Notes on probability

James E. Pustejovsky

September 14, 2014

A conceptual understanding of probability and random variables will be of great help in building an understanding of the theory that supports causal inference methodology. In particular, the concepts of *conditional distribution*, *conditional expectation*, and *conditional independence* will play a central role. These notes will provide a far-from-rigorous overview of these and related concepts. For more rigorous but still very accessible textbooks that cover this material, see Grinstead and Snell (1998) or Ash (2008). An online introductory course is also available at `http://projects.iq.harvard.edu/stat110/home`.

1 Random variables

A ***random variables*** is a quantity determined by the outcome of a chance process. For example, the random variable C_1 might represent the outcome of tossing a fair coin, with $C_1 = 1$ if the coin comes up heads and $C_1 = 0$ if the coin comes up tails. Another example that will play an important role in this course is a random variable, say Y, that represents a quantitative measurement (e.g., an achievement test score) on a student selected using simple random sampling from a large population of students.

We will often consider a set of several random variables that represent distinct aspects of a common chance process. For instance, multiple measurements might be taken on the randomly selected student, with Y representing an achievement test score, X representing

the student's level of academic motivation, and W representing the student's family income level. For another example, consider a set of n students in a simple randomized experiment. We can represent the process of random assignment by define a random variable T_i for each student, $i = 1, ..., n$, where $T_i = 1$ if student i is assigned to the treatment condition and $T_i = 0$ if student i is assigned to the control condition.

Random variables are sometimes defined in terms of other, simpler random variables. For example, we might represent the outcomes of tossing a fair coin eight times using the random variables $C_1, C_2, ..., C_8$. The total number of heads is then

$$H = C_1 + C_2 + \cdots + C_8,$$

which is itself a random variable. (H follows a binomial distribution with success probability $\frac{1}{2}$ and 8 total trials.)

The ***support*** of a random variable is the set of all unique values that it can take. For example, the random variable C_1 representing the outcome of a coin toss has support $\{0, 1\}$, while the random variable H representing the number of heads in eight coin flips has support $\{0, 1, 2, 3, 4, 5, 6, 7, 8\}$.

If the support of a random variable contains only a countable (either finite or countably infinite) number of elements, then it is a ***discrete*** random variable. C_1 and H are both discrete because their supports are finite (with 2 elements and 9 elements, respectively). As an example of a discrete random variable with a countably infinite support, imagine buying a lottery ticket every day until you finally win, and let the random variable L be the number of losing tickets that you buy. $L = 0$ if you win on the first try, $L = 1$ if you win on the second try, etc. For any fixed, positive number k, it is possible that it takes exactly k tries to win. Therefore, the support of L is *all* integers greater than or equal to zero: $\{0, 1, 2, 3, 4, ...\}$.

Continuous random variables have supports that include ranges of real numbers, rather than just distinct values. For example, a standard normal random variable can take on *any*

real number, and thus has support $(-\infty, \infty)$. If we consider a basketball game a chance process, then the total time of possession (in minutes) for the home team might be described by a random variable with support $(0, 90)$. Of course, continuous random variables represent a level of abstraction from real life measurements because no quantity can ever really measured to arbitrary precision. For instance, time of possession in a basketball game may only be measured to the second, so perhaps it would be more realistic to treat the home team's time of possession as a discrete random variable, with support $\{0{:}00, 0{:}01, 0{:}02, ..., 90{:}00\}$. As another example, consider a test score such as the SAT. We might think about this score as a continuous random variable (i.e., model it with a normal distribution), but in fact it can take on only integer values between 600 and 2400. Because we can usually apply this sort of discretization argument, it will suffice for our purposes to ignore the complications that come with continuous random variables. **In what follows, we therefore treat all random variables as discrete.**

1.1 Distribution functions

The ***distribution function*** of a (discrete) random variable describes the probability that the random variable will take on a specified value. The distribution function for a random variable X tells us $\Pr(X = k)$ for any $k \in \mathcal{S}_X$.[1] The distribution function can then be used to determine the probability that the random variable satisfies a more general criterion. For example, for the random variable H representing the number of heads in eight coin tosses, we could determine the probability of getting no more than 3 heads by summing up the probabilities for all the outcomes that satisfy this criteria:

$$\Pr(H \leq 3) = \sum_{k=0}^{3} \Pr(H = k).$$

[1] We will use the notation $\Pr(X = k)$ to denote the probability that the random variable X takes on the specific value k. We will use the notation $k \in \mathcal{S}_X$ to denote "a value k in the support set $\mathcal{S}_X$ of the random variable X," or "k in the support of X" for short.

One property of distribution functions is that they sum to one across the support of the random variable:

$$\sum_{k \in \mathcal{S}_X} Pr(X = k) = 1.$$

Consider again a random variable Y representing the achievement test score of a student selected at random from a large population of students. The distribution function of Y corresponds to the distribution of the test score in the population, and $\Pr(Y = k)$ corresponds to the proportion of all students in the population that scored exactly k on the achievement test.

The ***joint distribution*** of several random variables describes the probability that each random variable takes on certain specified values. For example, suppose the random variable P represents home-team time of possession in a basketball game and the random variable D represents the point differential (home team score minus visiting team score) in the same basketball game. The joint distribution of (P, D) describes the probability that P takes on a value j and D takes on a value k, for specified value j in the support of P and specified value k in the support of D. We will denote this joint distribution by $\Pr(P = j, D = k)$, for $j \in \mathcal{S}_P$ and $k \in \mathcal{S}_D$.

