6 Dependence

This chapter treats features of a joint distribution which give insight into the nature of dependence between random variables. Sections 6.1 and 6.2 concern conditional distributions and expectations in the discrete case. Then parallel formulae for the density case are developed in Section 6.3. Covariance and correlation are introduced in Section 6.4. All these ideas are combined in Section 6.5 in a study of the bivariate normal distribution.

6.1 Conditional Distributions: Discrete Case

This section translates into the language of random variables the conditioning ideas of Section 1.4. The dependence between two variables X and Y can be understood in terms of the marginal distribution of X and the conditional distribution of Y given $X = x$, which may be a different distribution for each possible value x of X. Given this information, the distribution of Y is found by the rule of average conditional probabilities, and the conditional distribution of X given $Y = y$ is found by Bayes' rule.

J. Pitman, *Probability*

Example 1. Number of successes in a random number of trials.

Suppose a fair die is rolled. Then as many fair coins are tossed as there are spots showing on the die.

Problem 1. Find the distribution of the number of heads showing among the coins.

Solution. Let Y denote the number of heads showing among the coins. The problem is to calculate the probabilities

$$P(Y = y) = P(y \text{ heads}) \qquad (y = 0, 1, 2, \ldots, 6)$$

Let X represent the number showing on the die. If $X = x$, that is to say the die rolls x, then x coins are tossed, so the chance of y heads given the die rolls x is given by the binomial formula for the probability of y successes in x trials with probability $1/2$ of success on each trial:

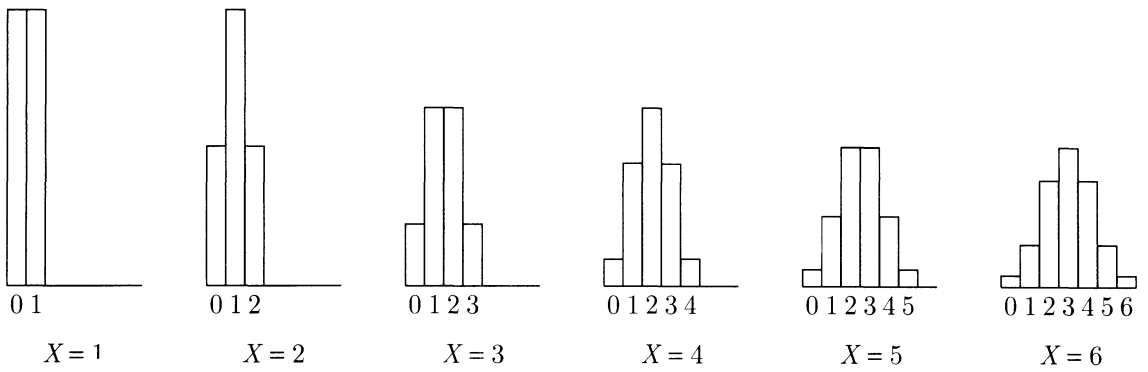
$$P(y \text{ heads} \,|\, \text{die rolls } x) = P(y \text{ heads in } x \text{ fair coin tosses}) = \binom{x}{y} 2^{-x}$$

where $\binom{x}{y} = 0$ if $x < y$. In random variable notation,

$$P(Y = y \,|\, X = x) = \binom{x}{y} 2^{-x}$$

This formula states that the *conditional distribution of* Y *given* $X = x$ is the binomial distribution with parameters $n = x$ and $p = 1/2$.

FIGURE 1. Conditional distribution of Y given $X = x$ for $x = 1, 2, \ldots, 6$ in Example 1.



The assumption that the die is fair specifies the unconditional distribution of X:

$$P(X = x) = P(\text{die rolls } x) = 1/6 \qquad (x = 1, 2, \ldots, 6)$$

These ingredients are combined by the rule of average conditional probabilities to give $P(Y = y)$, the unconditional probability of getting y heads:

$$\begin{aligned} P(Y = y) = P(y \text{ heads}) &= \sum_{x=1}^{6} P(\text{die rolls } x \text{ and } y \text{ heads}) \\ &= \sum_{x=1}^{6} P(y \text{ heads} \,|\, \text{die rolls } x) P(\text{die rolls } x) \\ &= \sum_{x=1}^{6} P(Y = y \,|\, X = x) P(X = x) \\ &= \frac{1}{6} \sum_{x=1}^{6} \binom{x}{y} 2^{-x} \qquad (0 \le y \le 6) \end{aligned}$$

where $\binom{x}{y} = 0$ if $x < y$. For example,

$$P(Y = 0) = \frac{1}{6}\left[\frac{1}{2} + \frac{1}{2^2} + \cdots + \frac{1}{2^6}\right] = \frac{1}{6} \times \frac{63}{64} = \frac{63}{384}$$

$$P(Y = 4) = \frac{1}{6}\left[\binom{4}{4}\frac{1}{2^4} + \binom{5}{4}\frac{1}{2^5} + \binom{6}{4}\frac{1}{2^6}\right] = \frac{29}{384}$$

and so on. Continuing in this way we obtain $P(Y = y)$ for each $y = 0, 1, 2, \ldots, 6$, as shown in Table 1.

TABLE 1. Probability $P(Y = y)$ of getting y heads.

y	0	1	2	3	4	5	6
$P(Y = y)$	$\frac{63}{384}$	$\frac{120}{384}$	$\frac{99}{384}$	$\frac{64}{384}$	$\frac{29}{384}$	$\frac{8}{384}$	$\frac{1}{384}$

Example 1 introduces the important idea of conditional distributions.

Conditional Distribution of Y Given $X = x$

For each possible value x of X, as y varies over all possible values of y, the probabilities $P(Y = y \,|\, X = x)$ form a probability distribution, depending on x, called the *conditional distribution of Y given $X = x$*.

The given value x of X can be thought of a as *parameter* in the distribution of Y given $X = x$. In Example 1, the distribution of Y given $X = x$ is the binomial distribution with parameters $n = x$ and $p = 1/2$.

According to the rule of average conditional probabilities, the unconditional distribution of Y, found in Example 1, is the *average* or *mixture* of these conditional

Example 1 introduces the important idea of conditional distributions.

Conditional Distribution of Y Given X = x

For each possible value x of X, as y varies over all possible values of y, the probabilities $P(Y = y \mid X = x)$ form a probability distribution, depending on x, called the *conditional distribution of* Y *given* $X = x$.

The given value x of X can be thought of as a *parameter* in the distribution of Y given $X = x$. In Example 1, the distribution of Y given $X = x$ is the binomial distribution with parameters $n = x$ and $p = 1/2$.

According to the rule of average conditional probabilities, the unconditional distribution of Y, found in Example 1, is the *average* or *mixture* of these conditional distributions, with equal weights $1/6$ defined by the uniform distribution of X. This distribution of Y may be called the *overall*, *marginal*, or *unconditional* distribution of Y, to distinguish it from the conditional distributions used to calculate it. The key step in the calculation of Example 1 was the following:

Rule of Average Conditional Probabilities

$$P(Y = y) = \sum_x P(Y = y \mid X = x)P(X = x)$$

This is just a basic rule of probability expressed in random variable notation. The rule holds for every pair of discrete random variables X and Y defined in the same probabilistic setting. The method of finding the distribution of a random variable Y by using this formula is called *conditioning on the value of* X. Note that in the sum for $P(Y = y)$ the term

$$P(Y = y \mid X = x)P(X = x) = P(X = x, Y = y)$$

is the generic entry in the joint probability table for X and Y. See Table 2 for example. You can use the above formula to calculate the distribution of a random variable Y if you can find a random variable X such that you either know or can easily calculate:

(i) the distribution of X;

(ii) the conditional probabilities $P(Y = y \mid X = x)$ for all possible values x of X.

If Y is determined by some two-stage or multistage process, the distribution of Y can often be calculated this way by letting X be the result of the first stage.

Example 1. (Continued.)

As in the previous example, let Y be the number of heads in X fair coin tosses, where X is uniformly distributed on $\{1, \ldots, 6\}$.

Problem 1. Find the conditional distribution of X given $Y = y$ for $y = 0, 1, \ldots 6$.

Solution. The problem now is to find $P(X = x \mid Y = y)$ as x varies, for each possible value y of Y. These conditional probabilities are calculated using Bayes' rule, as in Section 1.5. All that is new here is the random variable notation and terminology. As a start, the division rule for conditional probabilities gives

$$P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

where the joint probabilities

$$P(X = x, Y = y) = P(Y = y \mid X = x)P(X = x)$$

are the individual terms in the sum used previously to calculate $P(Y = y)$. Substituting the values of $P(X = x)$ and $P(Y = y \mid X = x)$, the joint probabilities $P(X = x, Y = y)$ are displayed in Table 2.

TABLE 2. Joint distribution table for (X, Y).

		Possible values x for X						Marginal
		1	2	3	4	5	6	distn. of Y
Possible values y for Y	0	$\frac{1}{6}\frac{1}{2}$	$\frac{1}{6}\frac{1}{4}$	$\frac{1}{6}\frac{1}{8}$	$\frac{1}{6}\frac{1}{16}$	$\frac{1}{6}\frac{1}{32}$	$\frac{1}{6}\frac{1}{64}$	$\frac{63}{384}$
	1	$\frac{1}{6}\frac{1}{2}$	$\frac{1}{6}\frac{2}{4}$	$\frac{1}{6}\frac{3}{8}$	$\frac{1}{6}\frac{4}{16}$	$\frac{1}{6}\frac{5}{32}$	$\frac{1}{6}\frac{6}{64}$	$\frac{120}{384}$
	2	0	$\frac{1}{6}\frac{1}{4}$	$\frac{1}{6}\frac{3}{8}$	$\frac{1}{6}\frac{6}{16}$	$\frac{1}{6}\frac{10}{32}$	$\frac{1}{6}\frac{15}{64}$	$\frac{99}{384}$
	3	0	0	$\frac{1}{6}\frac{1}{8}$	$\frac{1}{6}\frac{4}{16}$	$\frac{1}{6}\frac{10}{32}$	$\frac{1}{6}\frac{20}{64}$	$\frac{64}{384}$
	4	0	0	0	$\frac{1}{6}\frac{1}{16}$	$\frac{1}{6}\frac{5}{32}$	$\frac{1}{6}\frac{15}{64}$	$\frac{29}{384}$
	5	0	0	0	0	$\frac{1}{6}\frac{1}{32}$	$\frac{1}{6}\frac{6}{64}$	$\frac{8}{384}$
	6	0	0	0	0	0	$\frac{1}{6}\frac{1}{64}$	$\frac{1}{384}$
	Marginal distn. of X	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

In column x of the table you see numbers proportional to the binomial $(x, 1/2)$ probabilities forming the conditional distribution of Y given $X = x$. The constant of proportionality is $1/6$, which is the marginal probability of $(X = x)$. Similarly, in row y of the table you see numbers proportional to the conditional distribution of X given $Y = y$. The conditional probabilities themselves are obtained by dividing the numbers in the row y by the constant factor $P(Y = y)$, their sum, which appears in the margin. For example, the conditional distribution of X given $Y = 2$ is displayed in Table 3.

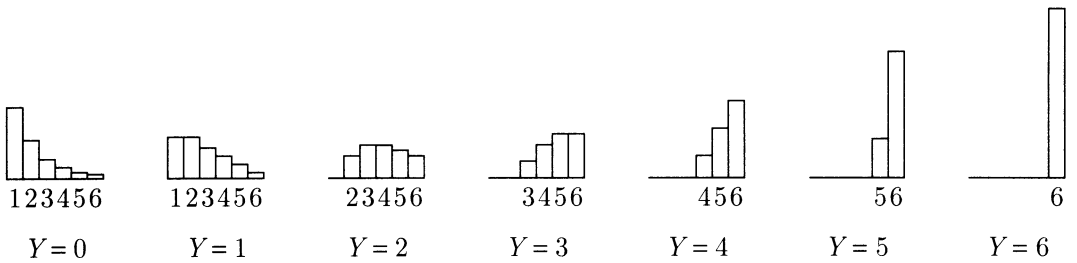
TABLE 3. Conditional distribution of X given $Y = 2$.

x	1	2	3	4	5	6
$P(X = x \mid Y = 2)$	0	$\frac{16}{99}$	$\frac{24}{99}$	$\frac{24}{99}$	$\frac{20}{99}$	$\frac{15}{99}$

So, given two heads, the number of coins tossed is equally likely to be either 3 or 4, and these are the most likely values.

Similar tables of the conditional distributions are easily made for other values y of Y. Here is a graphical display of all seven of these conditional distributions using histograms.

FIGURE 3. Conditional distribution of X given $Y = y$.



Exercises 6.1

1. Suppose I toss three coins. Some of them land heads and some land tails. Those that land tails I toss again. Let X be the number of heads showing after the first tossing, Y the total number showing after the second tossing, including the X heads appearing on the first tossing. So X and Y are random variables such that $0 \le X \le Y \le 3$ no matter how the coins land. Write out distribution tables and sketch histograms for each of the following distributions:

a) the distribution of X;

b) the conditional distribution of Y given $X = x$ for $x = 0, 1, 2, 3$;

c) the joint distribution of X and Y (no histogram in this case);

d) the distribution of Y;

e) the conditional distribution of X given $Y = y$ for $y = 0, 1, 2, 3$.

f) What is the best guess of the value of X given $Y = y$ for $y = 0, 1, 2, 3$? That is, for each y, choose x depending on y to maximize $P(X = x|Y = y)$.

g) Suppose the random experiment generating X and Y is repeated independently over and over again. Each time you observe the value of Y, and then guess the value of X using the rule found in f). Over the long run, what proportion of times will you guess correctly?

2. In a particular town 10% of the families have no children, 20% have one child, 40% have two children, 20% have three children, and 10% have four. Let T represent the total number of children, and G the number of girls, in a family chosen at random from this town. Assuming that children are equally likely to be boys or girls, find the distribution of G. Display your answer in a table and sketch the histogram.

3. Suppose the names of all the children in the town of Exercise 2 are put into a hat, and a name is picked out at random. So now a child is picked at random instead of a family being picked at random. Let U be the total number of children in the family of the child chosen at random.

a) Find the distribution of U. Why is this distribution different from the distribution of T in Exercise 2?

b) What is the probability that the child picked at random comes from a family consisting of two girls and a boy?

c) Is this the same as the probability that a family picked at random consists of two girls and a boy? Calculate and explain.

4. Let $A_1, \ldots, A_{20}$ be independent events each with probability $1/2$. Let X be the number of events among the first 10 which occur and let Y be the number of events among the last 10 which occur. Find the conditional probability that $X = 5$, given that $X + Y = 12$.

5. Let X_1 and X_2 be independent Poisson random variables with parameters λ_1 and λ_2.

a) Show that for every $n \geq 1$, the conditional distribution of X_1, given $X_1 + X_2 = n$, is binomial, and find the parameters of this binomial distribution.

b) The number of eggs laid by a certain kind of insect follows a Poisson distribution quite closely. It is known that two such insects have laid their eggs in a particular area. If the total number of eggs in the area is 150, what is the chance that the first insect laid at least 90 eggs? (State your assumptions, and give approximate decimal answer.)

6. Conditioning independent Poisson variables on their sum. Let N_i be independent Poisson variables with parameters λ_i. Think of the N_i as the number of points of a Poissson scatter in disjoint parts of the plane with areas λ_i, where the mean intensity is one point per unit area.

a) What is the conditional joint distribution of $(N_1, \ldots, N_m)$ given $N_1 + \ldots + N_m = n$? [*Hint*: See Exercise 5 for a special case.]

b) Suppose now that N has Poisson(λ) distribution, and given $N = n$ the conditional joint distribution of some m-tuple of random variables $(N_1, \ldots, N_m)$ is exactly what you found in part a). What can you conclude about the unconditional distribution of $(N_1, \ldots, N_m)$?

7. **Poissonization of the binomial distribution.** Let N have Poisson (λ) distribution. Let X be a random variable with the following property: for every n, the conditional distribution of X given $(N = n)$ is binomial (n, p).

a) Show that the unconditional distribution of X is Poisson, and find its parameter.

It is known that X-rays produce chromosome breakages in cells. The number of such breakages usually follows a Poisson distribution quite closely, where the parameter depends on the time of exposure, etc. For a particular dosage and time of exposure, the number of breakages follows the Poisson (0.4) distribution. Assume that each breakage heals with probability 0.2, independently of the others.

b) Find the chance that after such an X-ray, there are 4 healed breakages.

8. **Independence in Poissonization of the binomial distribution.** Suppose you roll a random number of dice. If the number of dice follows the Poisson (λ) distribution, show that the number of sixes is independent of the number of nonsixes. [*Hint*: Let N be the number of dice, X the number of sixes, and Y the number of nonsixes. Exercise 7 gives you the marginal distributions of X and Y. To show that the joint distribution of X and Y is the product of the marginals, show

$$P(X = x, Y = y) = P(N = x + y, X = x, Y = y)$$

and then use the multiplication rule.]

9. **Conditional independence.** Random variables X and Y are called *conditionally independent given* Z if given the value of Z, X, and Y are independent. That is,

$$P(X = x, Y = y \,|\, Z = z) = P(X = x \,|\, Z = z) P(Y = y \,|\, Z = z)$$

for all possible values x, y, and z. Prove that X and Y are conditionally independent given Z if and only if the conditional distribution of Y given $X = x$ and $Z = z$ is a distribution which depends only on z:

$$P(Y = y \,|\, X = x, Z = z) = P(Y = y \,|\, Z = z)$$

for all possible values x, y, and z. Give a further equivalent condition in terms of the conditional distribution of X given $Y = y$ and $Z = z$.

