω given that ω is rational? It seems that ω must be uniformly distributed on the set of rationals. If not, then which rationals should be more likely than others? But Section 3.6 shows that there is no uniform distribution on a countable set. On the one hand, the distribution seems like it must be uniform, and on the other hand, it cannot be uniform. □

The problem with Example 10.47 is that conditional expectation conditions on a random variable or a sigma-field, not on a single set of probability 0. At this point the reader may be wondering (1) whether this kind of problem could arise in practice and (2) whether one could solve the problem by formulating it as a conditional expectation given a random variable. The next example answers both of these questions.

Example 10.48. Conditioning on the equality of two continuous random variables Borel paradox In a clinical trial of an HIV vaccine, investigators discovered participants with genetically extremely similar viruses, suggesting that some trial participants either had sex with each other or with a common partner. We will refer to this as "sexual sharing." This raises a question about whether the usual methods of analyzing independent data can still be applied. Proschan and Follmann (2008) argue that under certain reasonable assumptions, a permutation test is still valid. In response to a reviewer's query, the authors pointed out that even in clinical trials with independent data, one can view the observations as conditionally dependent given certain information. For instance, a certain gene might protect someone against acquiring HIV even if that person engages in risky behavior. Conditioned on the presence/absence of that gene, the HIV indicators Y_i of different patients are correlated: the Y_i of two patients who both have the gene or both do not have the gene (concordant patients) are positively correlated, whereas the Y_i of two patients who have different gene statuses (discordant patients) are negatively correlated.

The gene status random variable is fairly simple because it takes only two values, but we might try to extend the same reasoning to a setting with continuous random variables. This is exactly the setting of Example 1.3. In that example, we postulate a linear model $Y = \beta_0 + \beta_1 X + \epsilon$ relating the continuous outcome Y to the continuous covariate X. For simplicity, we assume that $X \sim N(0,1)$. We then attempt to obtain the relationship between Y values of two people with the same X value. In formulation 2 of the problem, we imagine continually observing pairs (X, Y) until we find two pairs with the same X value. The first step toward determining the relationship between the two Y values is determining the distribution of the common X, i.e., the distribution of X_1 given that $X_1 = X_1$. Formulating the event $X_1 = X_2$ in different ways, $X_2 - X_1 = 0$ or $X_2/X_1 = 1$, gives different answers.

The problem is that in the plane, sets of the form $x_2/x_1 \leq a$ look quite different from sets of the form $x_2 - x_1 \leq b$, and this leads to different conditional distributions given X_2/X_1 versus given $X_2 - X_1$. The conditional distribution of X_1 given $X_1 = X_2$ is not well-defined. Again we cannot think in terms of conditioning on **sets** of probability 0, but only on random variables or sigma-fields. Because we can envision sets of probability 0 in more than one way as realizations from random variables, in general, there is not a unique way to define conditioning on an arbitrary set of probability 0. □

10.7.2 Substitution in Conditioning Expressions

Who among us has never used the following reasoning? To compute the distribution of $X + Y$, where X and Y are independent with respective distributions F and G, we argue that

$$\begin{aligned} P(X+Y \leq z) &= \int P(X+Y \leq z \,|\, X = x) dF(x) \\ &= \int P(x+Y \leq z \,|\, X = x) dF(x) \\ &= \int P(Y \leq z - x \,|\, X = x) dF(x) \end{aligned}$$

$$= \int G(z-x)dF(x). \tag{10.34}$$

The last step follows from the independence of X and Y. The second step replaces the random variable X with its value, x. Such arguments abound in statistics, and they can help us deduce results. Nonetheless care is needed in carrying out substitution in conditional expectations. This was not the case in elementary statistics courses because the definition of conditional expectation was much more specific—Expression (10.2). But the more general Definition 10.2 admits different versions of conditional expectation, and this can lead to confusion.

Example 10.49. Confusion from substitution of $X = x$ Compute $P(X \le x)$ by first conditioning on $X = x$: $P(X \le x \,|\, X = x)$. Given that $X = x$, $X \le x$ is guaranteed, so $P(X \le x \,|\, X = x) = 1$. By Proposition 10.10, $P(X \le x) = \mathrm{E}(1) = 1$. Something is clearly amiss if we have shown that for an arbitrary random variable X and value x, $P(X \le x) = 1$. The confusion stems from the two different xs, the argument of the distribution function and the value of the random variable. Had we used, say x_0 and x to denote the argument of the distribution function and value of X, respectively, we would have concluded that

$$P(X \le x_0 \,|\, X = x) = I(x \le x_0), \quad P(X \le x_0 \,|\, X) = I(X \le x_0).$$

When we take the expected value of this expression over the distribution of X, we reach the correct conclusion that $P(X \le x_0) = \mathrm{E}\{I(X \le x_0)\}$. □

There is potential for confusion whenever there is a random variable X and its value x in the same expression, as in $\mathrm{E}\{f(x, Y) \,|\, X = x\}$. The following example makes this more clear.