As another example, suppose that we randomly select a student from a population of students and measure Y representing the student's achievement test score, X representing the student's level of academic motivation, and W representing the student's family income level. The joint distribution of (Y, X, W) is equivalent to the population distribution of all three variables, and $\Pr(Y = k, X = j, W = i)$ is the proportion of students that score k on the achievement test, have level of academic motivation j, and have family income level i.

Just as with single distribution functions, joint distributions have the property that they sum to one across the supports of *all* of the component random variables:

$$\sum_{k \in \mathcal{S}_Y} \sum_{j \in \mathcal{S}_X} \sum_{i \in \mathcal{S}_W} \Pr(Y = k, X = j, W = i) = 1.$$

Joint distributions can be used to determine the probability of more general criteria. For example, one could calculate the probability of randomly drawing a student with an achievement score lower than 800 who has low motivation and whose family income less than \$50K per year by summing the joint distribution function over all values of $k \in \mathcal{S}_Y$ where $k < 800$ and all values of $i \in \mathcal{S}_W$ where $i < 50$:

$$\Pr(Y < 800, X = \text{“}low\text{”}, W < 50) = \sum_{k<800} \sum_{i<50} \Pr(Y = k, X = \text{“}low\text{”}, W = i).$$

The ***marginal*** distribution of a random variable is its distribution without reference to any other random variables. This term is useful for distinguishing the joint distribution of multiple random variables (e.g., the joint distribution $\Pr(Y = k, X = j, W = i)$) from the distribution of individual random variables (e.g., the marginal distribution of Y is $\Pr(y = k)$). The marginal distribution of a random variable can be determined from a joint distribution by summing over the supports of all the other variables involved. For example:

$$\Pr(Y = k) = \sum_{j \in \mathcal{S}_X} \sum_{i \in \mathcal{S}_W} \Pr(Y = k, X = j, W = i).$$

This works for joint distributions of subsets of the random variables too:

$$\Pr(Y = k, X = j) = \sum_{i \in \mathcal{S}_W} \Pr(Y = k, X = j, W = i).$$

1.2 Independence

Two random variables X and Y are ***independent*** if knowing the value of one variable provides no information that would let you predict the value of the other. This corresponds to the mathematical property that the joint distribution of X and Y is equal to the product of the marginal distributions:

$$\Pr(X = j, Y = k) = \Pr(X = j) \times \Pr(Y = k)$$

for any $j \in \mathcal{S}_X$ and $k \in \mathcal{S}_Y$. More generally, a set of several random variables $X_1, ..., X_n$ is ***mutually independent*** if their joint distribution is equal to the product of the separate distributions of each

$$\Pr(X_1 = k_1, X_2 = k_2, ..., X_n = k_n) = \Pr(X_1 = k_1) \times \Pr(X_2 = k_2) \times \cdots \times \Pr(X_n = k_n),$$

$k_i \in \mathcal{S}_{X_i}$ for $i = 1, ..., n$. A shorthand way of denoting that two random variables X and Y are independent is: $X \perp\!\!\!\perp Y$.

Again consider randomly sampling a student from a population of students and measuring that student's achievement test score Y and level of academic motivation X. If X and Y are independent, then this would imply that knowing a student's level of motivation would *not* help you to predict whether that student had a higher or lower test score; similarly, knowing a student's test score would *not* help you to predict whether the student was motivated. (Does it seem plausible that X and Y would be independent in this example?)

1.3 Exercises

Suppose that you randomly select a student from a large population of 4th grade students and record the student's sex and number of siblings. Let $B = 1$ if the selected student is male and $B = 0$ if the selected student is female. Let S record the student's number of siblings. The proportion of the population of students falling into each category of B and S is recorded in the table below.

	Number of siblings				
Sex	0	1	2	3	4
Female	.10	.18	.12	.07	.04
Male	.12	.18	.14	.03	.02

1. What is the support of S?

2. What is the support of B?

3. Find the probability that the randomly selected student will be male with two siblings.

4. Find $\Pr(B = 0, S \leq 2)$.

5. Find the probability that the randomly selected student will be an only child.

6. Find the marginal distribution of B.

7. Find the marginal distribution of S.

8. Are B and S independent?

2 Expectation and Variance

The distribution function of a random variable provides a complete characterization of its behavior. However, often times it is simpler and more useful to work in terms of summary characteristics of the random variable. Two such summary characteristics are expectation (the mean value of the random variable) and variance (the variability around the mean value).

2.1 Expectation

The ***expectation*** of a random variable measures its mean value. The expectation of X will be denoted as $\mathrm{E}(X)$, or sometimes as μ_X (read: "mu-X") for short. The expectation of a discrete random variable X can be calculated from the distribution function of X as

$$\mathrm{E}(X) = \sum_{k \in \mathcal{S}_X} k \Pr(X = k).$$

As can be seen in the formula above, the expectation of X is a weighted-average of all of the values in its support, where the weights are equal to the probability that X takes on that value. As an aside, note that if a random variable does not have finite support, then it is possible that its expectation may not exist (because the infinite sum does not converge to a finite value).

Again consider randomly sampling a student from a population of students and measuring the student's achievement test score Y. Since the distribution of Y is equivalent to the distribution of test scores in the population, $\mathrm{E}(Y)$ is equal to the mean achievement score across all students in the population.