10. **Conditional independence (continued).** Suppose as in Example 5 of Section 3.1 that two sequences of n draws with replacement are made from a box containing an unknown number of red tickets among a total of 10 tickets. Regard the number of red tickets in the box as the value of a random variable R, with probability distribution $P(R = r) = \pi_r$, $r = 0, 1, \ldots, 10$. Let X_1 be the number of red tickets in the first n draws, and X_2 the number in the second n draws. Assuming that X_1 and X_2 are conditionally independent and binomially distributed given $R = r$, find expressions for the following:

a) $P(R = r, X_1 = x_1, X_2 = x_2)$; b) $P(R = r \,|\, X_1 = x_1)$;

c) $P(X_2 = x_2 \,|\, R = r, X_1 = x_1)$; d) $P(X_2 = x_2 \,|\, X_1 = x_1)$.

e) Calculate numerical values for the conditional probabilities in d) assuming that $\pi_r = 1/11$ for $r = 0, 1, \ldots, 10$ and $n = 1$. Are X_1 and X_2 independent?

6.2 Conditional Expectation: Discrete Case

Conditional expectations are simply expectations relative to conditional distributions.

Conditional Expectation Given an Event

The *conditional expectation of a random variable Y given an event A*, denoted by $E(Y \mid A)$, is the expectation of Y under the conditional probability distribution given A:

$$E(Y \mid A) = \sum_{\text{all } y} yP(Y = y \mid A)$$

This is just the definition of $E(Y)$, with probabilities replaced by conditional probabilities given A. Intuitively, $E(Y \mid A)$ is the expected value of Y, given the information that event A has occurred.

Example 1. **Conditioning on at most 2 heads on 4 coin tosses.**

Let Y be the number of heads in four tosses of a fair coin. Calculate the conditional expectation of Y given 2 or less heads. What is the long-run interpretation of this quantity?

Solution. Here the conditioning event is $A = (Y \leq 2)$. Since Y has the binomial $(4, \frac{1}{2})$ distribution

$$P(Y = y) = \binom{4}{y} \Big/ 2^4 \qquad (y = 0 \text{ to } 4)$$

$$P(Y \leq 2) = (1 + 4 + 6)/16 = 11/16$$

Hence

$$P(Y = y \mid Y \leq 2) = \binom{4}{y} \Big/ 11 \qquad (y = 0, 1, 2)$$

and

$$E(Y \mid Y \leq 2) = \sum_{y=0}^{2} y \binom{4}{y} \Big/ 11 = (1 \cdot 4 + 2 \cdot 6)/11 = 16/11$$

The long-run interpretation is that if you repeatedly toss four fair coins, the long-run average number of heads, averaging only over the trials that produce 0, 1, or 2 heads, will be 16/11.

Properties of Conditional Expectation

For a fixed conditioning event A, conditional expectation has familiar properties of expectation like linearity. For instance, there is the addition rule

$$E(X+Y\,|\,A) = E(X\,|\,A) + E(Y\,|\,A)$$

and so on. For a fixed random variable Y, as A varies, there is a useful generalization of the rule of average conditional probabilities, a rule of *average conditional expectations*: If $A_1, \ldots, A_n$ is a partition of the whole outcome space, then

$$E(Y) = \sum_{i=1}^{n} E(Y\,|\,A_i)P(A_i)$$

In the special case when Y is an indicator random variable, say $Y = I_B$, the indicator of event B, this reduces to the rule of average conditional probabilities

$$P(B) = \sum_{i=1}^{n} P(B\,|\,A_i)P(A_i)$$

The general case can be derived from this special case by linear operations. It is most convenient for applications to express the general rule as follows, for the partition generated by values of a discrete random variable X:

Rule of Average Conditional Expectations

For any random variable Y with finite expectation and any discrete random variable X,

$$E(Y) = \sum_{\text{all } x} E(Y\,|\,X = x)P(X = x)$$

This formula is also called the *formula for $E(Y)$ by conditioning on X*. This formula gives a useful method of calculating expectations, as shown by the examples below. The next box introduces a useful short notation:

Definition of $E(Y|X)$

The *conditional expectation of Y given X*, denoted $E(Y\,|\,X)$, is the function of X whose value is $E(Y\,|\,X = x)$ if $X = x$.

Here $E(Y \mid X)$ is actually a random variable, since by definition it is a particular function of X, and a function of a random variable defines another random variable. It can be shown that $E(Y \mid X)$ is the best predictor of Y based on X, in the sense of mean-square error. That is to say, $E(Y \mid X)$ is the function $g(X)$ that minimizes the mean square prediction error $E[(Y - g(X))^2]$. See Exercise 17. Because $E(Y \mid X)$ is a random variable, it makes sense to consider its expectation. The result is stated in the next box.

Expectation is the Expectation of the Conditional Expectation

$$E(Y) = E\,[E(Y \mid X)]$$

This is a condensed form of the rule of average conditional expectations, obtained by application to $g(x) = E(Y \mid X = x)$ of the formula

$$E[g(X)] = \sum_{\text{all } x} g(x)P(X = x)$$

Examples

Example 2. **Tossing a random number of coins.**

As in Example 1 of the previous section, let Y be the number of heads in X tosses of a fair coin, where X is generated by a fair die roll.

Problem 1. Find the conditional expectation of Y given $X = x$.

Solution. Since the conditional distribution of Y given $X = x$ is binomial with parameters $n = x$ and $p = 1/2$, the conditional expectation of Y given $X = x$ is the mean of the binomial(n, p) distribution, that is np, for $n = x$ and $p = 1/2$:

$$E(Y \mid X = x) = x/2 \qquad (x = 1, 2, \ldots, 6)$$

Problem 2. Find $E(Y)$.

Solution. Since from the previous solution $E(Y \mid X) = X/2$, and $E(X) = 3.5$

$$E(Y) = E[E(Y \mid X)] = E(X/2) = E(X)/2 = (3.5)/2 = 1.75$$

Discussion. Of course, the expectation of Y can also be calculated from the distribution of Y, shown in Table 1 of Section 6.1. But the method of conditioning on X gives the result more quickly. Also, the method of computing $E(Y)$ by conditioning on a

suitable random variable X can be applied in problems where it is difficult to obtain a formula for the distribution of Y.

Problem 3. Find $E(X \mid Y = 2)$

Solution. There is no simple formula for $E(X \mid Y = y)$ as a function of y in this problem. But these conditional expectations can be calculated one by one from the various conditional distributions of X given $Y = y$ for $y = 0$ to 6. Using the conditional distribution of X given $Y = 2$ displayed in Table 3 of Section 6.1 gives

$$E(X \mid Y = 2) = (2{\times}16 + 3{\times}24 + 4{\times}24 + 5{\times}20 + 6{\times}15)/99 = 390/99 \approx 3.94$$

Example 3. Number of girls in a family.

Suppose the number of children in a family is a random variable X with mean μ, and given $X = n$ for $n \geq 1$, each of the n children in the family is a girl with probability p and a boy with probability $1 - p$.

Problem. What is the expected number of girls in a family?

Solution. Intuitively, the answer should be $p\mu$. To show this is correct, let G be the random number of girls in a family. Given $X = n$, G is the sum of n indicators of events with probability p, so

$$E(G \mid X = n) = np$$

Note that this is correct even for $n = 0$. By conditioning on X,

$$E(G) = \sum_n E(G \mid X = n)P(X = n) = p \sum_n nP(X = n) = p\mu$$

Remark. In short notation,

$$E(G \mid X) = pX$$

$$E(G) = E\,[E(G \mid X)] = E(pX) = pE(X)$$

Example 4. Success counts in overlapping series of trials.

Let S_n be the number of successes in n independent trials with probability p of success on each trial.

Problem. Calculate $E(S_m \mid S_n = k)$ for $m \leq n$.

Solution. Since $S_m = X_1 + \cdots + X_m$ where X_j is the indicator of success on the jth trial

$$E(S_m \mid S_n = k) = \sum_{j=1}^{m} E(X_j \mid S_n = k) \quad \text{where}$$

$$\begin{aligned} E(X_j \mid S_n = k) &= P(j\text{th trial is a success} \mid S_n = k) \\ &= \frac{P(j\text{th trial is a success, } S_n = k)}{P(S_n = k)} \\ &= \frac{P(j\text{th trial success, } k-1 \text{ of other } n-1 \text{ trials are successes})}{P(S_n = k)} \\ &= \frac{p\binom{n-1}{k-1}p^{k-1}(1-p)^{n-k}}{\binom{n}{k}p^k(1-p)^{n-k}} \qquad \begin{array}{l}\text{using independence and} \\ \text{the binomial distribution}\end{array} \\ &= \frac{k}{n} \quad \text{so} \\ E(S_m \mid S_n = k) &= \frac{mk}{n} \end{aligned}$$

Discussion. In short notation, the conclusion is that for $1 \le m \le n$

$$E(S_m \mid S_n) = \frac{m}{n} S_n$$

This is a rather intuitive formula. It says that given S_n successes in n trials, the number of successes to be expected in m of the trials is proportional to m. The formula can be derived in other ways. By symmetry, $E(X_j \mid S_n)$ must be the same for all j, and equal to $E(X_1 \mid S_n)$. Since

$$S_n = E(S_n \mid S_n) = \sum_{j=1}^{n} E(X_j \mid S_n) = nE(X_1 \mid S_n)$$

it follows that $E(X_1 \mid S_n) = S_n/n$ and hence

$$E(S_m \mid S_n) = \sum_{j=1}^{m} E(X_j \mid S_n) = mE(X_1 \mid S_n) = \frac{m}{n} S_n$$

This argument shows that formula

$$E(S_m \mid S_n) = \frac{m}{n} S_n \qquad (1 \le m \le n)$$

holds whenever S_n is a sum of n independent and identically distributed variables $X_1, \ldots, X_n$. In fact all that is required is that the variables $X_1, \ldots, X_n$ are *exchangeable*, as defined in Section 3.6. This is an example where a conditional expectation can be calculated using symmetry and linearity, even though there is no nice formula for the conditional distribution.

Treating a conditioned variable as a constant. When computing conditional probabilities or expectations given $X = x$, the random variable X may be treated as if it were the constant x. Intuitively, this is quite obvious: on the restricted outcome space $(X = x)$, the random variable X has only one value, namely, x. To illustrate, if g is a function of two random variables X and Y, the conditional distribution of $g(X, Y)$ given $X = x$, is the same as the conditional distribution of $g(x, Y)$ given $X = x$. And if g has numerical values

$$E[g(X,Y) \mid X = x] = E[g(x,Y) \mid X = x]$$

For instance

$$E[XY \mid X = x] = E[xY \mid X = x] = xE[Y \mid X = x]$$

which reads in short notation

$$E[XY \mid X] = XE[Y \mid X]$$

Another example is

$$E[aX + bY \mid X = x] = E[ax + bY \mid X = x] = ax + bE[Y \mid X = x]$$

which reads in short notation

$$E[aX + bY \mid X] = aX + bE[Y \mid X]$$

Example 5. Conditional expectation of a sum given one of the terms.
Suppose X and Y are independent.

Problem. Find $E(X + Y \mid X = x)$.

Solution.

$$\begin{aligned} E(X + Y \mid X = x) &= E(X \mid X = x) + E(Y \mid X = x) \\ &= x + E(Y) \end{aligned}$$

Here $E(X \mid X = x) = x$ because X may be treated as the constant x given $X = x$. And $E(Y \mid X = x)$ is the mean of the conditional distribution of Y given $X = x$, and by independence this is just the unconditional distribution of Y with mean $E(Y)$.

Exercises 6.2

1. Let X_1 and X_2 be the numbers on two independent fair-die rolls. Let X be the minimum and Y the maximum of X_1 and X_2. Calculate: a) $E(Y|X = x)$; b) $E(X|Y = y)$.

2. Repeat Exercise 1 above, with X_1 and X_2 independent and uniformly distributed on $\{1, 2, \ldots, n\}$.

3. Repeat Exercise 1 with X_1 and X_2 two draws without replacement from $\{1, 2, \ldots, n\}$.

4. An item is selected randomly from a collection labeled $1, 2, ..., n$. Denote its label by X. Now select an integer Y uniformly at random from $\{1, \ldots, X\}$. Find:

a) $E(Y)$; b) $E(Y^2)$; c) $SD(Y)$; d) $P(X + Y = 2)$.

5. Suppose an event A is independent of a pair of random variables X_1 and X_2, whose c.d.f's are F_1 and F_2. Define a random variable X by:

$$X = \begin{cases} X_1 & \text{if } A \text{ occurs} \\ X_2 & \text{if } A \text{ does not occur} \end{cases}$$

Find and justify formulae for:

a) the c.d.f. $F(x)$ of X, in terms of $F_1(x)$, $F_2(x)$, and $p = P(A)$;

b) $E(X)$ in terms of $E(X_1)$, $E(X_2)$, and p.

c) $Var(X)$ in terms of $E(X_1)$, $E(X_2)$, $Var(X_1)$, $Var(X_2)$ and p.

6. Suppose that N is a Poisson random variable with parameter μ. Suppose that given $N = n$, random variables $X_1, X_2...., X_n$ are independent with uniform $(0, 1)$ distribution. So there are a random number of X's.

a) Given $N = n$, what is the probability that all the X's are less than t?

b) What is the (unconditional) probability that all the X's are less than t?

c) Let $S_N = X_1 + \cdots + X_N$ denote the sum of the random number of X's. (If $N = 0$ then $S_N = 0$.) Find $P(S_N = 0)$. Explain.

d) Find $E(S_N)$.

7. Suppose that N is a counting random variable, with values $\{0, 1, \ldots, n\}$, and that given $(N = k)$, for $k \geq 1$, there are defined random variables $X_1, \ldots, X_k$ such that

$$E(X_j | N = k) = \mu \qquad (1 \leq j \leq k)$$

Define a random variable S_N by

$$S_N = \begin{cases} X_1 + X_2 + \cdots + X_k & \text{if } (N = k),\ 1 \leq k \leq n \\ 0 & \text{if } (N = 0) \end{cases}$$

Show that $E(S_N) = \mu E(N)$.

8. Suppose that each individual in a population produces a random number of children, and the distribution of the number of children has mean μ. Starting with one individual, show, using the result of Exercise 7, that the expected number of descendants of that individual in the nth generation is μ^n.

9. Let T_i be the place at which the ith good element appears in a random ordering of $N - k$ bad elements and k good ones. Use the results of Exercise 3.6.13 to calculate:

a) $E(T_1 | T_2 = j)$; b) $E(T_2 | T_1 = j)$;

c) $E(T_h|T_i = j)$ first for $h < i$, then for $h > i$.

10. What is the expected number of black balls among $n \leq b+w+d$ balls drawn at random from a box containing b black balls, w white balls, and d balls drawn at random from another box of b_0 black balls and w_0 white balls? Assume all draws are made without replacement.

11. A deck of cards is cut into two halves of 26 cards each. As it turns out, the top half contains 3 aces and the bottom half just one ace. The top half is shuffled, then cut into two halves of 13 cards each. One of these packs of 13 cards is shuffled into the bottom half of 26 cards, and from this pack of 39 cards, 5 cards are dealt. What is the expected number of aces among these 5 cards?

12. Conditional expectations in Polya's urn scheme. An urn contains 1 black and 2 white balls. One ball is drawn at random and its color noted. The ball is replaced in the urn, together with an additional ball of its color. There are now four balls in the urn. Again, one ball is drawn at random from the urn, then replaced along with an additional ball of its color. The process continues in this way.

a) Let B_n be the number of black balls in the urn just before the nth ball is drawn. (Thus B_1 is 1.) For $n \geq 1$, find $E(B_{n+1}|B_n)$.

b) For $n \geq 1$, find $E(B_n)$. [*Hint*: $E(B_1) = 1$; now use part a) and induction on n.]

c) For $n \geq 1$, what is the expected proportion of black balls in the urn just before the nth ball is drawn?

13. Conditioning on the number of successes in Bernoulli trials. Let $S_n = X_1 + \cdots + X_n$ be the number of successes in of n independent Bernoulli(p) trials $X_1, X_2, \ldots, X_n$.

a) For $1 \leq m \leq n$, show that the conditional distribution of S_m, the number of successes in the first m trials, given $S_n = k$, is identical to the distribution of the number of good elements in a random sample of size m without replacement from a population of k good and $n - k$ bad elements.

b) Use the result of a) to rederive the result of Example 4 that $E(S_m \,|\, S_n = k) = mk/n$.

c) Find $Var(S_m \,|\, S_n = k)$.

14. Sufficiency of the number of successes in Bernoulli trials. Let $S_n = X_1 + \cdots + X_n$ be the number of successes in n independent Bernoulli (p) trials $X_1, X_2, \ldots, X_n$. As a continuation of Exercise 13, show that conditionally given $S_n = k$, the sequence of zeros and ones $X_1, \ldots, X_n$ is distributed like an exhaustive sample without replacement from a population of k ones and $n - k$ zeros. [Note that this conditional distribution does not depend on p. In the language of statistics, when p is an unknown parameter S_n is called a *sufficient statistic* for p. If you want to estimate an unknown p given observed values of $X_1, \ldots, X_n$, and are committed to the assumption of Bernoulli (p) trials, it makes no sense to use any aspect of the data besides S_n in the estimation problem, because given $S_n = k$, the parameter p does not affect the distribution of the data at all. One natural estimate of p given the data is S_n/n, the observed proportion of successes. But other functions of S_n may be considered. See Exercise 6.3.15.]