Example 10.50. More confusion in substitution Let $f(x, y) = x + y$, and let X and Y be iid standard normals. Consider the calculation of $\mathrm{E}\{f(X, Y) \,|\, X = 0\}$ using the "rule" $\mathrm{E}\{f(0, Y) \,|\, X = 0\} = \mathrm{E}(0 + Y \,|\, X = 0) = 0 + \mathrm{E}(Y \,|\, X = 0)$. One version of $\mathrm{E}(Y \,|\, X)$ is $\mathrm{E}(Y) = 0$ because X and Y are independent. Using this version, we get $\mathrm{E}\{f(x, Y) \,|\, X = 0\} = 0 + 0 = 0$. But we can change the definition of $\mathrm{E}(Y \,|\, X)$ at the single value $X = 0$, and it will remain a version of $\mathrm{E}(Y \,|\, X)$. For instance,

$$\mathrm{E}(Y \,|\, X) = \begin{cases} 0 & \text{if } X \neq 0 \\ 1 & \text{if } X = 0. \end{cases} \tag{10.35}$$

Using this version, we get $\mathrm{E}\{f(0, Y) \,|\, X = 0\} = 0 + 1 = 1$. Of course, we could replace the value 1 in Equation (10.35) with any other value, so $\mathrm{E}\{f(0, Y) \,|\, X = 0)$ could be any value whatsoever.

The same problem holds if we replace the conditioning value 0 by any value x. In this example: $\mathrm{E}\{f(x, Y) \,|\, X = x\} = \mathrm{E}(x + Y \,|\, X = x) = x + \mathrm{E}(Y \,|\, X = x)$. Let $g(x)$ be an arbitrary Borel function. One version of $\mathrm{E}(Y \,|\, X)$ is

$$\mathrm{E}(Y \,|\, X) = \begin{cases} 0 & \text{if } X \neq x \\ g(x) - x & \text{if } X = x. \end{cases} \tag{10.36}$$

With this version, $\mathrm{E}\{f(x, Y) \,|\, X = x\} = x + g(x) - x = g(x)$. But $g(x)$ was arbitrary, so $\mathrm{E}\{f(x, Y) \,|\, X = x\}$ could literally be any Borel function of x. □

In light of the above examples, which of the following two statements is incorrect?

1. If $G(y \mid X = x)$ is a regular conditional distribution function of Y given $X = x$, then $\mathrm{E}\{f(X,Y) \mid X = x\} = \int f(x,y) dG(y \mid x)$.

2. $\mathrm{E}\{f(X,Y) \mid X = x\} = \mathrm{E}\{f(x,Y) \mid X = x\}$.

The statements seem equivalent, but they are not. The first involves a **specific** version of $\mathrm{E}\{f(x,Y) \mid X = x\}$, whereas $\mathrm{E}\{f(x,Y) \mid X = x\}$ in the second statement is **any** version of $\mathrm{E}\{f(x,Y) \mid X = x\}$. Thus, the second statement asserts that $\mathrm{E}\{f(X,Y) \mid X = x\} = \mathrm{E}\{f(x,Y) \mid X = x\}$ for **any** version of $\mathrm{E}\{f(x,Y) \mid X = x\}$. Example 10.50 is a counterexample.

Fortunately, the first statement is correct. More generally:

Proposition 10.51. Substitution is permitted once we select a specific version of the conditional distribution function *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be random vectors and* $\lambda(X,Y)$ *be a function such that* $|\lambda(\mathbf{X},\mathbf{Y})|$ *is integrable. If* $G(\mathbf{y} \mid \mathbf{x})$ *is a regular conditional distribution function of* $\mathbf{Y}$ *given* $\mathbf{X} = \mathbf{x}$*, then one version of* $\mathrm{E}\{\lambda(\mathbf{X},\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}\}$ *is* $\int \lambda(\mathbf{x},\mathbf{y}) dG(\mathbf{y} \mid \mathbf{x})$.

An immediate consequence of Proposition 10.51 is the following.