As another example, consider again the random variable C_1 that is equal to one if a coin comes up heads and equal to zero if a coin comes up tails. Suppose though that the coin is rigged so that the chance of a heads is some value $0 < p < 1$. C_1 is then a Bernoulli random variable with success probability p. The expectation of C_1 can be determined directly from

the formula:

$$\mathrm{E}(C_1) = 0 \times \Pr(C_1 = 0) + 1 \times \Pr(C_1 = 1) = \Pr(C_1 = 1) = p.$$

Three important properties of expectations have to do with the expection of a sum of several random variables and the expectation of a product of several random variables. First, the expectation of a sum of several random variables is equal to the sum of their expected values. Suppose we have random variables W, X, Y and arbitrary constants a, b, c. Then

$$\mathrm{E}\left(aW + bX + cY\right) = a\mathrm{E}(W) + b\mathrm{E}(X) + c\mathrm{E}(Y).$$

(This property also holds for an arbitrary number of random variables.) Second, the expectation of a constant (non-random) value is itself. It follows that $\mathrm{E}(Y + c) = \mathrm{E}(Y) + c$. Finally, if several random variables are mutually independent, then the expectation of their product is equal to the product of their expectations; however, this is not true if any of the variables are dependent. For instance, suppose that W, X, Y are mutually independent; then

$$E\left(W \times X \times Y\right) = \mathrm{E}(W) \times \mathrm{E}(X) \times \mathrm{E}(Y).$$

These properties are often useful for determining expectations. Consider the random variable H equal to the number of heads in eight coin flips, but now suppose that the coin is rigged so that the chance of a heads is p. We can use the fact that $H = \sum_{i=1}^{8} C_i$ for Bernoulli random variables $C_1, ..., C_8$ to easily determine the expectation of H:

$$\mathrm{E}(H) = \mathrm{E}\left(\sum_{i=1}^{8} C_i\right) = \sum_{i=1}^{8} \mathrm{E}(C_i) = \sum_{i=1}^{8} p = 8p.$$

2.2 Variance

The ***variance*** of a random variable is one measure of its variability, or the extent to which it deviates from its mean value. If a random variable has high variance, then it is noisy or unpredictable. The variance of a random variable X will be denoted as $\mathrm{Var}(X)$, or sometimes as σ_X^2 (read: "sigma-squared-X") for short. Mathematically, the variance X is defined as the expectation of the squared deviation of the random variable from its mean:

$$\mathrm{Var}(X) = \mathrm{E}\left[(X - \mu_X)^2\right] = \sum_{k \in \mathcal{S}_X} (k - \mu_X)^2 \Pr(X = k),$$

where $\mu_X = \mathrm{E}(X)$. Applying a little algebra and the properties of expectations, it can be seen that an equivalent expression for the variance is $\mathrm{Var}(X) = \mathrm{E}\,(X^2) - \mu_X^2$. Note that the variance of a random variable can never be negative because it is the expectation of the non-negative function $(X - \mu_X)^2$. (However, it is possible for a random variable with non-finite support to have infinite variance.)

Consider again the random variable C_1 in the rigged coin toss example. Recalling that $\mathrm{E}(C_1) = p$, the variance of C_1 can again be determined directly from the formula:

$$\mathrm{Var}(C_1) = (0-p)^2 \Pr(C_1 = 0) + (1-p)^2 \Pr(C_1 = 1) = p^2(1-p) + (1-p)^2 p = p(1-p).$$

The ***covariance*** of two random variables is a measure of the linear dependence between the variables. Higher covariance indicates that one variable can be more strongly predicted by a linear function of the other variable. Mathematically, the covariance of random variables X and Y is defined as

$$\mathrm{Cov}(X, Y) = \mathrm{E}\left[(X - \mu_X)\,(Y - \mu_Y)\right] = \mathrm{E}\,(XY) - \mu_X \mu_Y.$$

Note that the covariance of a random variable with itself is equivalent to the variance of the random variable: $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.

Having defined variance and covariance, we can now consider some of their algebraic properties.

- The variance of a random variable plus a constant is equal to the variance of the random variable alone; for arbitrary constant b, $\text{Var}(X + b) = \text{Var}(X)$.

- The variance of a sum of two random variables is equal to the sum of their variances and covariances:

$$\text{Var}(X + Y) = \text{Var}(X) + 2\text{Cov}(X, Y) + \text{Var}(Y).$$

- The variance of a constant times a random variable is equal to the square of the constant times the variance of the random variable: $\text{Var}(cX) = c^2\text{Var}(X)$.

- Combining the previous two properties, the covariance between two different sums of random variables is equal to the sum of the covariances between each combination of terms. For the constants $a_1, ..., a_m$ and $b_1, ..., b_n$ and the random variables $X_1, ..., X_m$ and $Y_1, ..., Y_n$:

$$\text{Cov}\left(\sum_{i=1}^{m} a_i X_i, \sum_{j=1}^{n} b_j Y_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j \text{Cov}(X_i, Y_j).$$

- The expectation of a product of two random variables (not necessarily independent) is equal to the product of their expectations plus their covariance: $\text{E}(XY) = \text{E}(X)\text{E}(Y) + \text{Cov}(X, Y)$. This property follows directly from the definition of covariance.

- If two random variables are independent, then their covariance is zero. However, zero covariances does not imply independence.

- It a set of random variables is mutually independent, then the variance of their sum is equal to the sum of their variances. For instance, suppose W, X, Y are mutually

independent and a, b, c are arbitrary constants. Then

$$\text{Var}(aW + bX + cY) = a^2\text{Var}(W) + b^2\text{Var}(X) + c^2\text{Var}(Y).$$

- The range of the covariance between two random variables is limited by their variances. In particular,

$$[\text{Cov}(X, Y)]^2 \leq \text{Var}(X) \times \text{Var}(Y).$$

  This relationship is known as the Cauchy-Schwarz inequality.