15. Let Π be a random proportion between 0 and 1, for example, the proportion of black balls in an urn picked at random from some population of urns. Let S be the number of successes in n Bernoulli trials, which given $\Pi = p$ are independent with probability p, for example, the number of black balls in n draws at random, with replacement from the urn picked at random.

a) Find a formula for $E(S)$ in terms of n and $E(\Pi)$.

b) Find a formula for $Var(S)$ in terms of n, $E(\Pi)$, and $Var(\Pi)$.

c) For given n and $E(\Pi) = p$, say, which distribution of Π makes $Var(S)$ as large as possible? Which as small as possible? Prove your answers using your answer to b).

16. Expectation of a product by conditioning. Let X and Y be random variables, and let h be a function of X. Show that

$$E[h(X)Y] = E[h(X)E(Y|X)]$$

[*Hint*: Look at $E(h(X)Y|X = x)$.] *Remark:* This identity, for indicator functions $h(x)$, is used in more advanced treatments of probability to define conditional expectations given a continuous random variable X.

17. Prediction by functions. Suppose you want to predict the value of a random variable Y. Instead of just trying to predict the value of Y by a constant, as was done in Section 3.2, suppose that some additional information pertinent to the prediction of Y is available. For instance, you might know the value of some other random variable X, whose joint distribution with Y is assumed known. The problem here is to predict the value of Y by a function of X, call it $g(X)$. Once the value x of X is known, the value $g(x)$ of $g(X)$ can be calculated and used to predict the unknown value of Y.
One measure of the goodness of the predictor $g(X)$ is its *mean square error* (MSE)

$$MSE(g(X)) = E[(Y - g(X))^2]$$

It is a measure of, on average, how far off the prediction is. Show that $g(X) = E(Y|X)$ minimizes the MSE. [*Hint*: Condition on the value of X

$$E[(Y - g(X))^2] = \sum_x E[(Y - g(X))^2|X = x]P(X = x)$$

and minimize each term in the sum separately.]

18. Conditional variance. Define $Var(Y|X)$, the *conditional variance of Y given X*, to be the random variable whose value, if $(X = x)$, is the variance of the conditional distribution of Y given $X = x$. So $Var(Y|X)$ is a function of X, namely $h(X)$, where $h(x) = E(Y^2|X = x) - [E(Y|X = x)]^2$. Show that

$$Var(Y) = E[Var(Y|X)] + Var[E(Y|X)]$$

In words, the variance is the expectation of the conditional variance plus the variance of the conditional expectation.

6.3 Conditioning: Density Case

This section treats conditional probabilities given the value of a random variable X with a continuous distribution. In the discrete case, the conditional probability of an event A, given that X has value x, is defined by

$$P(A \mid X = x) = \frac{P(A, X = x)}{P(X = x)}$$

whenever $P(X = x) > 0$. In the continuous case $P(X = x) = 0$ for every x, so the above formula gives the undefined expression $0/0$. This must be replaced, as in the usual calculus definition of a derivative dy/dx, by the following:

Infinitesimal Conditioning Formula

$$P(A \mid X = x) = \frac{P(A, X \in dx)}{P(X \in dx)}$$

Intuitively, $P(A \mid X = x)$ should be understood as $P(A \mid X \in dx)$, the chance of A given that X falls in a very small interval near x. It is assumed here that in the limit of small intervals this chance does not depend on what interval is chosen near x. So, like a derivative dy/dx, $P(A \mid X \in dx)$ is a function of x, hence the notation $P(A \mid X = x)$. In terms of limits,

$$P(A \mid X = x) = \lim_{\Delta x \to 0} P(A \mid X \in \Delta x) = \lim_{\Delta x \to 0} \frac{P(A, X \in \Delta x)}{P(X \in \Delta x)}$$

where Δx stands for an interval of length Δx containing the point x. It is assumed here that the limit exists, except perhaps for a finite number of exceptional points x such as endpoints of an interval defining the range of X, or places where the density of X has a discontinuity. See the book *Probability and Measure* by P. Billingsley for a rigorous treatment of conditioning on a continuously distributed variable.

Most often, the event A of interest is determined by some random variable Y, for instance, $A = (Y > 3)$. If (X, Y) has a joint density $f(x, y)$, then $P(A \mid X = x)$ can be found by integration of the conditional density of Y given $X = x$, defined as follows:

Conditional Density of Y given X = x

For random variables X and Y with joint density $f(x,y)$, for each x such that the marginal density $f_X(x) > 0$, the *conditional density of Y given $X = x$* is the probability density function with dummy variable y defined by

$$f_Y(y \mid X = x) = f(x,y)/f_X(x)$$

Intuitively, the formula for $f_Y(y \mid X = x)$ is justified by the following calculation of the chance of $(Y \in dy)$ given $X = x$:

$$\begin{aligned} P(Y \in dy \mid X = x) &= P(Y \in dy \mid X \in dx) \\ &= \frac{P(X \in dx, Y \in dy)}{P(X \in dx)} \\ &= \frac{f(x,y)\,dx\,dy}{f_X(x)\,dx} \\ &= f_Y(y \mid X = x)dy \end{aligned}$$

The formula $\int f(x,y)dy = f_X(x)$, the marginal density of X, implies that

$$\int f_Y(y \mid X = x)dy = 1$$

So for each fixed x with $f_X(x) > 0$, the formula for $f_Y(y \mid X = x)$ gives a probability density in y. This conditional density given x defines a probability distribution parameterized by x, called the *conditional distribution of Y given $X = x$*. In examples, this will often be a familiar distribution, for example, a uniform or a normal distribution, with parameters depending on x.

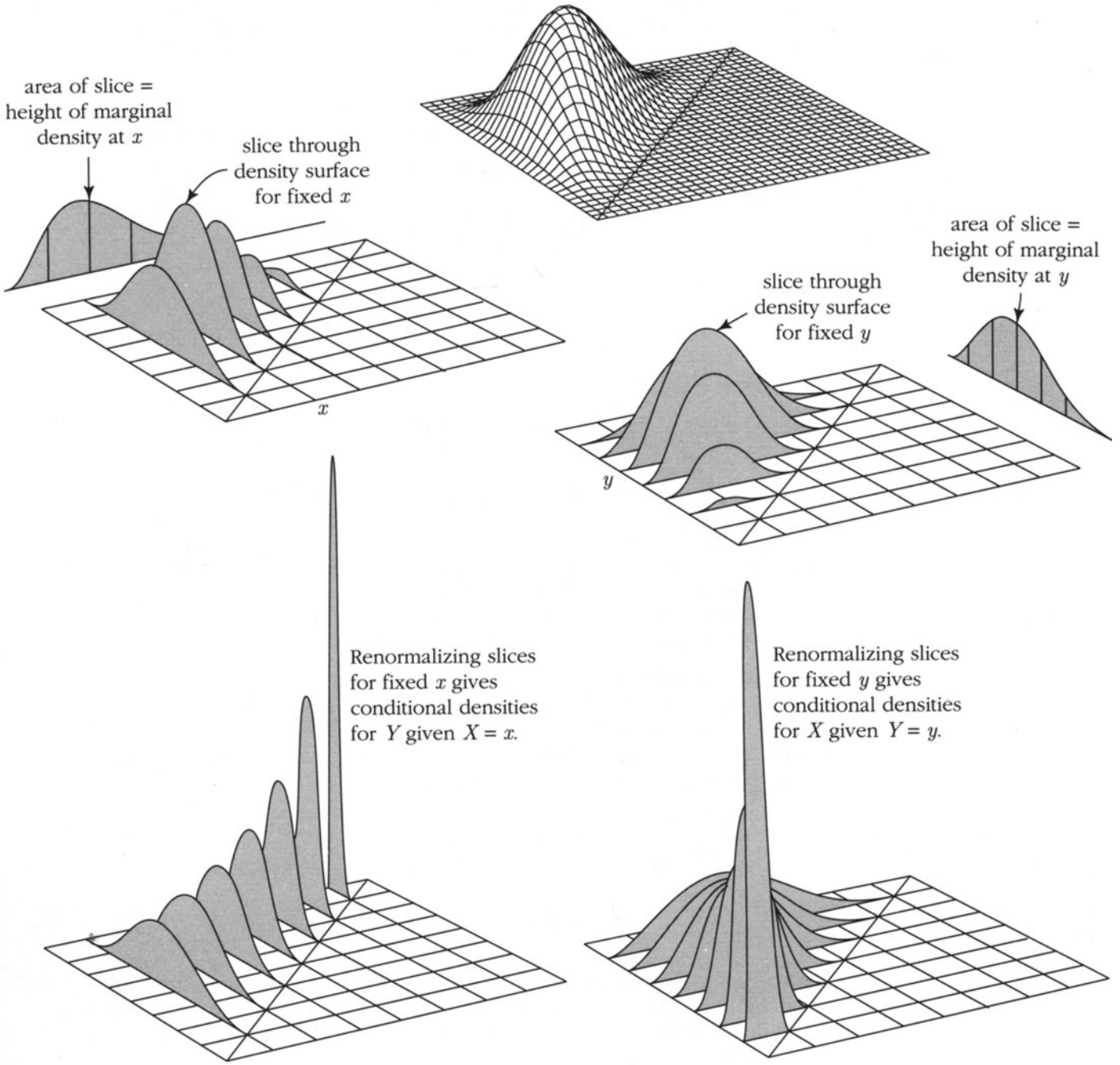
The conditional density of Y given $X = x$ can be understood geometrically by taking a vertical slice through the joint density surface at x, and renormalizing the resulting function of y by its total integral, which is $f_X(x)$. Conditional probabilities given $X = x$ of events determined by X and Y can be calculated by integrating with respect to this conditional density. For example

$$P(Y > b \mid X = x) = \int_b^\infty f_Y(y \mid X = x)dy$$

$$P(Y > 2X \mid X = x) = \int_{2x}^\infty f_Y(y \mid X = x)dy$$

Such expressions are obtained formally from their discrete analogs by replacing a sum by an integral, and replacing the probability of an individual point by the value of a density times an infinitesimal length. See the display at the end of this section for details of this analogy.

FIGURE 1. Joint, marginal, and conditional densities.



Key to Figure 1

Top: Joint density surface. This is a perspective projection of the surface

$$z = f(x, y)$$

defined by a particular joint density function $f(x, y)$.

Middle left: Slices for some values of X and the marginal density of X. Here are seven slices, or cross sections through the density surface for given values X ranging from 1/8 to 7/8. (The last two are so low that they are invisible.) The probability that X falls in a short interval of length Δ near x is the volume of such a slice of thickness Δ, which for small enough Δ is essentially Δ times the area of the slice at x. This area equals

$$\int f(x, y)dy = f_X(x)$$

the height of the *marginal density of X at x*, graphed at back. This marginal density shows how probability is distributed between slices according to the distribution of X. The heights of the vertical segments shown in the graph of the marginal density are proportional to the areas of corresponding slices.

Middle right: Slices for some values Y and the marginal density of Y. Here are perpendicular slices through the density surface for given values of Y. The area of the slice at y equals

$$\int f(x, y)\, dx = f_Y(y),$$

the height of the *marginal density of Y at y*, shown at right.

Bottom left: Conditional density of Y for some given values of X. Rescaling each section of the diagram above by its total area, the marginal density of X at x, gives the *conditional density of Y given $X = x$*, shown here using the same vertical scale as for the marginal densities in the middle diagrams. Given $X = x$, Y is distributed with density proportional to the section of the density surface $f(x, y)$ through x. Dividing by the total area of the section through x gives the conditional density of Y given $X = x$. Note how the shape of the two invisible sections in the middle left diagram can now be seen, due to the normalization of each section by its total area. The marginal density of Y (see middle right) is the average of all the conditional densities of Y given $X = x$ weighted according to the marginal distribution of X (middle left).

Bottom right: Conditional density of X for some given values of Y. These are interpreted just as above, with the roles of X and Y switched.

Example 1. Uniform on a triangle.

Problem. Suppose that a point (X, Y) is chosen uniformly at random from the triangle $\{(x, y) : x \geq 0, y \geq 0, x + y \leq 2\}$. Find $P(Y > 1 | X = x)$.

To illustrate the basic concepts, three slightly different solutions will be presented.

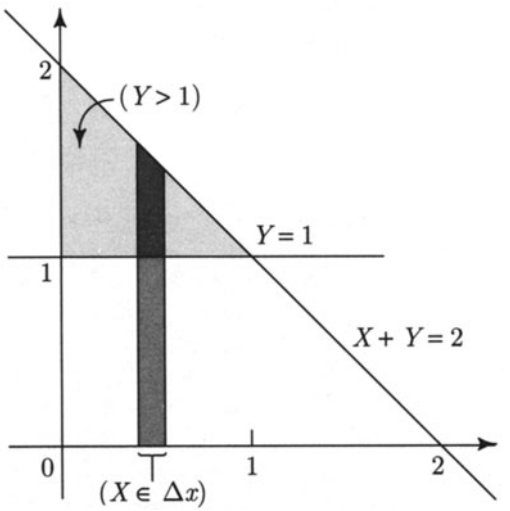
Solution 1. *Informal approach.* Intuitively, it seems obvious that given $X = x$, the random point (X, Y) should be regarded as uniformly distributed on the vertical line segment $\{(x, y) : y \geq 0, x + y \leq 2\}$ with length $2 - x$. This is the conditional distribution of (X, Y) given $X = x$. If x is between 0 and 1, the portion of this segment above $y = 1$ has length $(2 - x) - 1 = 1 - x$. Otherwise, no portion of the segment is above $y = 1$. So the answer is

$$P(Y > 1 | X = x) = \begin{cases} (1 - x)/(2 - x) & 0 \leq x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Solution 2. *Definition of conditional probability.* To see that Solution 1 agrees with the formal definition

$$P(Y > 1 | X = x) = \lim_{\Delta x \to 0} P(Y > 1 | X \in \Delta x)$$

look at the following diagram which shows the events $(Y > 1)$ and $(X \in \Delta x) = (x \leq X \leq x + \Delta x)$:



Since the triangle has area 2, the probability of an event is half its area. So, for $0 \leq x < 1, x + \Delta x \leq 1$, there are the exact formulae

$$P(X \in \Delta x) = \frac{1}{2}\Delta x(2 - x - \frac{1}{2}\Delta x)$$

$$P(Y > 1, X \in \Delta x) = \frac{1}{2}\Delta x(1 - x - \frac{1}{2}\Delta x)$$

Therefore, for $0 \leq x < 1$,

$$\begin{aligned} P(Y > 1 \mid X \in \Delta x) &= \frac{P(Y > 1, X \in \Delta x)}{P(X \in \Delta x)} \\ &= \frac{1 - x - \frac{1}{2}\Delta x}{2 - x - \frac{1}{2}\Delta x} \\ &\to \frac{1 - x}{2 - x} \quad \text{as} \quad \Delta x \to 0 \end{aligned}$$

This verifies the formula of Solution 1 for $0 \leq x < 1$. The formula for $x \geq 1$ is obvious because the event $(Y > 1, X \in \Delta x)$ is empty if $x \geq 1$.

Solution 3. *Calculation with densities.* Let us recalculate $P(Y > 1 \mid X = x)$ using the conditional density $f_Y(y \mid X = x)$. The uniform distribution on the triangle makes the joint density

$$f(x, y) = \begin{cases} 1/2 & x \geq 0, y \geq 0, x + y \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

So for $0 \leq x \leq 2$,

$$f_X(x) = \int_0^\infty f(x, y)dy = \int_0^{2-x} \frac{1}{2}dy = \frac{1}{2}(2 - x)$$

and

$$f_Y(y \mid X = x) = \begin{cases} \dfrac{f(x, y)}{f_X(x)} = \dfrac{1}{2 - x} & 0 \leq y \leq 2 - x \\ 0 & \text{otherwise} \end{cases}$$

That is, given $X = x$ for $0 \leq x \leq 2$, Y has uniform $(0, 2 - x)$ distribution, as is to be expected intuitively. So

$$P(Y > 1 \mid X = x) = \begin{cases} \displaystyle\int_1^{2-x} \frac{dy}{2 - x} = \frac{1 - x}{2 - x} & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

as before.

Discussion. The point of the first solution is that conditional distributions are often intuitively obvious, and once identified they can be used to find conditional probabilities very quickly. The second solution shows how this kind of calculation is justified by the formal definition. This method is not recommended for routine calculations. The third solution is essentially a more detailed version of the first. While rather pedantic

in the present problem, this kind of calculation is essential in more difficult problems where you cannot guess the answer by intuitive reasoning.

Rules for conditional densities. These are analogs of corresponding rules in the discrete case. Note that every concept defined by the distribution of a real-valued random variable Y, in particular, the notions of density function, distribution function, expectation, variance, moments, and so on, can be considered for conditional distributions, just as well as for unconditional ones. There is just an extra parameter, x, the given value of X.

When the density of X is known, and a conditional density for Y given $X = x$ is specified for each x in the range of X, the joint density of X and Y is calculated by the following rearrangement of the formula $f_Y(y \mid X = x) = f(x,y)/f_X(x)$.

Multiplication Rule for Densities

$$f(x,y) = f_X(x) f_Y(y \mid X = x)$$

Example 2. Gamma and uniform.

Suppose X has gamma $(2,\lambda)$ distribution, and that given $X = x$, Y has uniform $(0,x)$ distribution.

Problem 1. Find the joint density of X and Y.

Solution. By the definition of the gamma distribution

$$f_X(x) = \begin{cases} \lambda^2 x e^{-\lambda x} & x > 0 \\ 0 & x \le 0 \end{cases}$$

and from the uniform $(0,x)$ distribution of Y given $X = x$

$$f_Y(y \mid X = x) = \begin{cases} 1/x & 0 < y < x \\ 0 & \text{otherwise} \end{cases}$$

So by the multiplication rule for densities

$$f(x,y) = f_X(x) f_Y(y \mid X = x) = \begin{cases} \lambda^2 e^{-\lambda x} & 0 < y < x \\ 0 & \text{otherwise} \end{cases}$$

Problem 2. Find the marginal density of Y.