Corollary 10.52. Substitution and independent vectors *Suppose that* $\mathbf{X}$ *and* $\mathbf{Y}$ *are independent with* $\mathbf{Y} \sim G(\mathbf{y})$. *If* $|\lambda(\mathbf{X},\mathbf{Y})|$ *is integrable, then one version of* $\mathrm{E}\{\lambda(\mathbf{X},\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}\}$ *is* $\int \lambda(\mathbf{x},\mathbf{y}) dG(\mathbf{y})$.

Example 10.53. Permutation tests condition on outcome data In a clinical trial comparing means in the treatment and control arms, the treatment effect estimate is

$$\hat{\delta}(\mathbf{Z},\mathbf{Y}) = (1/n_T)\sum_{i=1}^{n} Z_i Y_i - (1/n_C)\sum_{i=1}^{n}(1 - Z_i)Y_i \tag{10.37}$$

where Z_i is 1 if patient i is assigned to treatment and 0 if control, and n_T and n_C are the numbers of patients assigned to treatment and control, respectively. That is, $\sum_{i=1}^{n} Z_i = n_T$. Under the null hypothesis, treatment has no effect on outcome, and the treatment vector $\mathbf{Z}$ is assumed independent of the outcome vector $\mathbf{Y}$. A regular conditional distribution function $F(\mathbf{z} \mid \mathbf{Y} = \mathbf{y})$ of $\mathbf{Z}$ given $\mathbf{Y} = \mathbf{y}$ is its unconditional distribution with mass function $p(\mathbf{z}) = \binom{n}{n_T}^{-1}$ for each string $\mathbf{z}$ of zeroes and ones with exactly n_T ones. Therefore, a regular conditional distribution function of $\hat{\delta}$ given $\mathbf{Y} = \mathbf{y}$ is

$$\begin{aligned} P\{\hat{\delta}(\mathbf{Z},\mathbf{Y}) \le d \mid \mathbf{Y} = \mathbf{y}\} &= \int I\{\hat{\delta}(\mathbf{z},\mathbf{y}) \le d\} dF(\mathbf{z} \mid \mathbf{y}) \\ &= \sum_{\mathbf{z}} I\{\hat{\delta}(\mathbf{z},\mathbf{y}) \le d\} p(\mathbf{z}) \text{ (Corollary 10.52)} \\ &= \frac{1}{\binom{n}{n_T}} \sum_{\mathbf{z}} I\{\hat{\delta}(\mathbf{z},\mathbf{y}) \le d\}. \end{aligned} \tag{10.38}$$

This is the permutation distribution of $\hat{\delta}$. □

It is sometimes possible to bypass arguments using conditional probability. This is considered better form, just as a direct proof is considered better form than a proof by contradiction. The following example illustrates how to bypass conditioning.

Example 10.54. An illustration of avoiding conditioning: Stein's method In a one-sample t-test setting with iid normal observations, the standard 95% confidence interval for the mean μ is $\bar{Y}_n \pm t_{n-1,\alpha/2} s_n/n^{1/2}$, where n is the sample size, s_n is the sample standard deviation, and $t_{n-1,\alpha/2}$ is the upper $\alpha/2$ point of a t-distribution with $n-1$ degrees of freedom. Notice that the width of the confidence interval is random because it depends on s_n. Even though the width tends to 0 almost surely as $n \to \infty$, even for large sample size n, there is some small probability that s_n is large enough to make the interval wide. Stein (1945) showed how to construct a confidence interval with a fixed and arbitrarily small width.

The first step of Stein's method is to take a subsample of size m, say $m = 30$. Let s_m be the sample standard deviation for this subsample. Once we observe the subsample, s_m is a fixed number. We choose the final sample size $N = N(s_m)$ as a certain Borel function of s_m to be revealed shortly. We then observe $N - m$ additional observations. Consider the distribution of

$$T = \frac{\bar{Y}_N - \mu}{s_m/N^{1/2}}.$$

We claim that T has a t-distribution with $m-1$ degrees of freedom, so that

$$P\left(\bar{Y}_N - t_{m-1,\alpha/2}\frac{s_m}{\sqrt{N}} \le \mu \le \bar{Y}_N + t_{m-1,\alpha/2}\frac{s_m}{\sqrt{N}}\right) = 1-\alpha.$$

That is,

$$\bar{Y}_N \pm t_{m-1,\alpha/2}\frac{s_m}{\sqrt{N}}$$

is a $100(1-\alpha)\%$ confidence interval for μ with width $2t_{m-1,\alpha/2}(s_m/n^{1/2})$. We can construct a confidence interval of arbitrarily small width w or less as follows. For the observed value s_m, simply choose N to be the smallest integer n such that $2t_{m-1,\alpha/2}(s_m/n^{1/2}) < w$. The confidence interval will then have width w or less. This device can also be used to construct a test whose power does not depend on the unknown value of σ, hence the title of Stein's (1945) paper.