Consider again the random variable H equal to the number of heads in eight coin flips, assuming that the coin is rigged so that the chance of a heads is p. Because the eight coin flips (represented by random variables $C_1, ..., C_8$ are independent, the variance of H is equal to the sum of the variances of the individual flips:

$$\text{Var}(H) = \text{Var}\left(\sum_{i=1}^{8} C_i\right) = \sum_{i=1}^{8} \text{Var}(C_i) = \sum_{i=1}^{8} p(1-p) = 8p(1-p).$$

2.3 Mean and variance of a sample average

Suppose that we randomly select n students from a large population of students and measure their achievement scores, so that $Y_1, ..., Y_n$ are mutually independent random variables representing the achievement test scores of students 1 through n. The students are selected from a common population, implying that $Y_1, ..., Y_n$ all have the same distribution function, same expectation, and same variance. Let $\mu_Y = \text{E}(Y_i)$ and let $\sigma_Y^2 = \text{Var}(Y_i)$. Since Y_i is a random draw from the population, μ_Y corresponds to the mean achievement score in the population and σ_Y^2 corresponds to the population variance in achievement scores.

Now let the random variable M be the average of Y_1 through Y_n:

$$M = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

M is a constant times a sum of independent random variables, and so we can use the properties of expectation and variance to evaluate the expectation of M:

$$\mathrm{E}(M) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left(Y_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mu_Y = \mu_Y.$$

As you might anticipate, the expected value of the mean of independent, identically distributed random variables is equal to their common expected value, i.e., the population mean. It is interesting to note that the above derivation does not rely on the fact that $Y_1, ..., Y_n$ are mutually independent.

Now consider the variance of M:

$$\mathrm{Var}(M) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left(Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma_Y^2 = \frac{\sigma_Y^2}{n}.$$

The variance of a mean of independent, identically distributed random variables is equal to the population variance divided by the sample size, a result that should be familiar from introductory statistics courses.

2.4 Exercises

Suppose that you randomly select a student from a large population of 4th grade students and record the student's sex and number of siblings. Let $B = 1$ if the selected student is male and $B = 0$ if the selected student is female. Let S record the student's number of siblings. The proportion of the population of students falling into each category of B and S is recorded in the table below (the numbers are the same as in Section 1.3.

	Number of siblings				
Sex	0	1	2	3	4
Female	.10	.18	.12	.07	.04
Male	.12	.18	.14	.03	.02

1. Find $\mathrm{E}(B)$ and $\mathrm{Var}(B)$.

2. Find $\mathrm{E}(S)$ and $\mathrm{Var}(S)$.

3. Find $\mathrm{Cov}(B, S)$.

4. Suppose that instead of just one student, you sample 12 students. Let $S_1, ..., S_{12}$ represent the number of siblings reported by each student (you can assume that these random variables are independent). Let $\bar{S} = \sum_{h=1}^{12} S_h/12$ be the average number of siblings in the sample. Find $\mathrm{E}(\bar{S})$ and $\mathrm{Var}(\bar{S})$.

5. (Challenge) Consider the random variables A, B, and D, where B is defined in terms of the other two random variables as $B = A + D$. Suppose that $\mathrm{Var}(A) = \mathrm{Var}(B)$. What does this imply about $\mathrm{Cov}(A, D)$ and $\mathrm{Var}(D)$?

6. (Challenge) For independent random variables X and Y, find an expression for $\mathrm{Var}(X \times Y)$.

3 Conditional probability

Conditional probability has to do with how the probability distribution of (one or more) random variables change if we have partial knowledge about them. Conditioning on a random variable means to fix its magnitude at a specified value–to hold it constant. For example, consider a population of students, each of whom have achievement test scores Y and family income levels W. Suppose that we know the joint distribution of (Y, W) in this population. In the absence of further information, the probability that a student randomly selected from this population has a test score higher than 1000 can be evaluated from the distribution function:

$$\Pr(Y > 1000) = \sum_{k>1000} k \Pr(Y = k).$$

Now suppose that we randomly select a student and learn that their family income level is \$80K per year. What is the probability that this student has a test score higher than 1000? To answer this question, we need conditional probability. Specifically, we need hold family income constant at \$80K, that is, to look at the distribution of test scores among the *subset* of the population of students with family incomes of exactly \$80K.

For random variables W and Y, the conditional probability that $Y = k$ given that $W = i$ is denoted $\Pr(Y = k|W = i)$; everything to the right of the | symbol specifies the information given. The ***conditional distribution*** of Y given $W = i$ is the function that provides this conditional probability for each value of $k \in \mathcal{S}_Y$. Mathematically, the conditional distribution of Y given $W = i$ is given by the ratio of the joint distribution of Y and W to the marginal distribution of W, both evaluated at $W = i$:

$$\Pr(Y = k|W = i) = \frac{\Pr(Y = k, W = i)}{\Pr(W = i)},$$

for $k \in \mathcal{S}_Y$ and $i \in \mathcal{S}_W$. The same formula works if the conditioning event is more general. For instance, if we knew only that W was less than or equal to some fixed value h, we could

consider

$$\Pr(Y = k|W \leq h) = \frac{\Pr(Y = k, W \leq h)}{\Pr(W \leq h)}.$$

Conditional distributions can even be defined given some information about Y. For instance, suppose if we knew that Y was less than or equal to some fixed value g, we could consider

$$\Pr(Y = k|Y \leq g) = \frac{\Pr(Y = k, Y \leq g)}{\Pr(Y \leq g)}.$$

The joint probability in the numerator is equal to $\Pr(Y = k)$ if $k \leq g$, and is otherwise equal to zero.