Solution. Integrating out x in the joint density gives the marginal density of Y: for $y > 0$

$$f_Y(y) = \int_0^\infty f(x,y)\,dx = \int_y^\infty \lambda^2 e^{-\lambda x}\,dx = \lambda e^{-\lambda y}$$

The density is of course 0 for $y \leq 0$. That is to say, Y has exponential (λ) distribution.

Problem 3. Show that X and Y have the same joint distribution as T_2 and T_1, where T_1 is the first arrival time and T_2 is the second arrival time in a Poisson arrival process with rate λ.

Solution. That X has the same distribution as T_2, and that Y has the same distribution as T_1, follows from the above calculation and the result of Section 4.2 that the ith arrival time in a Poisson process with rate λ has gamma(i, λ) distribution. That the *joint* distribution of X and Y is the same as the joint distribution of T_2 and T_1 requires a little more calculation, because a joint distribution is not determined by its marginals. The simplest way to verify this is to observe that for $0 < y < x$

$$P(T_1 \in dy, T_2 \in dx)$$

is the probability of no arrivals in the time interval $[0, y]$ of length y, one arrival in time dy, no arrivals in the time interval $[y + dy, x]$ of length $x - y - dy \approx x - y$, and finally one arrival in dx. By independence and Poisson distribution of counts in disjoint intervals, and neglecting a term of order $(dy)^2$, this event has probability

$$e^{-\lambda y}\, \lambda\, dy\, e^{-\lambda(x-y)}\, \lambda\, dx = \lambda^2 e^{-\lambda x} dy\, dx$$

Dividing the last expression by $dy\, dx$ shows that the joint density of (T_2, T_1) at (x, y) with $0 < y < x$ is identical to the joint density found in Problem 1. Since obviously $P(T_1 < T_2) = 1$, the joint density of (T_2, T_1) can be taken to be zero except if $0 < y < x$. Thus (T_2, T_1) has the same joint density function as (X, Y), hence the same joint distribution.

Problem 4. For T_1 and T_2 the first two arrival times in a Poisson process with rate λ, find the conditional distribution of T_1 given $T_2 = x$.

Solution. Since according to the solution of the previous problem, T_2 and T_1 have the same joint density as X and Y, found in Problem 1, the conditional distribution of T_1 given $T_2 = x$ is identical to the conditional distribution of Y given $X = x$, which was given at the start, that is to say, uniform on $(0, x)$.

Averaging Conditional Probabilities

For a random variable X with density f_X, the rule of average conditional probabilities becomes the following:

Integral Conditioning Formula

$$P(A) = \int P(A \,|\, X = x) f_X(x)\, dx$$

The integral breaks up the probability of A according to the values of X:

$$P(A \mid X = x) f_X(x)\, dx = P(A \mid X \in dx) P(X \in dx) = P(A, X \in dx)$$

Just as in the discrete case, $P(A \mid X = x)$ is often specified in advance by the formulation of a problem. Then $P(A)$ can be calculated by the integral conditioning formula, assuming also that the distribution of X is known. Bayes' rule then gives the conditional density of X given that A has occurred:

$$P(X \in dx \mid A) = \frac{P(X \in dx) P(A \mid X = x)}{P(A)} = \frac{f_X(x) P(A \mid X = x)}{P(A)}\, dx$$

The following example shows how the integral conditioning formula arises naturally by taking limits of discrete problems. In this example, as is often the case, the limits defined by integrals are much easier to work with than the discrete sums. The example makes precise the idea of independent trials with probabilty p of success in a setting where it makes clear sense to think of p as picked at random from some distribution before the trials are performed. In the first problem p is picked from a discrete uniform distribution on $N+1$ evenly spaced points in $[0,1]$. Passing to the limit as $N \to \infty$ leads to p that is uniformly distributed on $[0,1]$. Bayesian statisticians view this as a model for independent trials with unknown probability of success.

Example 3. Discrete uniform–binomial.

Suppose there are $N+1$ boxes labeled by $b = 0, 1, 2, \ldots, N$. Box b contains b black and $N - b$ white balls. A box is picked uniformly at random, and then n balls are drawn at random with replacement from whatever box is picked (the same box for each of the n draws). Let S_n denote the total number of black balls that appear among the n balls drawn.

Problem 1. Find the distribution of S_n.

Solution. Let Π denote the proportion of black balls in the box picked. Let G_N denote the grid of $N+1$ possible values p of Π:

$$G_N = \{0, \frac{1}{N}, \frac{2}{N}, \ldots, \frac{N-1}{N}, 1\}$$

For each $p \in G_N$ the binomial formula for n independent trials with probability p of success on each trial gives

$$P(S_n = k \mid \Pi = p) = \binom{n}{k} p^k (1-p)^{n-k}$$

Averaging with respect to the uniform distribution of Π over the $N+1$ values in G_N, and substituting $p = b/N$, gives the unconditional distribution of S_n:

$$P(S_n = k) = \sum_{p \in G_N} \binom{n}{k} p^k (1-p)^{n-k} \frac{1}{N+1} \tag{1}$$

$$= \binom{n}{k} \frac{1}{(N+1)N^n} \sum_{b=0}^{N} b^k (N-b)^{n-k}$$

It is hard to simplify this expression further. But the expression is easily evaluated for small values of n and N. To illustrate, for $N = n = 2$ the result is shown in the next table. The limiting behavior for large N is the subject of the next problem.

Distribution of S_2 for $N = 2$

k	0	1	2
$P(S_2 = k)$	$\frac{5}{12}$	$\frac{2}{12}$	$\frac{5}{12}$

Problem 2. For a fixed value of n, find the limiting distribution of S_n, the number of black balls that appear in n draws, as the number of boxes N tends to ∞.

Solution. Expression (1) for $P(S_n = k)$ is $\binom{n}{k}$ times a discrete approximation to the beta integral

$$B(k+1, n-k+1) = \int_0^1 p^k (1-p)^{n-k} dp$$

The approximation in (1) is obtained by taking the average value of the function $p^k(1-p)^{n-k}$ at $N+1$ evenly spaced points p, beween 0 and 1. In the limit as $N \to \infty$, the discrete average converges to the continuous integral. Using the expression for the beta integral in terms of the gamma function, and $\Gamma(m+1) = m!$ for integers m, gives

$$B(k+1, n-k+1) = \frac{\Gamma(k+1)\Gamma(n-k+1)}{\Gamma(k+1+n-k+1)} = \binom{n}{k}^{-1} \frac{1}{n+1} \tag{2}$$

The conclusion is that as $N \to \infty$

$$P(S_n = k) \to \binom{n}{k} \binom{n}{k}^{-1} \frac{1}{n+1} = \frac{1}{n+1}$$

for every $0 \le k \le n$. That is, the limiting distribution of S_n as $N \to \infty$ is uniform on $\{0, 1, \ldots, n\}$.

Example 4. Continuous uniform–binomial.

Suppose that Π is picked uniformly at random from $(0,1)$. Given that $\Pi = p$, let S_n be the number of successes in n independent trials with probability p of success on each trial.

Problem 1. Find the distribution of S_n.

Solution. By the limiting result obtained in the previous example as $N \to \infty$, the answer must be uniform on $\{0, 1, \ldots n\}$. This can be derived directly in the continuous model using the integral conditioning formula. Since the density of Π is $f_\Pi(p) = 1$ for $0 < p < 1$, and 0 otherwise,

$$\begin{aligned} P(S_n = k) &= \int P(S_n = k \,|\, \Pi = p) f_\Pi(p) dp \qquad (3) \\ &= \int_0^1 \binom{n}{k} p^k (1-p)^{n-k} dp \\ &= \frac{1}{n+1} \end{aligned}$$

by evaluation of the beta integral as in the previous problem.

Discussion. Note the close parallel between the expression (3) for $P(S_n = k)$ obtained by the integral conditioning formula for Π with uniform distribution on $(0,1)$, and the corresponding expression (1) for $P(S_n = k)$ in the previous example for Π with uniform distribution on the set of $N+1$ values in G_N. All that happens is that the sum is replaced by an integral, and $1/(N+1)$, which is both the probability of each point in G_N and the difference between adjacent points in G_N, is replaced by the calculus differential dp representing the probability that the uniform variable falls in an infinitesimal length dp near p.

Problem 2. Find the conditional distribution of Π given that $S_n = k$.

Solution. Using Bayes' rule, for $0 < p < 1$,

$$\begin{aligned} P(\Pi \in dp \,|\, S_n = k) &= \frac{P(\Pi \in dp) P(S_n = k \,|\, \Pi = p)}{P(S_n = k)} \\ &= (n+1) \binom{n}{k} p^k (1-p)^{n-k} dp \end{aligned}$$

This is the density at p of the beta distribution with parameters $k+1$ and $n-k+1$, times dp. Conclusion: the conditional distribution of Π given $S_n = k$ is beta$(k+1,\ n-k+1)$.

Problem 3. In the above setup, given that n trials have produced k successes, what is the probability that the next trial is a success?

Solution. Given $\Pi = p$ and $S_n = k$, the next trial is a success with probability p, by the assumption of independent trials with constant probability p of success given $\Pi = p$. Given just $S_n = k$, the value of Π is unknown. Rather, Π is a random variable with beta $(k+1,\ n-k+1)$ distribution. By the integral conditioning formula, the required

probability is the conditional expectation of Π given $S_n = k$, which is $(k+1)/(n+2)$, by the formula $a/(a+b)$ for the mean of the beta (a,b) distribution. In detail:

$$\begin{aligned} &P(\text{next trial a success} \mid S_n = k) \\ &= \int_0^1 P(\text{next trial a success} \mid S_n = k, \Pi = p) f_\Pi(p \mid S_n = k) dp \\ &= \int_0^1 p f_\Pi(p \mid S_n = k) dp = E(\Pi \mid S_n = k) = \frac{k+1}{n+2} \end{aligned}$$

Discussion. In particular, for $k = n$, given n successes in a row, the chance of one more success is $(n+1)/(n+2)$. This formula, for the probability of one more success given a run of n successes in independent trials with unknown success probability assumed uniformly distributed on $(0,1)$, is known as *Laplace's law of succession.* Laplace illustrated his formula by calculating the probability that the sun will rise tomorrow, given that it has risen daily for 5000 years, or $n = 1,826,213$ days. But this kind of application is of doubtful value. Both the assumption of independent trials with unknown p and the uniform prior distribution of p make little sense in this context.

Example 5. Simulation of uniform–binomial.

Suppose you have available a random number generator which you are willing to believe generates independent uniform $(0,1)$ variables $U_0, U_1, \ldots$.

Problem 1. How could you simulate a pair of values from the joint distribution of Π and S_n considered above, with Π uniform on $(0,1)$, and S_n binomial(n,p) given $\Pi = p$?

Solution. Set

$$\Pi = U_0, \text{ and } S_n = \sum_{i=1}^{n} I(U_i < U_0)$$

where $I(U_i < U_0)$ is an indicator variable that is 1 if $(U_i < U_0)$ and 0 otherwise. If $\Pi = p$, then $S_n = \sum_{i=1}^n I(U_i < p)$ is the sum of n independent indicator variables, each of which is 1 with probability p and 0 with probability $1-p$, exactly as required.

Problem 2. Use this construction to calculate $P(S_n = k)$ without integration.

Solution. By construction of S_n from $U_0, U_1, \ldots, U_n$

$(S_n = 0)$ if and only if U_0 is the smallest of the $U_0, U_1, \ldots, U_n$

$(S_n = 1)$ if and only if U_0 is the second smallest of the $U_0, U_1, \ldots, U_n$

$\ldots \qquad \ldots \qquad \ldots$

$(S_n = n)$ if and only if U_0 is the largest of the $U_0, U_1, \ldots, U_n$

Since all $(n+1)!$ possible orderings of the $U_0, U_1, \ldots, U_n$ are equally likely, each of these events has the same probability $1/(n+1)$.

Remark. This calculation is closely related to the distribution of order statistics treated in Section 4.6. For $j = 1, \ldots, n+1$, let $U_{(j)}$ denote the jth smallest of the $n+1$ variables $U_0, \ldots, U_n$. Then the event $S_n = j - 1$, that there are exactly $j - 1$ values U_i less than U_0, is identical to the event $U_{(j)} = U_0$, that the jth smallest of the U_i equals U_0. The solution of Problem 2 in Example 4 now translates into the following: the conditional distribution of U_0, or of $U_{(j)}$, given that $U_{(j)} = U_0$, is beta $(j, n - j + 2)$. By symmetry, the same is true for U_k instead of U_0 for any $1 \le k \le n$. Consequently, the distribution of $U_{(j)}$, the jth smallest of $n+1$ independent uniform $(0, 1)$ variables, is beta $(j, n - j + 2)$, independently of K, where K is the random index k such that $U_k = U_{(j)}$. This agrees with the result of Section 4.6, with the present $n + 1$ and j instead of n and k in that section.

Independence

In the continuous case, just as in the discrete case, it can be shown that each of the following conditions is equivalent to independence of random variables X and Y:

- the conditional distribution of Y given $X = x$ does not depend on x;
- the conditional distribution of X given $Y = y$ does not depend on y.

By integration with respect to the distribution of X, the common conditional distribution of Y given $X = x$ then equals the unconditional distribution of Y. That is to say, for all subsets B in the range of Y

$$P(Y \in B \mid X = x) = P(Y \in B)$$

Similarly for all subsets A in the range of X

$$P(X \in A \mid Y = y) = P(X \in A)$$

These are variations of the basic definition of independence of X and Y, which is

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

for all subsets A and B in the ranges of X and Y respectively. When X and Y have densities, X and Y are independent if and only if $f_Y(y \mid X = x) = f_Y(y)$ for all x and y, and again if and only if $f_X(x \mid Y = y) = f_X(x)$ for all x and y. So the general multiplication rule for densities reduces in this case to the formula

$$f(x, y) = f_X(x) f_Y(y)$$

for independent variables X and Y. This formula was applied in Section 5.2.

Conditional Expectations

The conditional expectation of Y given $X = x$, denoted $E(Y \mid X = x)$, is defined as the expectation of Y relative to the conditional distribution of Y given $X = x$. More generally, for a function g, assuming that Y has a conditional density $f_Y(y \mid X = x)$,

$$E[g(Y) \mid X = x] = \int g(y) f_Y(y \mid X = x) dy$$

Taking $g(y) = y$ gives $E(Y \mid X = x)$. And integrating the conditional expectation with respect to the distribution of X gives the unconditional expectation

$$E[g(Y)] = \int E[g(Y) \mid X = x] f_X(x)\, dx$$

These formulae are extensions to general functions g of the basic conditional probability formulae, which are the special cases when g is an indicator. As a general rule, all the basic properties of conditional expectations, considered in the discrete case in Section 6.2, remain valid in the density case.

Example 6. Uniform distribution on a triangle.

Problem. Suppose, as in Example 1, that (X, Y) is chosen uniformly at random from the triangle $\{(x, y) : x \geq 0, y \geq 0, x + y \leq 2\}$. Find $E(Y \mid X)$ and $E(X \mid Y)$.

Solution. As argued before, given $X = x$, for $0 < x < 2$, Y has uniform distribution on $(0, 2 - x)$. Since the mean of this conditional distribution is $(2 - x)/2$,

$$E(Y \mid X = x) = (2 - x)/2$$

In short notation

$$E(Y \mid X) = (2 - X)/2$$

Similarly, because joint density of X and Y is symmetric in x and y,

$$E(X \mid Y) = (2 - Y)/2$$

Conditioning Formulae: Discrete Case

Multiplication rule: The joint probability is the product of the marginal and the conditional

$$P(X = x, Y = y) = P(X = x)P(Y = y \mid X = x)$$

Division rule: The conditional probability of $Y = y$ given $X = x$ is

$$P(Y = y \mid X = x) = \frac{P(X = x, Y = y)}{P(X = x)}$$

Bayes' rule:

$$P(X = x \mid Y = y) = \frac{P(Y = y \mid X = x)P(X = x)}{P(Y = y)}$$

Conditional distribution of Y given $X = x$: Sum the conditional probabilities

$$P(Y \in B \mid X = x) = \sum_{y \in B} P(Y = y \mid X = x)$$

Conditional expectation of $g(Y)$ given $X = x$: Sum g against the conditional probabilities

$$E(g(Y) \mid X = x) = \sum_{\text{all } y} g(y)P(Y = y \mid X = x)$$

Average conditional probability:

$$P(B) = \sum_{\text{all } x} P(B \mid X = x)P(X = x)$$

$$P(Y = y) = \sum_{\text{all } x} P(Y = y \mid X = x)P(X = x)$$

Average conditional expectation:

$$E(Y) = \sum_{\text{all } x} E(Y \mid X = x)P(X = x)$$

Conditioning Formulae: Density Case

Multiplication rule: The joint density is the product of the marginal and the conditional

$$f(x,y) = f_X(x) f_Y(y \mid X = x)$$

Division rule: The conditional density of Y at y given $X = x$ is

$$f_Y(y \mid X = x) = \frac{f(x,y)}{f_X(x)}$$

Bayes' rule:

$$f_X(x \mid Y = y) = \frac{f_Y(y \mid X = x) f_X(x)}{f_Y(y)}$$

Conditional distribution of Y given $X = x$: Integrate the conditional density

$$P(Y \in B \mid X = x) = \int_B f_Y(y \mid X = x) dy$$

Conditional expectation of $g(Y)$ given $X = x$: Integrate g against the conditional density:

$$E(g(Y) \mid X = x) = \int g(y) f_Y(y \mid X = x) dy$$

Average conditional probability:

$$P(B) = \int P(B \mid X = x) f_X(x)\, dx$$

$$f_Y(y) = \int f_Y(y \mid X = x) f_X(x)\, dx$$

Average conditional expectation:

$$E(Y) = \int E(Y \mid X = x) f_X(x)\, dx$$

Exercises 6.3

1. Suppose X has uniform $(0,1)$ distribution and $P(A|X=x)=x^2$. What is $P(A)$?

2. Let X and Y have the following joint density:

$$f(x,y)=\begin{cases} 2x+2y-4xy & \text{for } 0\le x\le 1 \text{ and } 0\le y\le 1 \\ 0 & \text{otherwise} \end{cases}$$

a) Find the marginal densities of X and Y.

b) Find $f_Y(y|X=\frac{1}{4})$. c) Find $E(Y\,|\,X=\frac{1}{4})$.