To deduce that T follows a t-distribution with $m-1$ degrees of freedom, take $\mu = 0$ without loss of generality. Consider the conditional distribution of $\bar{Y}_N$ given s_m. Given s_m, $N = n$ is a fixed number. Thus, the conditional distribution of Y_N given s_m such that $N = n$ is the same as the conditional distribution of Y_n given s_m such that $N = n$. But we saw from Basu's theorem that $\bar{Y}_n$ is independent of $s_2^2, s_3^2, \ldots s_n^2$, so the conditional distribution of $Z_N = \bar{Y}_N/(\sigma/N^{1/2})$ given s_m such that $N = n$ is the same as the unconditional distribution of Z_n, namely standard normal. Therefore, unconditionally, Z_N is standard normal and independent of s_m. It follows that

$$T = \frac{\bar{Y}_N}{s_m/N^{1/2}} = \frac{\frac{\bar{Y}_N}{\sigma/N^{1/2}}}{s/\sigma} = \frac{Z_N}{\sqrt{\frac{(m-1)s_m^2/\sigma^2}{m-1}}} \tag{10.39}$$

is the ratio of a standard normal and the square root of an independent chi-squared $(m-1)$ random variable divided by its number of degrees of freedom. By definition, T has a t-distribution with $m-1$ degrees of freedom.

The above development is very helpful to **deduce** the distribution of T, but is somewhat awkward as a proof. Once we know the right answer, we can circumvent conditioning and provide a more appealing proof:

$$\begin{aligned}
P(\{Z_N \le z\} \cap \{s_m^2 \le u\}) &= \sum_{n=m}^{\infty} P\left(\left\{\frac{\bar{Y}_N}{\sigma/N^{1/2}} \le z\right\} \cap \{s_m^2 \le u\} \cap \{N = n\}\right) \\
&= \sum_{n=m}^{\infty} P\left(\left\{\frac{\bar{Y}_n}{\sigma/n^{1/2}} \le z\right\} \cap \{s_m^2 \le u\} \cap \{N = n\}\right) \\
&= \sum_{n=m}^{\infty} P\left(\frac{\bar{Y}_n}{\sigma/n^{1/2}} \le z\right) P\left(\{s_m^2 \le u\} \cap \{N = n\}\right) \\
&= \sum_{n=m}^{\infty} \Phi(z) P\left(\{s_m^2 \le u\} \cap \{N = n\}\right) \\
&= \Phi(z) \sum_{n=m}^{\infty} P\left(\{s_m^2 \le u\} \cap \{N = n\}\right) \\
&= \Phi(z) P(s_m^2 \le u). \qquad (10.40)
\end{aligned}$$

The third line follows from the independence of $\bar{Y}_n$ and s_m and the fact that $N = N(s_m)$ is a function of s_m. Equation (10.40) shows that the joint distribution function of (Z_N, s_m^2) is that of two independent random variables, the first of which is standard normal. Also, we know that $(m-1)s_m^2/\sigma^2$ is chi-squared with $m-1$ degrees of freedom. It follows that (10.39) is the ratio of a standard normal deviate to the square root of a chi-squared $(m-1)$ divided by its degrees of freedom. Therefore, T has a t-distribution with $m-1$ degrees of freedom.

Exercises

1. In Example 10.47, let $Y(\omega) = \omega$ and consider two different sigma-fields. The first, $\mathcal{A}_1$, is the sigma-field generated by $I(\omega \text{ is rational})$. The second, $\mathcal{A}_2$, is the sigma-field generated by Y. What are $\mathcal{A}_1$ and $\mathcal{A}_2$? Give regular conditional distribution functions for Y given $\mathcal{A}_1$ and Y given $\mathcal{A}_2$.

2. Show that the (X, Y) pairs in Example 10.48 exhibit quirky behavior even when X is a binary gene status random variable. More specifically, consider the following two ways of simulating pairs (X, Y). Method 1: generate a gene status (present or absent) for person 1, then assign that same value to person 2. Then generate the two Ys using the regression equation, with ϵ having a $\mathrm{N}(0, \sigma^2)$ distribution. Method 2: continue generating (X, Y) pairs until you find two pairs with the same X value. Show that the distribution of the common value of X is different using Method 1 versus Method 2.