Returning to the above example, we want to know $\Pr(Y > 1000|W = 80)$. If we know the joint distribution of (Y, W), then we can determine the marginal distribution of W, and thus find $\Pr(W = 80)$. Then

$$\Pr(Y > 1000|W = 80) = \frac{\Pr(Y > 1000, W = 80)}{\Pr(W = 80)} = \frac{\sum_{k>1000} \Pr(Y = k, W = 80)}{\Pr(W = 80)}.$$

Now suppose that we knew only that the student's family income is greater than \$60K, we could find the conditional probability that the student's test score is above 1000 by looking at the distribution of test scores in the subset of the population with family incomes larger than \$60K:

$$\Pr(Y > 1000|W > 60) = \frac{\Pr(Y > 1000, W > 60)}{\Pr(W > 60)} = \frac{\sum_{k>1000} \sum_{i>60} \Pr(Y = k, W = i)}{\sum_{i>60} \Pr(W = i)}.$$

Finally, suppose that we knew that the student's family income is \$80K and that their test score is less than 1300. We could find the conditional probability that the student's score is above 1000 by looking at the distribution of test scores in the subset of the population with test scores of less than 1300 (so ignoring all students that score 1300 or higher) and whose

family income is exactly \$80:

$$\begin{aligned}\Pr(Y > 1000 | W = 80, Y < 1300) &= \frac{\Pr(Y > 1000, W = 80, Y < 1300)}{\Pr(W = 80, Y < 1300)} \\ &= \frac{\sum_{1000<k<1300} \Pr(Y = k, W = 80)}{\sum_{k<1300} \Pr(Y = k, W = 80)}.\end{aligned}$$

Conditional distributions can be defined for several random variables together. For the random variables Y, X, W, the conditional distribution of Y and X, given $W = i$, is defined in terms of the joint distribution of all three variables and the marginal distribution of W:

$$\Pr(Y = k, X = j | W = i) = \frac{\Pr(Y = k, X = j, W = i)}{\Pr(W = i)}.$$

One reason that conditional distributions are interesting and useful is that they can be used to describe the joint distributions of a set of random variables. Specifically, joint distributions can be factored into the product of a conditional distributions. For the random variables Y, X, W, the joint distribution can be factored as

$$\Pr(Y = k, X = j, W = i) = \Pr(Y = k | X = j, W = i) \Pr(X = J | W = i) \Pr(W = i).$$

(It could also be factored in other ways, such as starting with the conditional distribution of X on Y and W.) This factoring is helpful if the conditional distributions are easier to describe or model.

3.1 Conditional distributions and independence

The conditional probability of Y given X describes how the probability of Y is affected by information about X (e.g., knowledge that $X = j$). If Y and X are independent, then knowing something about X provides no information about Y, and so the probability of Y should be unaffected. Mathematically, this can be seen by combining the definitions of

independence and conditional probability. If Y and X are independent, then

$$\Pr(Y = k|X = j) = \frac{\Pr(Y = k, X = j)}{\Pr(X = j)} = \frac{\Pr(Y = k)\Pr(X = j)}{\Pr(X = j)} = \Pr(Y = k).$$

For example, let Y represent the test score of a randomly selected student, and let Z represent the height of the student's teacher. The latter piece of information is totally irrelevant for purposes of predicting the student's test score. Therefore, the conditional distribution of the student's test score, given that the student's teacher is 5'7", is equal to the marginal distribution of test scores in the population of students: $\Pr(Y = k|Z = 5'7") = \Pr(Y = k)$ for $k \in \mathcal{S}_Y$.

3.2 Conditional independence

Two random variables X and Y are ***conditionally independent***, given that a third variable $W = i$, if the conditional distribution of X given $W = i$ is independent of the conditional distribution of Y given $W = i$. Recalling the definition of independence, conditional independence implies that

$$\Pr(Y = k, X = j|W = i) = \Pr(Y = k|W = i) \times \Pr(X = j|W = i).$$

Intuitively, this means that once we know the value of W, knowing the value of X does adds nothing to our abilility to predict Y and knowing the value of Y tells us nothing further about the value of X. Conditional independence is often stated symbolically as $X \perp\!\!\!\perp Y|W = i$. This statement is shortened to

$$X \perp\!\!\!\perp Y|W$$

if Y and X are conditionally independent for every $i \in \mathcal{S}_W$.

The conditional independence of X and Y, given W implies certain things about the

conditional distribution of Y, given X and W. In particular, If $X \perp\!\!\!\perp Y|W$ then

$$\Pr(Y = k|X = j, W = i) = \Pr(Y = k|W = i)$$

for all $k \in \mathcal{S}_Y$, $j \in \mathcal{S}_X$, and $i \in \mathcal{S}_W$.

3.3 Conditional expectation and variance

As with marginal distributions, it is often useful to summarize a conditional distribution using its expectation or variance. Both of these are defined just as for unconditional distributions. The ***conditional expectation*** of a random variable Y, given that another random variable X is equal to j, is calculated using the conditional distribution of Y given that $X = j$:

$$\mathrm{E}(Y|X = j) = \sum_{k \in \mathcal{S}_Y} k \Pr(Y = k|X = j).$$

This conditional expectation can be intepretted as the average value of Y across the subset of possible outcomes where $X = j$.