3. Let (X,Y) be as in Example 1. Find a formula for $P(Y\le y|X=x)$.

4. Suppose X,Y are random variables with joint density

$$f(x,y)=\begin{cases} \lambda^3 x e^{-\lambda y} & \text{for } 0<x<y \\ 0 & \text{otherwise} \end{cases}$$

a) Find the density of Y. What is $E(Y)$? b) Compute $E(X|Y=1)$.

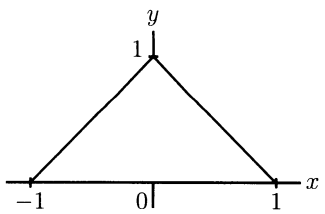
5. Suppose (X,Y) has uniform distribution on the triangle shown in the diagram. For x between -1 and 1, find:

a) $P(Y\ge\frac{1}{2}|X=x)$;

b) $P(Y<\frac{1}{2}|X=x)$;

c) $E(Y|X=x)$;

d) $Var(Y|X=x)$.



6. Suppose X,Y are random variables with joint density

$$f_{X,Y}(x,y)=\frac{1}{2\pi\sqrt{x(y-x)}}\,e^{-y/2} \qquad (0<x<y)$$

a) Find the distribution of Y. [*Hint*: For integration use the substitution $x=y\,s$.]

b) Compute $E(X|Y=1)$.

7. Suppose that Y and Z are random variables with the following joint density:

$$f(y,z)=\begin{cases} k(z-y) & \text{for } 0\le y\le z\le 1 \\ 0 & \text{otherwise} \end{cases}$$

for some constant k. Find:

a) the marginal distribution of Y; b) $P\left(Z<\frac{2}{3}|Y=\frac{1}{2}\right)$.

8. The random variable X has a uniform distribution on $(0,1)$. Given that $X=x$, the random variable Y is binomial with parameters $n=5$ and $p=x$.

a) Find $E(Y)$ and $E(Y^2)$. b) Find $P(Y=y \text{ and } x<X<x+dx)$.

c) Find the density of X given $Y=y$. Do you recognize it? If yes, as what?

9. Let A and B be events and let Y be a random variable uniformly distributed on $(0,1)$. Suppose that, conditional on $Y = p$, A and B are independent, each with probability p. Find:

a) the conditional probability of A given that B occurs;

b) the conditional density of Y given that A occurs and B does not.

10. Conditioning a Poisson process on the number of arrivals in a fixed time. Let T_1 and T_5 be the time of the first and fifth arrivals in a Poisson process with rate λ, as in Section 4.2.

a) Find the conditional density of T_1 given that there are 10 arrivals in the time interval $(0,1)$.

b) Find the conditional density of T_5 given that there are 10 arrivals in the time interval $(0,1)$.

c) Recognize the answers to a) and b) as named densities, and find the parameters.

11. Suppose X has uniform distribution on $(-1,1)$ and, given $X = x$, Y is uniformly distributed on $(-\sqrt{1-x^2}, \sqrt{1-x^2})$. Is (X,Y) then uniformly distributed over the unit disk $\{(x,y) : x^2 + y^2 < 1\}$? Explain carefully.

12. Suppose there are ten atoms, each of which decays by emission of an α-particle after an exponentially distributed lifetime with rate 1, independently of the others. Let T_1 be the time of the first α-particle emission, T_2 the time of the second. Find:

a) the distribution of T_1;

b) the conditional distribution of T_2 given T_1;

c) the distribution of T_2.

13. Let X and Y be independent random variables, X with uniform distribution on $(0,3)$, Y with Poisson (λ) distribution. Find:

a) a formula in terms of λ for $P(X < Y)$;

b) the conditional density of X given $X < Y$, and sketch its graph in the cases $\lambda = 1, 2, 3$;

c) $E(X|X < Y)$.

14. Bayesian sufficiency. Let $S_n = X_1 + \cdots + X_n$ be the number of successes in a sequence of n independent Bernoulli (p) trials $X_1, X_2, \ldots, X_n$ with unknown success probability p. Regard p as the value of a random variable Π whose prior distribution has some density $f(p)$ on $(0,1)$. Show that the conditional (posterior) distribution of Π given $X_1 = x_1, \ldots, X_n = x_n$, for any particular sequence of zeros and ones $x_1, \ldots, x_n$ with $x_1 + \cdots + x_n = k$, depends only the observed number of successes k in the n trials, and not on the order in which the k successes and $n-k$ failures appear. Deduce that this conditional distribution is identical to the posterior distribution of Π given $S_n = k$. [This is another expression of the fact that S_n is a sufficient statistic for p. See Exercise 6.2.14.]

15. Beta-binomial. As in Exercise 14 let $S_n = X_1 + \cdots + X_n$ be the number of successes in a sequence of n independent Bernoulli (p) trials $X_1, X_2, \ldots, X_n$, with unknown success probability p, regarded as the value of a random variable Π.

a) Suppose the prior distribution of Π is beta (r, s) for some $r > 0$ and $s > 0$. Show that the posterior distribution of Π given $S_n = k$ is beta $(r + k, s + n - k)$. [*Hint for quick solution:* It is enough to show that the posterior density is *proportional* to the beta $(r + k, s + n - k)$ density. See Chapter 4 Review Exercise 8.]

b) Using the fact that the total integral of the beta $(r + k, s + n - k)$ density is 1, find a formula for the unconditional probability $P(S_n = k)$.

c) Check your result in part b) agrees with the distribution of S_n found in Example 4 in the case $r = s = 1$.

d) For general r and s find the posterior mean $E(\Pi \mid S_n = k)$ and the posterior variance $Var(\Pi \mid S_n = k)$.

e) Suppose n is very large and the observed proportion of successes $\hat{p} = k/n$ is not very close to either 0 or 1. Show that no matter what r and s, provided n is large enough, $E(\Pi \mid S_n = k) \approx \hat{p}$ and $Var(\Pi \mid S_n = k) \approx \hat{p}(1 - \hat{p})/n$.

[It can be shown that the posterior distribution of Π given $S_n = k$ is approximately normal under the assumptions in e). So

for large enough n, the conditional distribution of the unknown value of p, given the observed proportion of successes $\hat{p}$ in n trials, is approximately normal with mean $\hat{p}$ and standard deviation $\sqrt{\hat{p}(1 - \hat{p})}/\sqrt{n}$,

regardless of the prior parameters r and s. The same conclusion holds for any strictly positive and continuous prior density $f(p)$ instead of a beta prior. In the long run, any reasonable prior opinion is overwhelmed by the data. The italicized assertion should be compared to the following paraphrase of the normal approximation to the binomial distribution:

for large enough n, the distribution of proportion of successes $\hat{p}$ in n trials, given the probability p of success on each trial, is approximately normal with mean p and standard deviation $\sqrt{p(1 - p)}/\sqrt{n}$.

While the assertions are very similar, and both true, it is not a trivial matter to pass from one to the other. There is a big conceptual difference between, on the one hand, the distribution of $\hat{p}$ for a fixed and known value of p, which has a clear frequency interpretation in terms of repeated blocks of n trials with the same p, and on the other hand, the posterior distribution of p given $\hat{p}$, which while intuitive from a subjective standpoint, is almost impossible to interpret in terms of long-run frequencies. Long-run frequency of what? The problem is that for large n, in any model of repeated blocks of n trials, the exact value of $\hat{p}$ observed in the first block will typically not be observed even once again until after a very large number of blocks have been examined. The number of blocks required to find the first repeat is of order $\sqrt{n}$ if the same p is used in each block, and order n if p is randomized for each block using the prior distribution: this is because the probability of the most likely values of $\hat{p}$ is of order $1/\sqrt{n}$ in the first case, by the normal approximation to the binomial, and order $1/n$ in the second case, as typified when the prior is uniform on $(0, 1)$ and the distribution of $\hat{p}$ is uniform on the $n + 1$ possible multiples of $1/n$. Either way, it is hard to make a convincing frequency interpretation of the conditional distribution of p given an exact observed value of $\hat{p}$.]

16. Negative binomial distribution for number of accidents. Consider a large population of individuals subject to accidents at various rates. Suppose the empirical distribution of accident rates over the whole population is well approximated by the gamma (r, α) distribution for some $r > 0$ and $\alpha > 0$. Suppose that given an individual has

accident rate λ per day, the number of accidents that individual has in t days has Poisson (λt) distribution. Let Λ be the accident rate and N be the number of accidents in t days for an individual picked at random from this population. So Λ has gamma (r, α) distribution, and given $\Lambda = \lambda$, N has Poisson (λt) distribution.

a) Show by integration that

$$P(N = k) = \frac{\Gamma(r+k)}{\Gamma(r)k!} p^r q^k \quad (k = 0, 1, 2 \ldots) \text{ where } p = \alpha/(t+\alpha), \quad q = t/(t+\alpha)$$

b) Evaluate $\Gamma(r+k)/\Gamma(r)$ as a product of k factors. Deduce that if r is a positive integer, the distribution of N is the same as the distribution of the number of failures before the rth success in Bernoulli (p) trials, as found in Section 3.4.

[In general, the distribution of N defined in a) is called the *negative binomial*(r, p) distribution, now defined for arbitrary $r > 0$ and $0 < p < 1$. The terminology is explained by the following relation between this distribution and the binomial expansion for the negative power $-r$.]

c) Show, either by conditioning on Λ, or from a) and b), that N has generating function

$$E(z^N) = p^r(1 - zq)^{-r} \qquad (|z| < 1)$$

d) Find $E(N)$ and $E(N^2)$ in terms of r and p by conditioning on Λ. Deduce a formula for $Var(N)$. Check for integer r that your results agree with those obtained in Section 3.4.

e) Derive $E(N)$ and $Var(N)$ another way by differentiating the generating function. (Refer to Exercise 3.4.22.)

f) Show that for each integer $k \geq 0$, the conditional density of Λ given $N = k$ is a gamma density, and find its parameters.

17. Sums of independent negative binomial variables. Consider, as in Exercise 16, a large population of individuals subject to accidents at various rates. Suppose now that an individual picked at random from the population is subject to one kind of accident at rate Λ_1 per day, and another kind of accident at rate Λ_2 per day, where Λ_1 and Λ_2 are independent gamma variables with parameters (r_1, α) and (r_2, α) for some $\alpha > 0$. Assume that given $\Lambda_1 = \lambda_1$ and $\Lambda_2 = \lambda_2$ the two types of accidents occur according to independent Poisson processes with rates λ_1 and λ_2. Let N_1 and N_2 be the numbers of accidents of these two kinds the individual has in t days.

a) Describe the joint distribution of N_1 and N_2.

b) What is the distribution of $N_1 + N_2$? [*Hint*: No calculation required. Use results about sums of independent random variables with gamma or Poisson distributions.] Check your conclusion is consistent with the mean and variance formulae of Exercise 16.

c) Suppose $X_i, 1 \leq i \leq k$ are k independent random variables, and that X_i has negative binomial (r_i, p) distribution for some $r_i > 0$, $0 < p < 1$. What is the distribution of $X_1 + \cdots + X_n$? Explain carefully how your conclusion follows from parts a) and b).

d) Derive the result of c) another way using generating functions [see Chapter 3 Review Exercise 34].

6.4 Covariance and Correlation

Covariance is a quantity which appears in calculation of the variance of a sum of possibly dependent random variables. This quantity is useful in variance calculations, but like variance is hard to interpret intuitively. Correlation is a standardized covariance which is easier to interpret. It provides a measure of the degree of linear dependence between two variables. In Section 3.3, the formula

$$Var(X+Y) = Var(X) + Var(Y) \qquad \text{if } X \text{ and } Y \text{ are independent}$$

was derived from the more general formula

$$Var(X+Y) = Var(X) + Var(Y) + 2E\left[(X-\mu_X)(Y-\mu_Y)\right]$$

where $\mu_X = E(X)$ and $\mu_Y = E(Y)$. For independent random variables, the last term vanishes. In general, for two random variables X and Y with finite second moments, there is the following:

Definition of Covariance

The *covariance of X and Y*, denoted $Cov(X,Y)$, is the number

$$Cov(X,Y) = E\left[(X-\mu_X)(Y-\mu_Y)\right]$$

where $\mu_X = E(X), \quad \mu_Y = E(Y)$

Alternative Formula

$$Cov(X,Y) = E(XY) - E(X)E(Y)$$

Variance of a Sum

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$

Proof of alternative formula for covariance. Expand

$$(X-\mu_X)(Y-\mu_Y) = XY - \mu_X Y - X\mu_Y + \mu_X\mu_Y$$

and take expectations.□

Variance. Notice that $Cov(X,X) = Var(X)$, so these formulae for covariance are extensions of old formulae for variance.

Independence. If X and Y are independent then $Cov(X,Y) = 0$.

Warning. $Cov(X,Y) = 0$ *does not* imply X and Y are independent. See Exercises.

Indicators

Let $X = I_A$ be the indicator of event A, and $Y = I_B$ the indicator of another event B. These could be events in any outcome space, where there is given a probability distribution P. In this case

$$XY = I_A I_B = I_{AB}$$

is the indicator of the intersection of the events A and B. Thus

$$E(I_A) = P(A); \qquad E(I_B) = P(B); \qquad E(I_A I_B) = P(AB)$$

$$Cov(I_A, I_B) = P(AB) - P(A)P(B)$$

This covariance is

positive	iff	$P(AB) > P(A)P(B)$, when A and B are called *positively dependent*;
zero	iff	$P(AB) = P(A)P(B)$, when A and B are *independent;*
negative	iff	$P(AB) < P(A)P(B)$, when A and B are called *negatively dependent.*

In the case of positive dependence, learning that B has occurred increases the chance of A:

$$P(A|B) > P(A) \quad \text{and vice versa} \quad P(B|A) > P(B)$$

For negative dependence, learning that B has occurred decreases the chance of A:

$$P(A|B) < P(A) \quad \text{and vice versa} \quad P(B|A) < P(B)$$

These formulations of positive and negative dependence are easily seen to be equivalent to those in the box, by using the formula for $P(A|B)$, and rearranging inequalities. The most extreme case of positive dependence is if A is a subset of B, with $0 < P(A) \leq P(B) < 1$. Then, given that A occurs, B is certain to occur. In this case, given that B occurs, A is more likely to occur than before

$$P(A|B) = P(AB)/P(B) = P(A)/P(B) > P(A)$$

The most extreme case of negative dependence is if A and B are mutually exclusive events B with $P(A) > 0$ and $P(B) > 0$. Then, given that A occurs, B cannot occur, and vice versa.

Example 1. Draws with and without replacement.

Consider two draws at random from a box of b black balls and w white balls, where $b > 0$, $w > 0$. Let Black_i and White_i denote the events of getting a black or a white ball on the ith draw, $i = 1, 2$. Then you can check that the dependence between pairs of these events from different draws is affected by whether the sampling is done with or without replacement, as shown in the following table.

Dependence Between Events on Different Draws

Pairs of events	Sampling with replacement	Sampling without replacement
$\text{Black}_1, \text{Black}_2$	independent	− dependent
$\text{Black}_1, \text{White}_2$	independent	+ dependent
$\text{White}_1, \text{White}_2$	independent	− dependent
$\text{White}_1, \text{Black}_2$	independent	+ dependent

The Sign of the Covariance

As a general rule, the sign of $Cov(X, Y)$ is *positive* if above-average values of X tend to be associated with above-average values of Y, and below-average values of X with below-average values of Y. The random variable $(X - \mu_X)(Y - \mu_Y)$ is then most likely positive, with a positive expectation.

The sign of $Cov(X, Y)$ is *negative* if above-average values of X tend to be associated with below-average values of Y, and vice versa. Then $(X - \mu_X)(Y - \mu_Y)$ is most likely negative, with a negative expectation.

$Cov(X, Y)$ is *zero* only in special cases when there is no such association between the variables X and Y. Then $(X - \mu_X)(Y - \mu_Y)$ has positive values balanced by negative values, and expected value zero.

While the sign of the covariance can be interpreted as above, its magnitude is hard to interpret. It is easier to interpret the *correlation of X and Y*, denoted here by $Corr(X, Y)$, which is defined as follows:

Definition of Correlation

$$Corr(X, Y) = \frac{Cov(X, Y)}{SD(X)SD(Y)}$$

Assume now that neither X nor Y is a constant, so $SD(X)SD(Y) > 0$. The sign of $Cov(X, Y)$ is then the same as the sign of $Corr(X, Y)$.

Conditions for X and Y to be Uncorrelated

The following three conditions are equivalent:

$$Corr(X,Y) = 0$$

$$Cov(X,Y) = 0$$

$$E(XY) = E(X)E(Y)$$

in which case X and Y are called *uncorrelated.* Independent variables are uncorrelated, but uncorrelated variables are not necessarily independent.