10.7.3 Weak Convergence of Conditional Distributions

In this section we investigate whether weak convergence of joint distributions translates into weak convergence of conditional distributions and vice versa.

Let (X_n, Y_n) be a sequence of random variables with joint distribution function $H_n(x, y)$ and marginal distribution functions $F_n(x)$ and $G_n(y)$. Write $H_n(x, y)$ as

$$H_n(x, y) = F_n(x) P(Y_n \le y \,|\, X_n \le x). \qquad (10.41)$$

We can define $P(Y_n \le y \,|\, X_n \le x)$ to be $\Phi(Y)$ when $F_n(x) = 0$, so factorization (10.41) holds even when $F_n(x) = 0$ because the left and right sides are both 0. Furthermore, if H_n

converges weakly to some joint distribution function $H(x, y)$, then the marginal distribution $F_n(x)$ also converges weakly to the corresponding marginal distribution function $F(x)$ by the Mann-Wald theorem (Theorem 6.59) because $\lambda(x, y) = x$ is a continuous function of (x, y) (alternatively, one could invoke the Cramer-Wold device, Corollary 8.43). By this fact and the factorization (10.41), if $H_n(x, y) \xrightarrow{D} H(x, y)$ and (x, y) is a continuity point of $H(x, y)$ such that $H(x, y) > 0$ (which implies that x is a continuity point of $F(x)$ and $F(x) > 0$), then the left side of Equation (10.41) converges to $H(x, y)$ if and only if the right side converges to $F(x)H(x, y)$. That is, weak convergence of H_n is equivalent to convergence of the marginal distribution $F_n(x)$ plus convergence of $P(Y_n \leq y \,|\, X_n \leq x)$ at each continuity point (x, y) of $H(x, y)$ such that $H(x, y) > 0$.

The next question is whether weak convergence of the joint distribution function is equivalent to weak convergence of the marginal distribution function plus weak convergence of the conditional distribution function $K_n(y \,|\, X_n = x)$. More specifically, our question is:

$$\text{Is } H_n(x, y) \xrightarrow{D} H(x, y) \text{ equivalent to } F_n(x) \xrightarrow{D} F(x) \text{ plus } K_n(y \,|\, x) \xrightarrow{D} K(y \,|\, x)?$$

This differs from the development of the preceding paragraph because we are now considering the conditional distribution of Y_n given $X_n = x$ rather than the conditional distribution of $Y_n \,|\, X_n \leq x$. Before proceeding, we note that we have already seen one setting in which this holds, namely when (X_n, Y_n) are independent.

Let $H_n(x, y)$ and $K_n(y \,|\, x)$ denote the joint and conditional distribution of (X_n, Y_n) and Y_n given $X_n = x$ respectively. Suppose that F_n converges weakly to a distribution function F and $K_n(y \,|\, x)$ converges weakly to a conditional distribution function $K(y \,|\, x)$. Does it follow that the joint distribution function $H_n(x, y)$ converges weakly to $F(x)K(y \,|\, x)$? Likewise, suppose that $H_n(x, y)$ converges weakly to the joint distribution function $H(x, y)$. Does it follow that the conditional distribution function $K_n(y \,|\, x)$ converges weakly to the corresponding conditional distribution function $K(y \,|\, x) = H(x, y)/\int_{x=-\infty}^{\infty} dH(x, y)$? Unfortunately, the answer to both of these questions is no. The following example shows that even in a relative simple setting in which X_n and Y_n take on only two possible values and the limiting joint and conditional distributions are point masses, those point masses may disagree.

Example 10.55. Possible disconnect between convergence of joint and conditional distributions Let

$$X_n = \begin{cases} 0 & \text{with probability } 1/n \\ 1 & \text{with probability } 1 - 1/n \end{cases}$$

and let $Y_n = I(X_n > 0)$. The conditional distribution of Y_n given $X_n = 0$ is a point mass at 0. Therefore, the conditional distribution of Y_n given $X_n = 0$ converges weakly to a point mass at 0. On the other hand, (X_n, Y_n) converges in probability to $(1, 1)$. Therefore, the joint distribution of (X_n, Y_n) converges weakly to a point mass at $(1, 1)$. □

Definition 10.56. Strong convergence *A sequence of distribution functions F_n is said to converge strongly to distribution F if $F_n(x) \to F(x)$ for every $x \in R$.*