The conditional variance of Y, given that $X = j$, is also calculated using the conditional distribution. Like the marginal variance, it measures the average variability of Y around its mean, but now the mean is conditional on $X = j$ and the average is also conditional on $X = j$. Letting $\mu_{Y|X=j}$ denote $\mathrm{E}(Y|X = j)$,

$$\mathrm{Var}(Y|X = j) = \mathrm{E}\left[\left(Y - \mu_{Y|X=j}\right)^2 \middle| X = j \right] = \mathrm{E}\left(Y^2|X = j\right) - \mu^2_{Y|X=j}.$$

Like the conditional expectation, conditional variance can be interpretted as the variance of Y across the subset of possible outcomes where $X = j$.

Again consider randomly sampling a student from a population of students and measuring that students achievement test score Y, level of academic motivation X, and family income W. If we know that the student's level of motivation is 8 (on a 0-to-10 scale), then a good

prediction for her achievement test score is the average test score of all students who have the same level of motivation: $\mathrm{E}(Y|X = 8)$. If we also knew that her family income is \$40K, then we could incorporate this information as well and potentially obtain a better prediction: the average test score of all students who have the same level of motivation and the same level of family income, $\mathrm{E}(Y|X = 8, W = 40)$. How accurate would this prediction be? One way to quantify its predictive accuracy is through the conditional variance $\mathrm{Var}(Y|X = 8, W = 40)$, which measures the variability in the sub-population of students with $X = 8$ and $W = 40$.

3.4 Iterated expectations and variance decomposition

Conditional expectations and conditional variances can also be defined with respect to a random variable, rather than with respect to a specific value of a random variable. In the former case, we write $\mathrm{E}(Y|X)$ and $\mathrm{Var}(Y|X)$, without specifying a value for X. These conditionals have a subtly different interpretation than $\mathrm{E}(Y|X = j)$ and $\mathrm{Var}(Y|X = j)$. Rather than corresponding to fixed numbers, they are themselves random variables: the conditional expectation and conditional variance at a randomly selected value of X. As random variables, they have distribution functions. Specifically, the distribution function of $\mathrm{E}(Y|X)$ is equal to the value $\mathrm{E}(Y|X = j)$ with probability $\Pr(X = j)$, for $j \in \mathcal{S}_X$; similarly, $\mathrm{Var}(Y|X)$ is equal to $\mathrm{Var}(Y|X = j)$ with probability $\Pr(X = j)$.

These *random* conditional expectations and variances follow two very useful relationships. First, the expectation of $\mathrm{E}(Y|X)$ is simply Y. This fact is known as the ***law of iterated expectations***. From the definition of expectations,

$$\mathrm{E}\left[\mathrm{E}(Y|X)\right] = \sum_{j \in \mathcal{S}_X} \mathrm{E}(Y|X = j) \Pr(X = j) = \mathrm{E}(Y).$$

Evaluating the expectation of a conditional expectation essentially involves averaging in two steps: First, find the mean value of Y for each possible value of X; second, take a weighted average of these conditional averages, with weights corresponding to the probabilities of X.

Consider a population of students, and let $B = 1$ if a randomly selected student is male, with $B = 0$ otherwise. Suppose that we know that the average achievement test score of female students is $\mathrm{E}(Y|B = 0) = 1210$ and the average achievement test score of male students is $\mathrm{E}(Y|B = 1) = 1185$. Suppose that the population is 51% female. We can find the overall average achievement test score across all students by taking a weighted average of the average score for females and the average score for males, with weights corresponding to the population proportions of females and males, respectively:

$$\begin{aligned}\mathrm{E}(Y) &= \mathrm{E}(Y|B = 0)\Pr(B = 0) + \mathrm{E}(Y|B = 1)\Pr(B = 1) \\ &= 1210 \times .51 + 1185 \times .49 \\ &= 1197.75\end{aligned}$$

The second relationship states that the overall variance of a random variable can be decomposed into two parts, the expected value of the conditional variance and the variance of the conditional expectation:

$$\mathrm{Var}(Y) = \mathrm{E}\left[\mathrm{Var}(Y|X)\right] + \mathrm{Var}\left[\mathrm{E}(Y|X)\right].$$

This formula is called the ***variance decomposition***. The formula makes apparent why conditioning on available information is often useful. Consider predicting a randomly selected student's test score Y using the overall population average test score $\mathrm{E}(Y)$. $\mathrm{Var}(Y)$ is a measure of the inaccuracy of this prediction. If instead we predicted the student's test score Y using information about her level of motivation X, the prediction would be $\mathrm{E}(Y|X)$ and the expected inaccuracy of this prediction would be $\mathrm{E}\left[\mathrm{Var}(Y|X)\right]$. To the extent that motivation is a good predictor of test scores, then $\mathrm{E}(Y|X)$ will vary depending on X. The larger is this variability $\mathrm{Var}\left[\mathrm{E}(Y|X)\right]$, the smaller the expected inaccuracy $\mathrm{E}\left[\mathrm{Var}(Y|X)\right]$ because their sum is fixed at $\mathrm{Var}(Y)$.

3.5 Exercises

Suppose that you randomly select a student from a large population of 4th grade students and record the student's sex and number of siblings. Let $B = 1$ if the selected student is male and $B = 0$ if the selected student is female. Let S record the student's number of siblings. The proportion of the population of students falling into each category of B and S is recorded in the table below (the numbers are the same as in Sections 1.3 and 2.4.