Let X^* and Y^* now denote X and Y rescaled to standard units. So

$$X^* = (X - \mu_X)/SD(X) \quad \text{and} \quad Y^* = (Y - \mu_Y)/SD(Y)$$

Then

$$E(X^*) = E(Y^*) = 0 \quad \text{and} \quad SD(X^*) = SD(Y^*) = 1$$

by the scaling properties of E and SD. And you can check that

$$Corr(X,Y) = Cov(X^*,Y^*) = E(X^*Y^*)$$

So correlation is a kind of standardized covariance that is unaffected by changes of origin or units of measurement. See Exercises.

Correlations are between -1 and +1

$$-1 \le Corr(X,Y) \le 1$$

no matter what the joint distribution of X and Y.

Proof. Since $E(X^{*2}) = E(Y^{*2}) = 1$

$$0 \le E(X^* - Y^*)^2 = 1 + 1 - 2E(X^*Y^*)$$

$$0 \le E(X^* + Y^*)^2 = 1 + 1 + 2E(X^*Y^*)$$

Thus $-1 \le E(X^*Y^*) \le 1$, and $Corr(X,Y) = E(X^*Y^*)$ by the preceding discussion. □

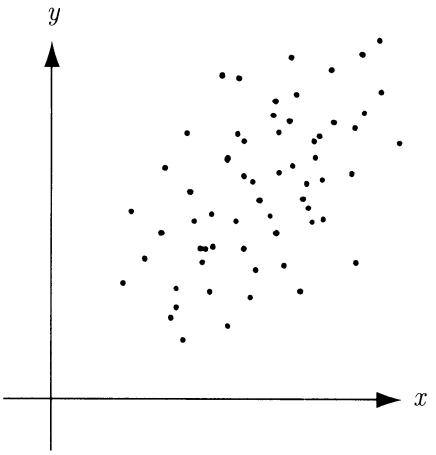
Correlations of ± 1. The proof that correlations are between ±1 shows $Corr(X,Y) = +1$ if and only if $E(X^* - Y^*)^2 = 0$, that is, if and only if $X^* = Y^*$ with probability one. This means there are constants a and b with $a > 0$ such that

$$Y = aX + b$$

with probability 1. That is to say, a correlation of $+1$ indicates a deterministic linear relationship between X and Y with positive slope. Similarly, a correlation of -1 indicates a deterministic linear relationship between X and Y with negative slope. Correlations between -1 and $+1$ indicate intermediate degrees of linear association between the two variables.

Example 2. Empirical correlations.

Like expectation and variance, covariance and correlation are generalizations to random variables of corresponding notions for empirical variables. Suppose $(x_1, y_1), \ldots, (x_n, y_n)$ is a list of n pairs of numbers, and (X,Y) is one of these pairs picked uniformly at random. Then the joint distribution of (X,Y) puts probability $1/n$ at each of the pairs, as suggested by the scatter diagram:



$$E(X) = \bar{x} \quad \text{and} \quad SD(X) = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2}$$

and similarly for Y instead of X. Also

$$E(XY) = \frac{1}{n}\sum_{k=1}^{n} x_k y_k \qquad \text{so}$$

$$Cov(X,Y) = E(XY) - E(X)E(Y) \qquad \text{and} \qquad Corr(X,Y) = \frac{Cov(X,Y)}{SD(X)SD(Y)}$$

can be computed from the list of number pairs. If the list of number pairs is a list of empirical measurements, or a sample of some kind, these may be called empirical or sample quantities. These quantities are all defined in terms of averages, which may be expected to converge to theoretical expectations as the sample size n increases, under conditions of random sampling. For example, the empirical correlation of n observed values of independent random variables (X_1, Y_1), $(X_2, Y_2), \ldots, (X_n, Y_n)$, all with the same joint distribution, will most probably be close to the theoretical correlation of X_1 and Y_1, provided n is sufficiently large. Thus a correlation in a theoretical model is often estimated by an empirically observed correlation based on a random sample. In particular, the empirical correlation of two variables over a large population can be estimated this way by the procedure of random sampling.

Example 3. Correlation and distribution of the sum.

This example shows in a simple case how the distribution of the sum of random variables X and Y is affected by their correlation. Suppose a gambler can bet on the value of a number U chosen uniformly at random from the numbers $1, 2, \ldots, 8$. The gambler can choose any set A of four numbers, such as $A = \{1, 2, 3, 4\}$, and place an even-money bet of \$1 on A. So the gambler wins \$1 if $U \in A$, and loses \$1 if $U \in A^c$. Let $\$X$ denote the gambler's net gain from this contract. Then, X has value $+1$ if $U \in A$, -1 if $U \in A^c$. In terms of indicators,

$$X = 2I_A - 1$$

Clearly $E(X) = 0$. The bet is fair no matter what set A the gambler chooses, because $P(A) = P(A^c) = 1/2$ for every set of four numbers A.

Suppose now that in addition to placing a bet on A, the gambler is also free to place at the same time a similar bet on a second set of four numbers B, for example $B = \{1, 3, 5, 7\}$. Let

$$Y = 2I_B - 1$$

denote the net gain to the gambler from this second bet. Then the gambler's overall gain from the placement of the two bets is the sum

$$S = X + Y$$

Notice that the distribution of X and the distribution of Y are the same, uniform on $\{-1, 1\}$, regardless of the gambler's choice of sets A and B. But the distribution of S is affected by the degree of dependence between X and Y, which is governed in turn by the amount of overlap between A and B. Clearly, $E(S)$ is zero no matter what the choice of A and B. But $SD(S)$ is affected by the gambler's choice of A and B. This standard deviation gives an indication of the likely size of the fluctuation in the gambler's fortune due to the combined bet.

Problem. Find how the standard deviation of S is determined by the choice of A and B.

Solution. Use the addition rule for variance

$$\begin{aligned} Var(S) &= Var(X) + Var(Y) + 2\,Cov(X,Y) \\ &= 2 + 2\,Corr(X,Y) \end{aligned}$$

because $SD(X) = SD(Y) = 1$, so $Corr(X,Y) = Cov(X,Y)$ in this case. Because $X = 2I_A - 1$, $Y = 2I_B - 1$, and the correlation coefficient is unchanged by linear transformations,

$$Corr(X,Y) = Corr(I_A, I_B) = \frac{Cov(I_A, I_B)}{SD(I_A)SD(I_B)} = \left(P(AB) - \frac{1}{4}\right) \Big/ \frac{1}{4} = 4P(AB) - 1$$

This used $P(A) = P(B) = 1/2$, which makes $SD(I_A) = SD(I_B) = 1/2$. Using the earlier expression for $Var(S)$ this gives

$$SD(S) = \sqrt{8P(AB)} = \sqrt{\#(AB)}$$

where $\#(AB)$ is the number of points in the intersection of A and B, so $P(AB) = \#(AB)/8$.

Discussion. The formula shows that the larger the overlap between A and B, the larger will be the likely size of the fluctuation in the gambler's fortune as a result of betting on both A and B. This is intuitively clear if you think about the following special cases:

Case $\#(AB) = 0$, $Corr(X,Y) = -1$, $SD(S) = 0$. This means $B = A^c$. Then $Y = -X$, because whatever is gained on one bet is lost on the other. So $S = X + Y = 0$. This is a strategy of extreme hedging, with zero result.

Case $\#(AB) = 1$, $Corr(X,Y) = -1/2$, $SD(S) = 1$. Intuitively, this is still hedging. The two bets tend to cancel each other.

Case $\#(AB) = 2$, $Corr(X,Y) = 0$, $SD(S) = \sqrt{2}$. In this case A and B are independent. Therefore, so too are the indicator random variables I_A and I_B, and the random variables $X = 2I_A - 1, Y = 2I_B - 1$ representing the net gains from the two bets. So the net effect of betting on both A and B in one game is the same as the effect of betting on A in one game, then betting on A again in a second game, independent of the first. The distribution of S in this case is the familiar binomial $(2, 1/2)$ distribution, but centered at 0 and rescaled by a factor of 2, because

$$S = X + Y = 2(I_A + I_B) - 2$$

where $I_A + I_B$ is the number of successes in two independent trials with probability $1/2$ of success on each trial, with binomial $(2, 1/2)$ distribution. The appearance of $\sqrt{2}$ as the standard deviation in this case illustrates the square root law for the standard deviation of the sum of $n = 2$ independent variables.

Case $\#(AB) = 3$, $Corr(X,Y) = 1/2$, $SD(S) = \sqrt{3}$. This is a bolder strategy.

Case $\#(AB) = 4$, $Corr(X,Y) = 1$, $SD(S) = 2$. Now $A = B$. All the gambler's eggs are in one basket. This is the boldest strategy for the gambler, effectively doubling the stake on A from \$1 to \$2.

Example 4. Red and black.

Let N_R be the number of reds that appear, N_B the number of blacks, in n spins of a roulette wheel that has proportion r of its numbers red, proportion b black, and the rest of its numbers green. (So $r + b < 1$. For a Nevada roulette wheel, as described at the end of Section 1.1, $r = b = 18/38$.)

Problem. Find $Corr(N_R, N_B)$.

Solution. Notice first, without calculation, that the answer ought to be negative for the usual case with $r+b \approx 1$. If $r+b = 1$ (no green numbers on the wheel) then $N_B = n - N_R$ which makes $Corr(N_R, N_B) = -1$. For $r + b \approx 1$ this relation is still approximately correct, so you should expect a correlation close to -1. Since N_R is a binomial (n, r) random variable,

$$E(N_R) = nr \text{ and } SD(N_R) = \sqrt{nr(1-r)}$$

and similarly for N_B, with b instead of r. Since

$$Cov(N_R, N_B) = E(N_R N_B) - E(N_R)E(N_B)$$

to calculate

$$Corr(N_R, N_B) = \frac{Cov(N_R, N_B)}{SD(N_R)SD(N_B)}$$

the only missing ingredient is $E(N_R N_B)$. You might try to calculate this from the joint distribution of N_R and N_B, but you will find this a frightful task. It is difficult to calculate even the variance of N_R directly from its binomial distribution, and the covariance with N_B is worse. The way around this difficulty is to use the connection between $Cov(N_R, N_B)$ and the variance of $N_R + N_B$

$$Var(N_R + N_B) = Var(N_R) + Var(N_B) + 2\,Cov(N_R, N_B)$$

The point is that $N_R + N_B$ is just the number of spins which are either red or black, which is a binomial $(n, r + b)$ random variable, with variance $n(r + b)(1 - r - b)$. Rearrange the equation and substitute all the variances to get

$$Cov(N_R, N_B) = \frac{1}{2}n\,[(r + b)(1 - r - b) - r(1 - r) - b(1 - b)] = -nrb,$$

hence,

$$Corr(N_R, N_B) = \frac{-nrb}{\sqrt{nr(1-r)}\sqrt{nb(1-b)}} = -\sqrt{\frac{rb}{(1-r)(1-b)}}$$

Discussion. In particular, for a Nevada roulette wheel,

$$r/(1-r) = b/(1-b) = 18/20 = 0.9 \qquad \text{so}$$

$$Corr(N_R, N_B) = -0.9$$

Note the interesting fact that the correlation does not depend at all on the number of spins n, only on the proportions of red and black. Also, the correlation is always negative, no matter what the proportions r and b.

Example 5. Correlations in the multinomial distribution.

Suppose the joint distribution of $(N_1, \ldots, N_m)$ is multinomial with parameters n and $(p_1, \ldots, p_m)$.

Problem. Find $Corr(N_i, N_j)$.

Solution. Call results in category i red, results in category j black, and results in all other categories green. Then the joint distribution of N_i and N_j is is the same as the joint distribution of N_R and N_B in the previous problem, for $r = p_i$, $b = p_j$. Since the correlation between two variables is determined by their joint distribution (by definition of correlation and the change of variable principle) this choice of r and b makes $Corr(N_i, N_j) = Corr(N_R, N_B)$. That is to say, from the solution of the previous problem,

$$Corr(N_i, N_j) = -\sqrt{\frac{p_i p_j}{(1-p_i)(1-p_j)}}$$

Correlation and Conditioning

An important connection between the ideas of correlation and conditioning is brought out by the following example.

Example 6. Sharkey's Casino.

At Sharkey's Casino the roulette wheels spin an average of one thousand times a day. Every day, Sharkey records the total numbers of red and black spins for the day on a computer. One day he notices that over the years he has been keeping data, the correlation between the number of reds and number of blacks has come out around

+0.8, rather than around −0.9 as predicted by the above calculation. Sharkey is very concerned that his roulette wheels are not obeying the laws of chance, and that someone might take advantage of it.

Problem. Should Sharkey get new roulette wheels?

Solution. Despite the fact that no matter what the number of spins n, the correlation between numbers of reds and blacks is −0.9, this does not imply that the same is true for a random number of spins, say N, the number of spins in a day picked at random at Sharkey's. While the expected value of N may be estimated as 1000 based on the long-run average of 1000 spins a day, it is reasonable to expect some spread in the distribution of N due to fluctuations in the number of customers and the rate of play. Since to a first approximation $N_B \approx \frac{18}{38}N$, $N_R \approx \frac{18}{38}N$, both N_B and N_R are positively correlated with N. If there is enough spread in the distribution of N, this will make for a positive correlation between N_B and N_R. So Sharkey need not be concerned, provided his data give a standard deviation of N consistent with a correlation of +0.8 between N_R and N_B.

To find the precise relation between $SD(N)$ and $Corr(N_R, N_B)$, for N_R and N_B, now numbers of reds and blacks in a random number N of spins, use the formula

$$Cov(N_R, N_B) = E(N_R N_B) - E(N_R)E(N_B)$$

where each expectation can be computed by conditioning on N. First, if N is treated as a constant, then by previous calculations,

$$E(N_R) = Nr \qquad E(N_B) = Nb$$

$$E(N_R N_B) = E(N_R)E(N_B) + Cov(N_R, N_B) = N^2 rb - Nrb$$

For random N, these are *conditional* expectations given N. But since expectations are expectations of conditional expectations, this gives

$$\begin{aligned} E(N_R) &= E(N)r, \qquad E(N_B) = E(N)b \\ E(N_R N_B) &= E(N^2)rb - E(N)rb, \qquad \text{hence} \\ Cov(N_R, N_B) &= E(N_R N_B) - E(N_R)E(N_B) \\ &= rb\left[E(N^2) - E(N) - [E(N)]^2\right] \\ &= rb\left[Var(N) - E(N)\right] \end{aligned}$$

In particular, $Cov(N_R, N_B)$ will be positive provided $Var(N) > E(N)$. Thus for $E(N) = 1000$, if $SD(N) > \sqrt{1000} \approx 32$, there will be a positive correlation between N_R and N_B. The same method of calculation gives

$$Var(N_B) = b^2\, Var(N) + b(1-b)E(N)$$

For $b = r$ this gives

$$\begin{aligned} Corr(N_R, N_B) &= \frac{b^2[Var(N) - E(N)]}{b^2\, Var(N) + b(1-b)E(N)} \\ &= \frac{9\, Var(N) - 9000}{9\, Var(N) + 10,000} \qquad \text{for} \quad b = \frac{18}{38}, \quad E(N) = 1,000. \end{aligned}$$

If $Var(N) = 0$ this simplifies to -0.9 as before. But as $Var(N)$ increases the correlation increases, and approaches 1 for large values of $Var(N)$. Set $Corr(N_R, N_B) = \rho$ and solve for $SD(N) = \sqrt{Var(N)}$ to get

$$\begin{aligned} SD(N) &= \sqrt{\frac{9000 + 10,000\rho}{9(1-\rho)}} \\ &= \sqrt{\frac{17,000}{9 \times 0.2}} \qquad \text{for } \rho = 0.8 \\ &\approx 100 \end{aligned}$$

So a correlation of 0.8 between N_R and N_B is consistent with a standard deviation of about 100 for the number of spins per day. Provided that is the case, Sharkey need not be concerned.

Discussion. The example makes the important point that two variables, like N_R and N_B, may be positively correlated due to association with some third variable, like N, even if there is zero or negative correlation between the two variables for a fixed value of N. Here is another example. For children of a fixed age, the correlation between height and reading ability would most likely come out around zero. But if you looked at children of ages from 5 to 10, there would be a high positive correlation between height and reading ability, because both variables are closely associated with age. For data variables, looking at distributions or relationships between some variables for a fixed value of another variable, N say, is called *controlling for* N. In a probability model the corresponding thing is *conditioning on* N. Whether or not you condition or control on one variable typically has major effects on relationships between other variables.

The calculations in the example show in general that for two mutually exclusive outcomes in independent trials, like red and black at roulette, the counts of results of the two kinds that occur in any fixed number of trials will be negatively correlated. If the number of trials N is random, the two counts will be positively or negatively correlated according to whether $Var(N) > E(N)$ or $Var(N) < E(N)$. In the case where $Var(N) = E(N)$, the two counts will be uncorrelated. In particular this is the case if N has a Poisson distribution. Then the two counts are actually independent. See Exercise 6.1.8.

Variance of a Sum of *n* Variables

The general formula involving covariance for the variance of a sum of two random variables has the following extension to n variables. The formula shows that the simple addition rule for the variance of a sum of independent random variables works just as well for uncorrelated ones, but in general there are $\binom{n}{2}$ covariance terms to be considered as well.

Variance of a Sum of *n* Variables

$$Var\left(\sum_k X_k\right) = \sum_k Var(X_k) + 2\sum_{j<k} Cov(X_j X_k)$$

where $\sum_k$ denotes a sum of n terms from $k = 1$ to n, and $\sum_{j<k}$ denotes a sum of $\binom{n}{2}$ terms indexed by j and k with $1 \le j < k \le n$.