Theorem 10.57. *(Sethuraman)* **Weak convergence of marginal distribution and strong convergence of conditional distribution implies weak convergence of joint distribution** *If F_n converges weakly to F and $K_n(y \,|\, x)$ converges strongly to $K(y \,|\, x)$, then $H_n(x, y)$ converges weakly to $H(x, y)$.*

10.8 Conditional Distribution Functions As Derivatives

There is another intuitive way to try to define a conditional distribution function $H(y \,|\, x)$ of Y given $X = x$. Condition on $x - \Delta < X < x + \Delta$ and let $\Delta \to 0$:

$$\begin{aligned} H(y \,|\, x) &= \lim_{\Delta \to 0} P(Y \le y \,|\, x - \Delta < X < x + \Delta) \\ &= \lim_{\Delta \to 0} \frac{P(\{Y \le y\} \cap \{x - \Delta < X < x + \Delta\})}{P(x - \Delta < X < x + \Delta)}. \end{aligned} \tag{10.42}$$

The first question is whether this limit always exists and is finite. Notice that both the numerator and denominator are increasing functions of Δ, so each decreases to some limit as $\Delta \to 0$. If $P(X = x) = p > 0$, then by the continuity property of probability, the denominator tends to p, while the numerator tends to $P(X = x, Y \le y)$. Thus, the limit in (10.42) is $P(X = x, Y \le y)/p$. On the other hand, if $P(X = x) = 0$, then the numerator and denominator of Expression (10.42) both tend to 0 as $\Delta \to 0$. Nonetheless, the ratio cannot "blow up" because the numerator can never exceed the denominator. That is, the ratio in Expression (10.42) cannot exceed 1, so cannot have an infinite limit. Still, there could be two or more distinct limit points as $\Delta \to 0$, in which case the limit would not exist.

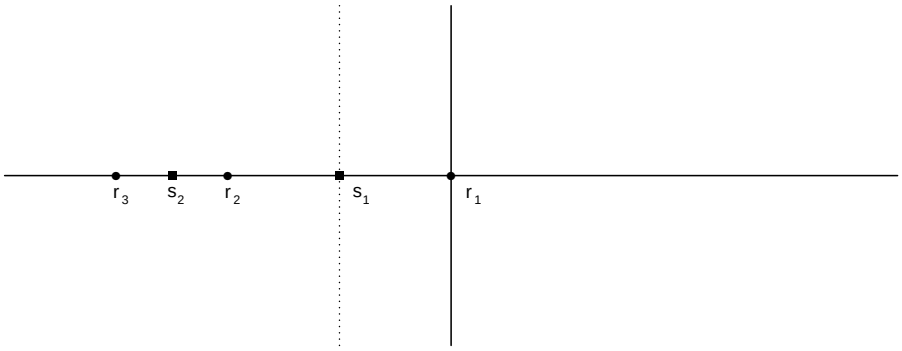

Figure 10.3: Points r_n (circles) and s_n (squares) in Example 10.58. Points strictly to the left of vertical lines represent $X < \Delta_n$ when Δ_n is r_n (solid line) or s_n (dashed line).

Example 10.58. Let $r_n = 1/2^n$, $n = 1, 2, \ldots$, and let $s_n = (r_n + r_{n+1})/2$, so that $r_{n+1} < s_n < r_n$, $n = 1, 2, \ldots$ (see Figure 10.3). Let X take value r_n with probability $1/2^{n+1}$ and s_n with probability $1/2^{n+1}$, $n = 1, 2, \ldots$ Define Y to be -1 if $X = r_n$ for some n, and 1 if $X = s_n$ for some n. Consider Expression (10.42) for $x = 0$, $y = 0$. Suppose that $\Delta_n = r_n$. Conditioned on $0 - \Delta_n < X < 0 + \Delta_n$, X can take any of the values $r_{n+1}, r_{n+2}, \ldots$ or $s_n, s_{n+1}, \ldots$ The conditional probability that X is one of $r_{n+1}, r_{n+2}, \ldots$ is

$$\begin{aligned} P(X \in \{r_{n+1}, r_{n+2}, \ldots\} \,|\, X < r_n) &= \frac{\sum_{i=n+1}^{\infty} P(X = r_i)}{\sum_{i=n+1}^{\infty} P(X = r_i) + \sum_{i=n}^{\infty} P(X = s_i)} \\ &= \frac{\sum_{i=n+1}^{\infty} 1/2^{i+1}}{\sum_{i=n+1}^{\infty} 1/2^{i+1} + \sum_{i=n}^{\infty} 1/2^{i+1}} \end{aligned}$$

$$= \frac{1/2^{n+1}}{1/2^{n+1} + 1/2^n} = 1/3.$$

Thus, $P(Y \leq 0 \,|\, -\Delta_n < X < \Delta_n) = 1/3$ for $\Delta_n = r_n$. If we now let $n \to \infty$, $P(Y \leq 0 \,|\, -\Delta_n < X < \Delta_n) \to 1/3$.