	Number of siblings				
Sex	0	1	2	3	4
Female	.10	.18	.12	.07	.04
Male	.12	.18	.14	.03	.02

1. Create a table that lists $\Pr(S = h|B = i)$ for each $h \in \mathcal{S}_S$ and $i \in \mathcal{S}_B$.
2. Create a table that lists $\Pr(B = i|S = h)$ for each $i \in \mathcal{S}_B$ and $h \in \mathcal{S}_S$.
3. On the basis of these tables, how can you determine whether S and B are independent?
4. Find $\Pr(S \leq 2|B = 1)$.
5. Find $\Pr(B = 1|S > 0)$.
6. Find $\mathrm{E}(S|B = i)$ and $\mathrm{Var}(S|B = i)$ for each $i \in \mathcal{S}_B$.
7. How can you calculate $\mathrm{E}(S)$ from $\mathrm{E}(S|B = i)$?
8. Find $\mathrm{E}(B|S = h)$ and $\mathrm{Var}(B|S = h)$ for each $h \in \mathcal{S}_S$.
9. How can you calculate $\mathrm{E}(B)$ from $\mathrm{E}(B|S = h)$?
10. What is the support of $\mathrm{E}(S|B)$?
11. What is the support of $\mathrm{E}(B|S)$?
12. Find $\mathrm{Var}\left[\mathrm{E}(B|S)\right]$?

13. (Challenge) Prove that $X \perp\!\!\!\perp Y|W$ implies that $\Pr(Y = k|X = j, W = i) = \Pr(Y = k|W = i)$ by factoring the joint distribution of Y, W, X and applying the definition of conditional independence.

14. (Challenge) A sketch-proof of the law of iterated expectations is given below. Write down the principle or definition that justifies each step.

$$
\begin{aligned}
\mathrm{E}\left[\mathrm{E}(Y|X)\right] &= \sum_{j \in \mathcal{S}_X} \mathrm{E}(Y|X = j) \Pr(X = j) \\
&= \sum_{j \in \mathcal{S}_X} \sum_{k \in \mathcal{S}_Y} k \Pr(Y = k|X = j) \Pr(X = j) \\
&= \sum_{j \in \mathcal{S}_X} \sum_{k \in \mathcal{S}_Y} k \Pr(Y = k, X = j) \\
&= \sum_{k \in \mathcal{S}_Y} k \sum_{j \in \mathcal{S}_X} \Pr(Y = k, X = j) \\
&= \sum_{k \in \mathcal{S}_Y} k \Pr(Y = k) \\
&= \mathrm{E}(Y)
\end{aligned}
$$

3.6 Treatment assignment

Consider a set of n students who participate in a simple randomized experiment. Of the participants, n_A will be assigned to the control condition and $n_B = n - n_A$ will be assigned to the treatment condition, all with equal probabilities of assignment. Let $p = n_B/n$ be the proportion of students assigned to treatment. Define the random variables $T_1, ..., T_n$, where $T_i = 1$ if student i is assigned to the treatment condition and $T_i = 0$ if student i is assigned to the control condition. The probability that student i is assigned to the treatment condition is $\Pr(T_i = 1) = \mathrm{E}(T_i) = p$. However, note that the treatment assignments are not actually independent (because the sample size in each condition is fixed). Instead,

$$\Pr(T_i = 1 | T_j = 1) = \mathrm{E}(T_i | T_j = 1) = \frac{n_B - 1}{n - 1}$$
$$\Pr(T_i = 1 | T_j = 0) = \mathrm{E}(T_i | T_j = 0) = \frac{n_B}{n - 1}.$$

These can be written compactly as $\mathrm{E}(T_i|T_j) = (n_B - T_j)/(n-1)$. From this, we can determine $\mathrm{Cov}(T_i, T_j)$. First, if $i = j$ then $\mathrm{Cov}(T_i, T_j) = \mathrm{Var}(T_j) = \pi(1 - \pi)$. Second, for $i \neq j$,

$$\mathrm{E}(T_i T_j) = \mathrm{E}\left[\mathrm{E}(T_i T_j | T_j)\right] = \mathrm{E}\left[T_j \mathrm{E}(T_i | T_j)\right] = \frac{1}{n-1} \mathrm{E}\left[T_j (n_B - T_j)\right] = \frac{p(n_B - 1)}{n - 1}.$$

It follows that (for $i \neq j$),

$$\mathrm{Cov}(T_i, T_j) = \frac{p(n_B - 1)}{n - 1} - p^2 = \frac{-p(1-p)}{n - 1}.$$

The fact that $\mathrm{Cov}(T_i, T_j)$ is non-zero implies that treatment assignments T_i and T_j are not independent.

Say that each participating student would have test score a_i if assigned to the control condition, but would score b_i if assigned to the treatment condition. Note that these are fixed values, not random variables, because we are not sampling students from a larger population (we might say that the n participating students comprise the entire population). Define the

average outcomes under each condition as

$$\mu_A = \frac{1}{n}\sum_{i=1}^{n} a_i, \qquad \mu_B = \frac{1}{n}\sum_{i=1}^{n} b_i,$$

the variances under each condition as

$$\sigma_A^2 = \frac{1}{n-1}\sum_{i=1}^{n}(a_i - \mu_A)^2, \qquad \sigma_B^2 = \frac{1}{n-1}\sum_{i=1}^{n}(b_i - \mu_b)^2,$$

and the covariance of the outcomes under each condition as

$$\sigma_{AB} = \frac{1}{n-1}\sum_{i=1}^{n}(a_i - \mu_A)(b_i - \mu_b).$$

The observed test score for student i can be written as $Y_i = (1 - T_i)a_i + T_i b_i$, which is a random variable because it depends on T_i. The average outcome of students assigned to the control condition and the treatment condition can be written as

$$\bar{y}_A = \frac{1}{n_A}\sum_{i=1}^{n}(1 - T_i)Y_i, \qquad \bar{y}_B = \frac{1}{n_B}\sum_{i=1}^{n} T_i Y_i.$$

The usual estimate of the average treatment effect is the difference between the average outcomes in each condition, $\bar{y}_B - \bar{y}_A$. Using only the expressions for the mean and covariance of the treatment assignments $T_1, ..., T_n$, we can determine $\mathrm{E}\left(\bar{y}_B - \bar{y}_A\right)$ and $\mathrm{Var}\left(\bar{y}_B - \bar{y}_A\right)$.