Proof: The variance of the sum is by definition the expectation of

$$\left[\sum_k X_k - E(\sum_k X_k)\right]^2 = \left[\sum_k X_k - \sum_k \mu_k\right]^2 \qquad \text{where } \mu_k = E(X_k)$$

$$= \left[\sum_k (X_k - \mu_k)\right]^2$$

$$= \sum_k (X_k - \mu_k)^2 + 2\sum_{j<k} (X_j - \mu_j)(X_k - \mu_k)$$

by the algebraic identity

$$\left(\sum a_k\right)^2 = \sum_k a_k^2 + 2\sum_{j<k} a_j a_k$$

applied to $a_k = X_k - \mu_k$. Now use the linearity of expectation and the definition of $Cov(X_j, X_k)$. In the sum over all $j < k$, there are exactly $\binom{n}{2}$ terms, one for each way of choosing two indices $j < k$ from the set $\{1, 2, \ldots, n\}$. □

Example 7. Variance of sample averages.

Let $x(1), x(2), \ldots, x(N)$ be a list of N numbers. Think of $x(k)$ as representing the height of the kth individual in a population of size N. Let

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x(k) \quad \text{and} \quad \sigma^2 = \frac{1}{n}\sum_{k=1}^{n} [x(k) - \bar{x}]^2$$

So $\bar{x}$ is the *population mean*, and σ^2 is the *population variance*. Let $X_1, X_2, \ldots, X_n$ be the heights obtained in a random sample of size n from this population. More formally, for $i = 1, 2, \ldots, n$, the ith height in the sample is $X_i = x(K_i)$, where $K_1, K_2, \ldots, K_n$ is a random sample of size n from the index set $\{1, 2, \ldots, N\}$. This random sample might be taken either with replacement or without replacement. Either way, each random index K_i has uniform distribution over $\{1, 2, \ldots, N\}$, by symmetry. So each X_i is distributed according to the distribution of the list of heights in the total population, with

$$E(X_i) = \bar{x} \quad \text{and} \quad SD(X_i) = \sigma \qquad (i = 1, 2, \ldots, n)$$

Let

$$\bar{X}_n = (X_1 + X_2 + \ldots + X_n)/n$$

be the *sample average*. This is the average height of individuals in the sample of size n. Note that this is a random variable: repeating the sampling procedure will typically produce a different sample average. Whereas $\bar{x}$, the population average, is a constant. Since $E(X_i) = \bar{x}$ for $i = 1, 2, \ldots, n$, the rules of expectation imply that also

$$E(\bar{X}_n) = \bar{x}$$

still no matter whether the sampling is done with or without replacement. In the case with replacement, the random variables X_i are independent, all with standard deviation σ, so

$$SD(\bar{X}_n) = \sigma/\sqrt{n} \qquad \text{(with replacement)}$$

by the square root law of Section 3.3. So the average height in a random sample of size n is most likely only a few multiples of $\sigma/\sqrt{n}$ away from the population average $\bar{x}$. If σ can be bounded or estimated, this gives an indication of the quality of the sample average $\bar{X}_n$ as an estimator of the unknown population average $\bar{x}$.

Intuitively, for sampling without replacement, $\bar{X}_n$ should provide a better estimate of $\bar{x}$ than for sampling with replacement. In this case, the random variables $X_1, \ldots, X_n$ turn out to be negatively correlated, which affects the formula for $SD(\bar{X}_n)$. The problem is how to correct for the dependence.

Problem. Calculate $SD(\bar{X}_n)$ for sampling without replacement.

Solution. Let $S_n = X_1 + \cdots + X_n$, so $\bar{X}_n = S_n/n$. Then

$$\begin{aligned} Var(S_n) &= \sum_j Var(X_j) + 2\sum_{j<k} Cov(X_j, X_k) \\ &= n\sigma^2 + n(n-1)\,Cov(X_1, X_2), \end{aligned}$$

because $Cov(X_j, X_k) = Cov(X_1, X_2)$ by the symmetry of sampling without replacement discussed in Section 3.6: (X_j, X_k) is for every $j < k$ a simple random sample of size 2, with the same distribution as (X_1, X_2). This formula for $Var(S_n)$ holds for every sample size n with $1 \leq n \leq N$. But for $n = N$

$$S_N = x_1 + x_2 + \cdots + x_N$$

is constant, because in a complete sample of the population each element appears exactly once, so the sum defining S_N is just the sum on the right done in a random order. Thus $Var(S_N) = 0$. Comparison with the previous formula for $Var(S_n)$, in the case where $n = N$, shows

$$Cov(X_1, X_2) = -\sigma^2/(N-1)$$

hence

$$Var(S_n) = n\sigma^2 \left[1 - \frac{n-1}{N-1}\right]$$

and

$$SD(\bar{X}_n) = SD(S_n)/n = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

Discussion. This shows that the standard deviation for the average in sampling without replacement is the corresponding standard deviation for sampling with replacement, reduced by the *correction factor* $\sqrt{\frac{N-n}{N-1}}$. The same is true for the sum as well as the average, by scaling.

The same correction factor appears in the formula for the variance of the hypergeometric distribution, calculated in Section 3.6. Though covariances are not used in that calculation, it is still a special case of the current example, with $x_j = 0$ or 1 for every j.

It is remarkable that the same correction factor works no matter what the distribution of the empirical variable x. The correction factor takes care of the slight negative correlation between terms, which also does not depend on the distribution of x:

$$Corr(X_j, X_k) = \frac{Cov(X_j, X_k)}{SD(X_j)SD(X_k)} = -1/(N-1)$$

The correlation is negative because observation of a large value of X_j removes a large value from the population, and tends to make large values of X_k less likely. Similarly, small values of X_j tend to make small values of X_k less likely. This means there is a greater tendency for the deviations $X_j - E(X_j)$ to cancel each other out

for sampling without replacement than for sampling with replacement, when these deviations are independent. This reduces the likely size of the deviation for the sum

$$S_n - E(S_n) = \sum_{j=1}^{n} (X_j - E(X_j))$$

Ultimately, for $n = N$, the deviation of S_N is zero, which was the key to calculating the correction factor.

Bilinearity of Covariance

The following formulae for covariances of linear combinations of variables are easily derived from the definition. These formulae can often be used to simplify covariance calculations.

$$Cov(X, Y + Z) = Cov(X, Y) + Cov(X, Z)$$

$$Cov(W + X, Y) = Cov(W, Y) + Cov(X, Y)$$

For constants a and b

$$Cov(aX, Y) = a\,Cov(X, Y) \quad \text{and} \quad Cov(X, bY) = b\,Cov(X, Y)$$

and so on for linear combinations of several variables. For example

$$Cov(aW+bX, cY+dZ)=a\,c\,Cov(W, Y)+a\,d\,Cov(W, Z)+b\,c\,Cov(X, Y)+b\,d\,Cov(X, Z)$$

To summarize:

Covariance is Bilinear

$$Cov\left(\sum_i a_i X_i\ ,\ \sum_j b_j Y_j\right) = \sum_i \sum_j a_i\, b_j\, Cov(X_i, Y_j)$$

Here the a_i and b_j are arbitrary constants. If there are n terms in the sum over i and m terms in the sum over j there are nm terms in the double sum on the right side. Taking $n = m$, $a_i = b_i = 1$ and $X_i = Y_i$ for $1 \le i \le n$, this formula reduces to the formula for the variance of $\sum_i X_i$.

Exercises 6.4

1. Suppose A, B are two events such that $P(A) = 0.3$, $P(B) = 0.4$, and $P(A \cup B) = 0.5$.

a) Find $P(A|B)$. b) Are A and B independent, positively or negatively dependent?

c) Find $P(A^c B)$. d) Let $X = I_A$, $Y = I_B$. Find $Corr(X, Y)$.

2. Use the formula $P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$ to prove:

a) if $P(A|B) = P(A|B^c)$ then A and B are independent;

b) if $P(A|B) > P(A|B^c)$ then A and B are positively dependent;

c) if $P(A|B) < P(A|B^c)$ then A and B are negatively dependent.

Now prove the converses of a), b), and c).

3. Suppose that the failures of two components are positively dependent. If the first component fails, does that make it more or less likely that the second component works? What if the first component works?

4. Let (X, Y) have uniform distribution on the four points $(-1, 0), (0, 1), (0, -1), (1, 0)$. Show that X and Y are uncorrelated but not independent.

5. Let X have uniform distribution on $\{-1, 0, 1\}$ and let $Y = X^2$. Are X and Y uncorrelated? Are X and Y independent? Explain carefully.

6. Let X_1 and X_2 be the numbers on two independent fair die rolls, $X = X_1 - X_2$ and $Y = X_1 + X_2$. Show that X and Y are uncorrelated, but not independent.

7. Let X_2 and X_3 be indicators of independent events with probabilities $1/2$ and $1/3$, respectively.

a) Display the joint distribution table of $X_2 + X_3$ and $X_2 - X_3$.

b) Calculate $E(X_2 - X_3)^3$.

c) Are X_2 and X_3 uncorrelated? Prove your answer.

8. You have N boxes labeled Box1, Box2, ..., BoxN, and you have k balls. You drop the balls at random into the boxes, independently of each other. For each ball the probability that it will land in a particular box is the same for all boxes, namely $1/N$. Let X_1 be the number of balls in Box1 and X_N be the number of balls in BoxN. Calculate $Corr(X_1, X_N)$.

9. Suppose n cards numbered $1, 2, \ldots, n$ are shuffled and k of the cards are dealt. Let S_k be the sum of the numbers on the k cards dealt. Find formulae in terms of n and k for:

a) the mean of S_k; b) the variance of S_k.

10. **Overlapping counts.** A fair coin is tossed 300 times. Let H_{100} be the number of heads in the first 100 tosses, and H_{300} the total number of heads in the 300 tosses. Find $Corr(H_{100}, H_{300})$.

11. Let T_1 and T_3 be the times of the first and third arrivals in a Poisson process with rate λ. Find $Corr(T_1, T_3)$.

12. Suppose α, β, γ denote the proportions of Democrats (D), Republicans (R) and Others (O) in a large population of voters. (So $0 \leq \alpha, \beta, \gamma \leq 1$ and $\alpha+\beta+\gamma = 1$.) An individual is selected at random from the population. Write $X = 1, Y = 0, Z = 0$ if that individual is D, write $X = 0, Y = 1, Z = 0$ if the individual is R and write $X = 0, Y = 0, Z = 1$ if the individual is O. Find:

a) $E(X)$, $E(Y)$; b) $Var(X)$, $Var(Y)$; c) $Cov(X, Y)$.

Suppose next that n individuals are selected independently and randomly with replacement from the population. The total number of D's may be written, $D_n = X_1 + ... + X_n$. Similarly let $R_n = Y_1 + ... + Y_n$. and let $O_n = Z_1 + ... + Z_n$. Let $D_n - R_n$ denote the excess of D's over R's selected. Find d) $E(D_n - R_n)$; e) $Var(D_n - R_n)$.

13. Let A and B be two possible results of a trial, not necessarily mutually exclusive. Let N_A be the number of times A occurs in n independent trials, N_B the number of times B occurs in the same n trials. True or false and explain: If N_A and N_B are uncorrelated, then they are independent.

14. Show that for any two random variables X and Y

$$|SD(X) - SD(Y)| \leq SD(X + Y) \leq |SD(X) + SD(Y)|$$

15. Covariance is bilinear. Show from the definition of covariance that:

a) $Cov(X, Y + Z) = Cov(X, Y) + Cov(X, Z)$

b) $Cov(W + X, Y) = Cov(W, Y) + Cov(X, Y)$

c) $Cov(\sum_i X_i, \sum_j Y_j) = \sum_i \sum_j Cov(X_i, Y_j)$

d) Use c) to rederive the formula for $Cov(N_R, N_B)$ in Example 6.

16. Invariance of the correlation coefficient under linear transformations. Show that for arbitrary random variables X and Y, and constants a, b, c, d with $a \neq 0$, $c \neq 0$,

$$Corr(aX + b, cY + d) = \begin{cases} Corr(X, Y) & \text{if } a \text{ and } c \text{ have the same sign} \\ -Corr(X, Y) & \text{if } a \text{ and } c \text{ have opposite signs.} \end{cases}$$

Thus the correlation coefficients are affected only by the sign of a linear change of variable. They are therefore unaffected by shifts of origin or changes of units.

17. Show that for indicator random variables I_A and I_B of events A and B

$$Corr(I_A, I_B) = Corr(I_{A^c}, I_{B^c}) = -Corr(I_A, I_{B^c}) = -Corr(I_{A^c}, I_B)$$

Deduce that if A and B are positively dependent, then so are A^c and B^c, but A and B^c are negatively dependent, as are A^c and B.

18. Random variables $X_1, \ldots, X_n$ are *exchangeable* if their joint distribution is the same, no matter what order they are presented (see Section 3.6). Show that if $X_1, \ldots, X_n$ are exchangeable, then

$$Var(\sum_{k=1}^{n} X_k) = n\, Var(X_1) + n(n-1)\, Cov(X_1, X_2)$$

19. A box contains 5 nickels, 10 dimes, and 25 quarters. Suppose 20 draws are made at random without replacement from this box. Let X be the total sum obtained in these 20 draws. Calculate: a) $E(X)$; b) $SD(X)$;

c) $P(X \leq \$3)$ using the normal approximation.

d) Can you imagine why these calculations might give results inconsistent with long-run repetitions of the sampling experiment? For each of a) and c), say whether your reasoning would suggest higher or lower long-run averages.

20. Correlation and conditioning. A random variable X assumes values x_1 and x_2 with probabilities p_1 and p_2, where $p_1 + p_2 = 1$. Given $X = x_i$, random variable Y has mean equal to μ_i and SD equal to σ_i. Find formulae in terms of x_i, p_i, μ_i, and σ_i, $i = 1, 2$, for the following quantities:

a) $E(X)$; b) $E(Y)$; c) $SD(X)$; d) $SD(Y)$; e) $Cov(X, Y)$; f) $Corr(X, Y)$.

Indicate how these formulae could be generalized to the case of X with n possible values $x_1, \ldots, x_n$.

21. A box contains 5 red balls and 8 blue ones. A random sample of size 3 is drawn *without* replacement. Let X be the number of red balls and let Y be the number of blue balls selected. Compute: a) $E(X)$; b) $E(Y)$; c) $Var(X)$; d) $Cov(X, Y)$.

22. Suppose there were m married couples, but that d of these $2m$ people have died. Regard the d deaths as striking the $2m$ people at random. Let X be the number of surviving couples. Find:

a) $E(X)$; b) $Var(X)$.

23. Linear prediction and the correlation coefficient. For random variables X and Y, the *linear prediction problem* for predicting Y based on knowledge of X is the problem of finding a linear function of X, $\beta X + \gamma$, which minimizes the *mean square* of the prediction error

$$MSE = E[Y - (\beta X + \gamma)]^2$$

(Compare with Exercise 6.2.17 where the predictor of Y could be an arbitrary function of X.) This exercise derives the basic formulae for the best linear predictor according to this criterion.

a) Expand out the MSE using algebra, and regard it as a quadratic function of γ and β with coefficients involving the numbers $E(X)$, $E(Y)$, $E(XY)$, etc.

b) Differentiate this function with respect to γ to show that for fixed β, the unique γ which minimizes the MSE is $\hat{\gamma}(\beta) = E(Y) - \beta E(X)$. What is the resulting minimal MSE called when $\beta = 0$?

c) Consider now the MSE as a function of β, with $\gamma = \hat{\gamma}(\beta)$ the best γ for the given β. Differentiate this function with respect to β, and show that it is minimized at $\hat{\beta} = Cov(X, Y)/Var(X)$ where it is assumed that $Var(X) > 0$.

d) Deduce that the unique pair (β, γ) which minimizes the MSE is $(\hat{\beta}, \hat{\gamma}(\hat{\beta}))$.

e) Let $\hat{Y} = \hat{\beta}X + \hat{\gamma}$ now denote this best linear predictor. Show that

$$E(\hat{Y}) = E(Y); \quad Var(\hat{Y}) = \hat{\beta}^2 Var(X); \quad E[\hat{Y}(Y - \hat{Y})] = 0$$

f) Deduce that the variance of Y can be decomposed into the sum of the variance of the best predictor $\hat{Y}$ and the minimum MSE according to the formula

$$Var(Y) = Var(\hat{Y}) + E[(Y - \hat{Y})^2]$$

with $Var(\hat{Y}) = \rho^2 Var(Y)$ and $E[(Y - \hat{Y})^2] = (1 - \rho^2) Var(Y)$ where $\rho = Corr(X, Y)$.

g) It is customary to express the slope $\hat{\beta}$ of the best linear predictor $\hat{Y} = \hat{\beta}X + \hat{\gamma}$ in terms of ρ. Show that $\hat{\beta} = \rho SD(Y)/SD(X)$ and that the intercept $\hat{\gamma}$ is then uniquely determined by the requirement that the line $y = \hat{\beta}x + \hat{\gamma}$ passes through the point $(E(X), E(Y))$.

h) Let $Y^* = (Y - E(Y))/SD(Y)$, $X^* = (X - E(X))/SD(X)$. Show that the best linear predictor of Y^* based on X^* is just ρX^*. So the correlation coefficient ρ is simply the slope of the best linear predictor when the variables are expressed in standard units.

6.5 Bivariate Normal

The radially symmetric bivariate normal distribution corresponding to independent normal variables was considered in Section 5.3. This section uses the tools of previous sections to analyze correlated normal variables by making a linear transformation to the simpler case of independent variables.