On the other hand, if $\Delta_n = s_n$, then X can take any of the values $r_{n+1}, r_{n+2}, \ldots$ or $s_{n+1}, s_{n+2}, \ldots$ Moreover, X is equally likely to be one of the r_i or one of the s_i. Therefore, if $\Delta_n = s_n$, then $P(Y \leq 0 \,|\, -\Delta_n < X < \Delta_n) = 1/2$. We have shown that $P(Y \leq 0 \,|\, -\Delta < X < 0)$ tends to $1/3$ if $\Delta \to 0$ along the path r_n, and tends to $1/2$ if $\Delta \to 0$ along the path s_n. Thus, the limit in Expression (10.42) does not exist at $x = 0$, $y = 0$. The fact that the limit may not exist highlights one problem with using Expression (10.42) as the definition of the conditional distribution of Y given $X = x$. □

Even if the limit in Expression (10.42) exists for a given x, it need not be a distribution function in y. The following example illustrates this.

Example 10.59. Let X be uniformly distributed on $(0, 1)$, and let $r \in (0, 1)$. Define Y to be $1/(X - r)$ if $X \neq r$, and 0 if $X = r$. Consider the conditional distribution of Y given $r - \Delta < X < r + \Delta$. Given $r - \Delta < X < r + \Delta$, X is as likely to be in $(r - \Delta, r)$ as it is to be in $(r, r + \Delta)$. Either way, $|Y| = 1/|X - r|$ is very large if Δ is very small; its sign is negative if $X \in (r - \Delta, r)$ and positive if $X \in (r, r + \Delta)$. Thus, Y is virtually guaranteed to be $\leq y$ if $X \in (r - \Delta, r)$ and virtually guaranteed to be $> y$ if $X \in (r, r + \Delta)$ for tiny Δ. Therefore, $P(Y \leq y \,|\, r - \Delta < X < r + \Delta) \to 1/2$ as $\Delta \to 0$ for each y. Clearly, $H(y \,|\, x)$ as defined by Expression (10.42) is not a distribution function in y because it does not satisfy $\lim_{y \to -\infty} H(y \,|\, x) = 0$ and $\lim_{y \to \infty} H(y \,|\, x) = 1$.

Because the limit in Expression (10.42) may either not exist or not be a distribution function in y for some x, it is problematic to define the conditional distribution function by Expression (10.42). Nonetheless, in Example 10.58, the set of x points such that the limit in Expression (10.42) does not exist has probability 0. In Example 10.59, the set of x points at which Expression (10.42) fails to converge to a distribution function has probability 0. These are not accidents.

Proposition 10.60. *(Pfanzagl, 1979)* **Existence of $H(y \,|\, x)$ of Expression (10.42)** *For any random variables (X, Y), the set N of x points such that (10.42) either fails to exist or fails to be a distribution function in y has $P(X \in N) = 0$.*

Thus, one actually could take Expression (10.42) as a definition of the conditional distribution function, and define the conditional distribution function to be some arbitrary F if the limit either does not exist or is not a distribution function in y for a given x. Although intuitive, this definition is avoided because it makes proofs more difficult.

Exercises

1. Suppose that (X, Y) has joint distribution function $F(x, y)$, and X has marginal distribution function $G(x)$. Suppose that, for a given (x, y), $\partial F / \partial x$ exists, and $G'(x)$ exists and is nonzero. What is Expression (10.42)?

10.9 Appendix: Radon-Nikodym Theorem

Theorem 10.61. Radon-Nikodym theorem *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space with $\mu(\Omega) < \infty$. If ν is a measure that is absolutely continuous with respect to μ (i.e., $A \in \mathcal{F}$ and $\mu(A) = 0$ implies that $\nu(A) = 0$ for each set $A \in \mathcal{F}$), then there exists a nonnegative measurable function $f(\omega)$ such that $\nu(A) = \int_A f(\omega) d\mu(\omega)$ for each $A \in \mathcal{F}$. The function f is unique in the sense that if g also has this property, then $g(\omega) = f(\omega)$ except on a set of μ measure 0.*

The function f described in the theorem is called the *Radon-Nikodym derivative* of ν with respect to μ.