First consider $\mathrm{E}(\bar{y}_B)$. Substituting in for Y_i,

$$\begin{aligned}
\mathrm{E}\left(\bar{y}_B\right) &= \mathrm{E}\left[\frac{1}{n_B}\sum_{i=1}^{n} T_i\left[(1-T_i)a_i + T_i b_i\right]\right] \\
&= \frac{1}{n_B}\mathrm{E}\left[\sum_{i=1}^{n} T_i b_i\right] \\
&= \frac{1}{n_B}\sum_{i=1}^{n} \mathrm{E}(T_i) b_i \\
&= \frac{1}{n_B}\sum_{i=1}^{n} \frac{n_B}{n} b_i \\
&= \frac{1}{n}\sum_{i=1}^{n} b_i = \mu_B.
\end{aligned}$$

Along the same lines, we find that $\mathrm{E}\left(\bar{y}_A\right) = \mu_A$. Thus, the expected value of the usual treatment effect estimate is $\mathrm{E}\left(\bar{y}_B - \bar{y}_A\right) = \mu_B - \mu_A$, i.e., the treatment effect estimate is unbiased for the difference in average outcomes under each condition.

Second, consider $\text{Var}(\bar{y}_B)$:

$$\begin{aligned}
\text{Var}\left(\bar{y}_B\right) &= \frac{1}{n_B^2}\text{Var}\left[\sum_{i=1}^{n} T_i b_i\right] \\
&= \frac{1}{n_B^2}\sum_{i=1}^{n}\sum_{j=1}^{n} b_i b_j \text{Cov}(T_i, T_j) \\
&= \frac{1}{n_B^2}\left[\sum_{i=1}^{n} b_i^2 \text{Var}(T_i) + \sum_{i=1}^{n}\sum_{j\neq i} b_i b_j \text{Cov}(T_i, T_j)\right] \\
&= \frac{1}{n_B^2}\left[\sum_{i=1}^{n} b_i^2 p(1-p) + \sum_{i=1}^{n}\sum_{j\neq i} b_i b_j \frac{p(1-p)}{n-1}\right] \\
&= \frac{p(1-p)}{n_B^2}\left[\sum_{i=1}^{n} b_i^2 - \frac{1}{n-1}\sum_{i=1}^{n}\sum_{j\neq i} b_i b_j\right] \\
&= \frac{p(1-p)}{n_B^2}\left[\frac{n}{n-1}\sum_{i=1}^{n} b_i^2 - \frac{1}{n-1}\sum_{i=1}^{n}\sum_{j=1}^{n} b_i b_j\right] \\
&= \frac{p(1-p)n}{n_B^2(n-1)}\left[\sum_{i=1}^{n} b_i^2 - n\mu_B^2\right] \\
&= \frac{p(1-p)n}{n_B^2}\sigma_B^2 \\
&= (1-p)\frac{\sigma_B^2}{n_B}.
\end{aligned}$$

Along the same lines, it can be established that $\text{Var}(\bar{y}_A) = p\frac{\sigma_A^2}{n_A}$ and $\text{Cov}(\bar{y}_A, \bar{y}_B) = \frac{\sigma_{AB}}{n}$. Putting the terms together,

$$\text{Var}(\bar{y}_B - \bar{y}_A) = (1-p)\frac{\sigma_B^2}{n_B} + p\frac{\sigma_A^2}{n_A} + 2\frac{\sigma_{AB}}{n}.$$

Because we never observe a_i and b_i on the same unit, σ_{AB} cannot be estimated. However, an upper bound for σ_{AB} is $\sigma_A\sigma_B$. Substituting in the upper bound, the variance of $\bar{y}_B - \bar{y}_A$ is bounded above by

$$\text{Var}(\bar{y}_B - \bar{y}_A) \leq (1-p)\frac{\sigma_B^2}{n_B} + p\frac{\sigma_A^2}{n_A} + 2\frac{\sigma_A\sigma_B}{n} = \frac{\sigma_B^2}{n_B} + \frac{\sigma_A^2}{n_A} - \frac{(\sigma_B - \sigma_A)^2}{n},$$

which can be estimated from the observed data. Note that this is less than or equal to the variance estimator in a two-sample t-test (assuming unequal variances). Note also that if the variances are equal ($\sigma_A^2 = \sigma_B^2$), then

$$\text{Var}(\bar{y}_B - \bar{y}_A) \leq \sigma_A^2 \left(\frac{1}{n_A} + \frac{1}{n_B} \right),$$

which is the usual variance estimator in a two-sample t-test, assuming equal variances. If the treatment effect is constant, so that $b_i - a_i = \delta$ for $i = 1, ..., n$, then $\sigma_{AB} = \sigma_A^2 = \sigma_B^2$, and so $\text{Var}(\bar{y}_B - \bar{y}_A)$ is exactly equal to the upper bound.

References

Ash, R. (2008). *Basic probability theory.* Mineola, NY: Dover Publications.

Grinstead, C., & Snell, J. (1998). *Introduction to probability.* Retrieved from http://www.dartmouth.edu/ chance/teaching_aids/books_articles/probability_book/amsbo