To apply this theorem in the setting of conditional expectation, note that if $\mathrm{E}(|Y|) < \infty$, $\nu(A) = \mathrm{E}\{|Y|I(A)\}$ for $A \in \mathcal{F}$ defines a finite measure on $\mathcal{F}$. Therefore, there exists a nonnegative, measurable function $Z(\omega)$ such that $\nu(A) = \int_A Z(\omega) d\mu(\omega)$. That is, $\mathrm{E}\{|Y|I(A)\} = \mathrm{E}\{ZI(A)\}$. The same argument can be used for $\mathrm{E}\{Y^+ I(A)\}$ and $\mathrm{E}\{Y^- I(A)\}$.

10.10 Summary

Let Y be a random variable with $\mathrm{E}(|Y|) < \infty$, $\mathcal{A} \subset F$ be a sigma-field, and $Z = \mathrm{E}(Y \,|\, \mathcal{A})$.

1. **Definition** Z is unique up to equivalence with probability 1, and is defined by:
   (a) Z is $\mathcal{A}$-measurable.
   (b) $\mathrm{E}\{ZI(A)\} = \mathrm{E}\{YI(A)\}$ for all $A \in \mathcal{A}$.
2. **Function of X** If $\mathcal{A} = \sigma(\mathbf{X})$, the sigma-field generated by the random vector $\mathbf{X}$, then there is an extended Borel function f such that $Z = f(\mathbf{X})$ a.s.
3. **Geometry and prediction** Suppose that $Y \in L^2$, and let $L^2(\mathcal{A})$ be the space of $\mathcal{A}$-measurable random variables with finite second moment.
   (a) Z is the projection of Y onto $L^2(\mathcal{A})$. I.e., $Y - Z \perp X$ for each $X \in L^2(\mathcal{A})$.
   (b) Z minimizes $\mathrm{E}(Y - \hat{Y})^2$ over all $\hat{Y} \in L^2(\mathcal{A})$.
4. **Conditional distributions**
   (a) There is a probability measure $\mu(B, \omega)$ such that:
      i. For fixed $B \in \mathcal{B}$, $\mu(B, \omega)$ is a version of $P(Y \in B \,|\, \mathcal{A})$.
      ii. For fixed ω, $\mu(B, \omega)$ is a probability measure on the Borel sets $B \in \mathcal{B}$.
   (b) One version of $\mathrm{E}\{f(Y) \,|\, \mathcal{A}\}$ is $\int f(y) d\mu(y, \omega)$.
   (c) Inequalities for expectation hold for conditional expectation as well, but we must add "almost surely."
5. **Important conditional mean, variance, and covariance identities**
   If $\mathrm{E}(|Y|) < \infty$,

$$\begin{aligned} \mathrm{E}\{\mathrm{E}(Y \,|\, \mathcal{A})\} &= \mathrm{E}(Y) \\ \mathrm{E}(Y \,|\, \mathcal{A}) &= \mathrm{E}\{\mathrm{E}(Y \,|\, \mathcal{C}) \,|\, \mathcal{A}\} \quad \text{for } \mathcal{A} \subset \mathcal{C}. \end{aligned}$$

If $\mathrm{E}(Y^2) < \infty,\ \mathrm{E}(Y_1^2) < \infty,\ \mathrm{E}(Y_2^2) < \infty,$

$$\begin{aligned} \mathrm{var}(Y) &= \mathrm{var}\{\mathrm{E}(Y \mid \mathcal{A})\} + \mathrm{E}\{\mathrm{var}(Y \mid \mathcal{A})\} \\ \mathrm{cov}(Y_1, Y_2) &= \mathrm{cov}\{\mathrm{E}(Y_1 \mid \mathcal{A}), \mathrm{E}(Y_2 \mid \mathcal{A})\} + \mathrm{E}\{\mathrm{cov}(Y_1, Y_2 \mid \mathcal{A})\}. \end{aligned}$$

6. **Paradoxes**

   (a) Conditioning on a null set N makes sense only in a wider context of random variables/sigma-fields. N might be expressible in different ways (e.g., $X - Y = 0$ or $X/Y = 1$) that give different answers. Therefore, we condition on random variables or sigma-fields, not on a specific set of probability 0.

   (b) Substituting the value of a random variable in a conditioning statement is valid using a pre-specified regular conditional distribution function, but need not be valid for every version of the conditional expectation.