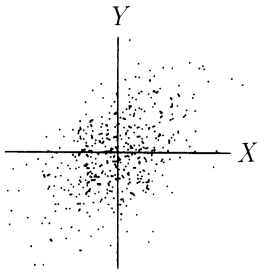
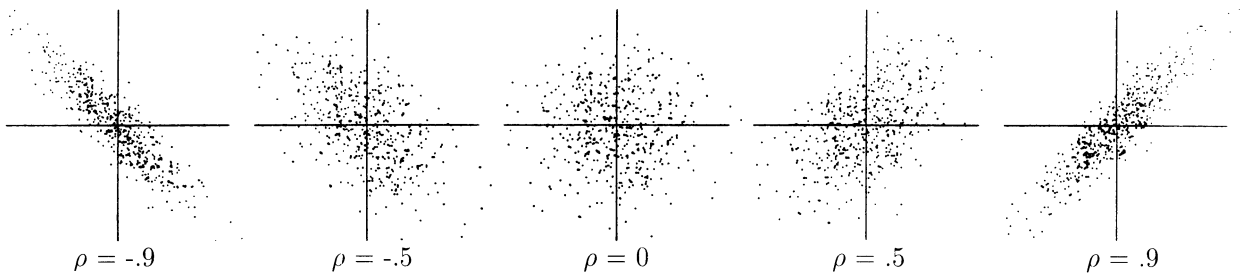


FIGURE 1. Bivariate normal scatter.
The diagram shows points picked at random from a bivariate distribution, in which the coordinates X and Y each have the same normal distribution, but are not independent. The two variables are *positively correlated*, which makes the cloud elliptical, sloping upward to the right and downwards to the left.

Clouds of data like this are very common in statistical analysis. They were first examined by the British scientist Francis Galton (1822–1911), who studied relations between variables like a father's height and his son's height. To display visually how two variables are related, a *scatter diagram* like Figure 1 may be used. In such a diagram, data pairs are represented by plotting a point at the coordinates of each pair. The hereditary connection between a father's height and his son's height makes the variables positively correlated—taller fathers tend to have taller sons, taller sons tend to have taller fathers. But the relation is not a rigid one, since the son's height is not a deterministic function of his father's height. The dependence between the two variables is more interesting and subtle. When variables are measured in their standard units, this dependence shows up in a scatter diagram as a tendency to form an elliptical cloud along a diagonal. The cloud has a major axis along the line $Y = X$ at 45° to the axes in the case of positive correlation, and a major axis along the perpendicular line $Y = -X$ in the case of negative correlation.

FIGURE 2. Bivariate normal scatters for various correlations ρ.



The object now is to describe this kind of dependence between variables by representing correlated normal variables as linear functions of independent ones. This is a powerful technique which is the basis for much statistical analysis of two or more

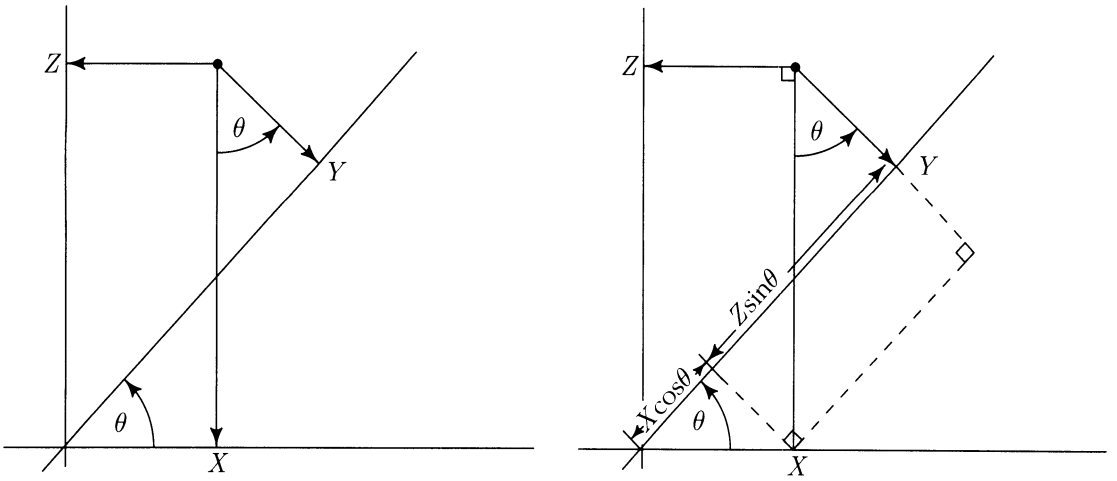
variables. A basic ingredient is the correlation coefficient, denoted here by ρ, often also by r:

$$\rho = Corr(X, Y) = E(X^*Y^*)$$

where X^* is X in standard units, and Y^* is Y in standard units. This correlation ρ is a theoretical quantity, defined by expected values or integrals with respect to a bivariate distribution. In practice, such correlations are usually estimated by the corresponding empirical correlation obtained from data, with the empirical distribution of a data list $(x_1, y_1), \ldots, (x_n, y_n)$ instead of the theoretical distribution, and averages instead of expectations.

Constructing Correlated Normal Variables

To get a pair of correlated standard normal variables X and Y, start with a pair of independent standard normal variables, say X and Z. Let Y be the projection of (X, Z) onto an axis at an angle θ to the X-axis, as in the left-hand diagram:



By the geometry of the right-hand diagram

$$Y = X\cos\theta + Z\sin\theta$$

By rotational symmetry of the joint distribution of X and Z, the distribution of Y is standard normal. Thus

$$\begin{aligned} E(X) &= E(Y) = E(Z) = 0 \\ SD(X) &= SD(Y) = SD(Z) = 1 \\ \rho(X,Y) = E(XY) &= E[X(X\cos\theta + Z\sin\theta)] \\ &= E(X^2)\cos\theta + E(XZ)\sin\theta \\ &= \cos\theta \end{aligned}$$

since $E(X^2) = 1$, and $E(XZ) = E(X)E(Z) = 0$ by independence of X and Z. To summarize, X and Y are standard normal variables with correlation $\rho = \cos\theta$. Note the special cases

$\theta = 0$	when $\rho = 1$	$Y = X$
$\theta = \pi/2$	when $\rho = 0$	$Y = Z$ is independent of X
$\theta = \pi$	when $\rho = -1$	$Y = -X$

For each ρ between -1 and 1, there is an angle $\theta = \arccos\rho$, which makes X and Y have correlation ρ. Then $\cos\theta = \rho$, $\sin\theta = \sqrt{1-\rho^2}$, and

$$Y = \rho X + \sqrt{1-\rho^2}Z$$

where X and Z are independent normal (0, 1). The joint distribution of X and Y so defined is the *standard bivariate normal distribution with correlation* ρ.

Standard Bivariate Normal Distribution

X and Y have standard bivariate normal distribution with correlation ρ if and only if

$$Y = \rho X + \sqrt{1-\rho^2}Z$$

where X and Z are independent standard normal variables.

Marginals. Both X and Y have standard normal distribution.

Conditionals. Given $X = x$, Y has normal $(\rho x, 1-\rho^2)$ distribution. Given $Y = y$, X has normal $(\rho y, 1-\rho^2)$ distribution.

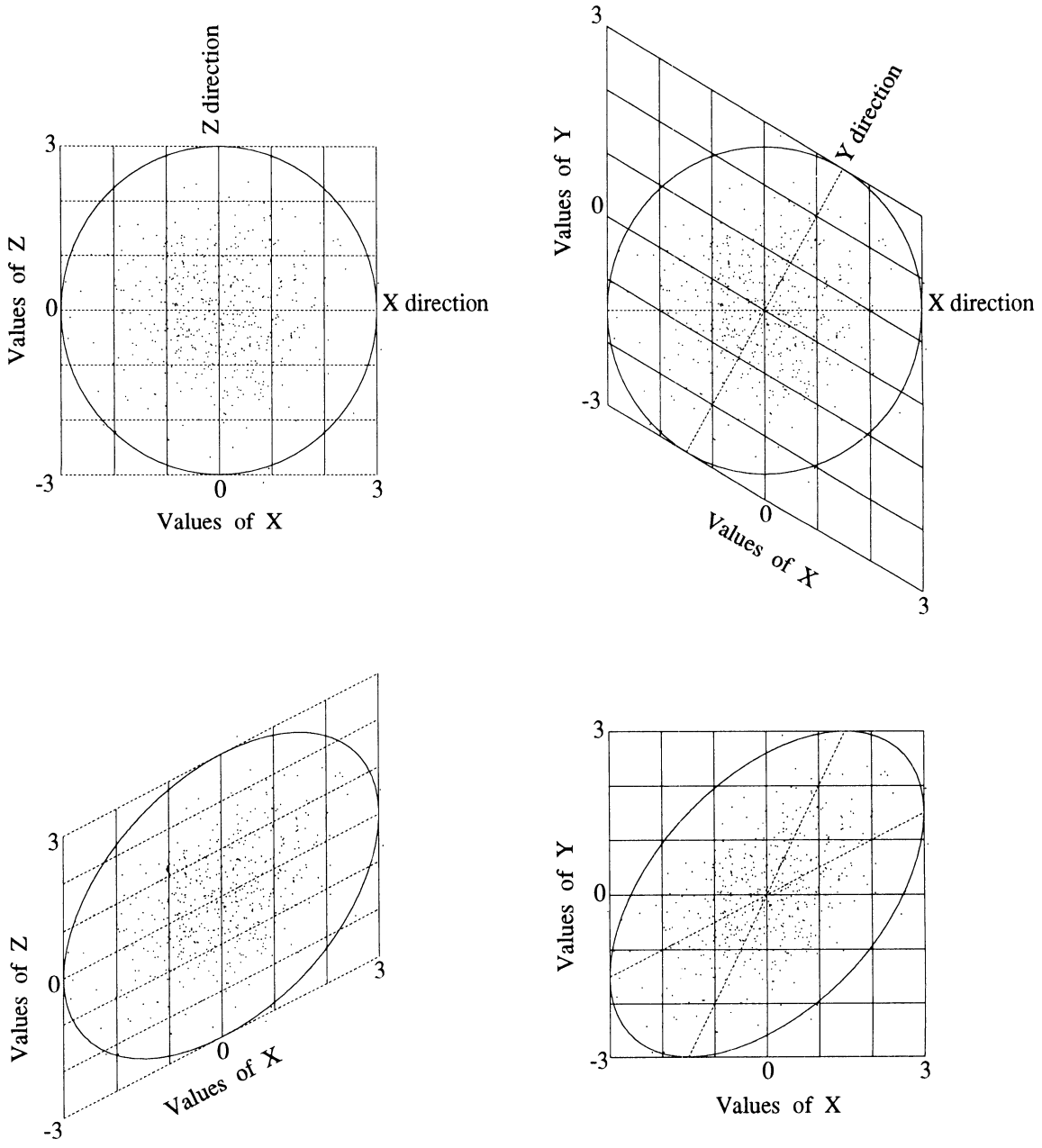
Joint density. The joint density of X and Y is

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right\}$$

Independence. For X and Y with standard bivariate normal distribution, X and Y are independent if and only if $\rho = 0$.

The next two pages display the geomerty of linear transformation from (X, Z) to (X, Y). Following these pages is a discussion of the results presented in the above box.

FIGURE 3. **Geometry of the bivariate normal distribution.** Properties of the standard bivariate normal distribution with correlation ρ may be understood in terms of the simplest case $\rho = 0$ by the geometry of the linear transformation $(X, Z) \mapsto (X, Y)$, displayed here for $\theta = 60°$, so $\rho = \cos\theta = \frac{1}{2}$, $\sqrt{1-\rho^2} = \sin\theta = \frac{\sqrt{3}}{2}$ and $Y = \frac{1}{2}X + \frac{\sqrt{3}}{2}Z$.



Key to Figure 3.

Top left panel. This shows a computer-generated scatter of 500 points picked at random according to the joint distribution of X and Z, plotted in the usual way with rectangular X and Z coordinates. This is a roughly circular cloud, due to the rotational symmetry of the distribution of two independent standard normals. The circle is the contour of constant density for (X, Z), of radius 3 standard units, containing 98.9% of the probability. The vertical lines represent the events $X = 0, \pm 1, \pm 2, \pm 3$. The dashed horizontal lines represent $Z = 0, \pm 1, \pm 2, \pm 3$.

Top right panel. This is the same scatter in the (X, Z) plane, but with the diagonal lines $Y = 0, \pm 1, \pm 2, \pm 3$. The Y direction is the dotted line at angle $\theta = 60°$ to the horizontal X direction. The diagonals $Y = \textit{constant}$ are at angle θ to the vertical lines $X = \textit{constant}$.

Bottom right panel. This is the image of the top right panel after shearing and shrinking to represent X and Y by new rectangular axes. Each point in the top scatter is transformed into one in the bottom scatter. Thus the cloud becomes a random scatter of 500 points picked at random according to the bivariate normal distribution of X and Y, with correlation $\rho = \cos\theta$. Think of the lines in the top right panel as a lattice of rigid rods attached by pins. Keep the vertical axis $X = 0$ fixed, and shear the lattice so the diagonals become horizontal. This makes a lattice of squares of side $1/\sin\theta$. Now shrink everything by a factor of $\sin\theta$ to get the bottom-right panel.

The shearing which turns the diamonds into squares turns the circle into an ellipse, with major axis on the 45-degree line through the new origin. This is an ellipse of constant density for (X, Y). The images of the dotted lines in the old X and Y directions are the dotted lines $Y = \rho X$ and $X = \rho Y$. These are the *regression lines* discussed further in the next paragraph.

Bottom left panel. This is the image of the top left panel by the same transformation from (X, Z) to (X, Y). The ellipse and the cloud of points are the same as in the bottom right panel. But now the lines representing $X = 0, \pm 1, \pm 2, \pm 3$ are shown, along with those representing $Z = 0, \pm 1, \pm 2, \pm 3$. The line $Z = 0$ plays a particularly important role. This is the *regression line*. The equation of this line $Z = 0$ in the (X, Y) plane is

$$Y = \rho X$$

where ρ is the correlation. Geometrically, this is the line of midpoints of vertical sections of the ellipse. Statistically, it is the best predictor of Y based on X.

The properties of the standard bivariate normal distribution stated in the box on page 451 all follow from the basic representation

$$Y = \rho X + \sqrt{1-\rho^2}Z \tag{1}$$

in terms of independent standard normal X and Z.

Conditionals. The formula for the distribution of Y given $X = x$ is immediate from (1). Conditioning on X does not affect the distribution of Z. And given $X = x$ you can treat X in (1) as the constant x, so Y is then just a linear transformation of the standard normal variable Z with coefficients involving ρ and x. This gives the conditional distribution of Y given $X = x$. The distribution of X given $Y = y$ follows by symmetry, or from (1′) below.

Symmetry. The standard bivariate normal distribution of (X, Y) is symmetric with respect to switching X and Y. This can be seen from the formula for the joint density, which is a symmetric function of x and y, or from the geometric description of X and Y. This symmetry is obscured in formula (1) however. You should check as an exercise that (1) has a dual

$$X = \rho Y + \sqrt{1-\rho^2}\,Z' \tag{1′}$$

where Z' is a linear combination of X and Z that is independent of Y.

Joint density. The derivation of this is an exercise: Write out the formulae for the marginal and conditional densities, multiply, and simplify. There is no point remembering this formula. Rather, take the following:

Advice. Do not attempt to compute bivariate normal probabilities or expectations by integrating against the joint density. It is always simpler to rewrite the problem in terms of independent variables X and Z, using (1). This technique is used in all the examples below.

Bivariate Normal Distribution

Random variables U and V have *bivariate normal distribution* with parameters μ_U, μ_V, σ_U^2, σ_V^2, and ρ if and only if the standardized variables

$$X = (U - \mu_U)/\sigma_U \qquad Y = (V - \mu_V)/\sigma_V$$

have standard bivariate normal distribution with correlation ρ. Then

$$\rho = Corr(X, Y) = Corr(U, V)$$

and U and V are independent if and only if $\rho = 0$.

Examples

The point of the following examples is to show how any problem involving random variables U and V with a bivariate normal distribution can be solved by a simple three-step procedure:

- **Step 1.** Express U and V in terms of the standardized variables X and Y.
- **Step 2.** Write $Y = \rho X + \sqrt{1-\rho^2}Z$ to reduce the problem to one involving two independent standard normal variables X and Z.
- **Step 3.** Solve the reduced problem involving X and Z by exploiting independence or rotational symmetry.

Example 1. Fathers and sons.

Galton's student Karl Pearson carried out a study on the resemblances between parents and children. He measured the heights of 1078 fathers and sons, and found that the sons averaged one inch taller than the fathers:

Fathers:	mean height: 5′9″	SD: 2″
Sons:	mean height: 5′10″	SD: 2″
	correlation: 0.5	

Problem 1. Predict the height of the son of a father who is 6′2″ tall.

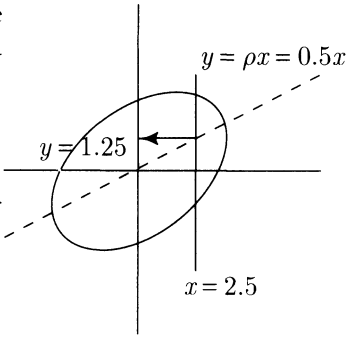
Solution. Assume that the data are approximately bivariate normal in distribution. Then the parameters can be estimated by the corresponding empirical measurements.

Let X be the father's height in standard units, and Y be the son's height in standard units. The assumption of a bivariate normal distribution makes

$$Y = \rho X + \sqrt{1-\rho^2}Z$$

where Z is standard normal independent of X. The natural prediction for Y given $X = x$ is

$$E(Y|X = x) = \rho x$$



Here the given value of X is

$$x = 6'2'' \text{ converted to standard units}$$
$$= (6'2'' - 5'9'')/2'' = 2.5 \text{ standard units}$$

So the predicted value of Y is

$$E(Y|X = x) = 0.5 \times 2.5 = 1.25 \text{ standard units},$$

That is,

$$\begin{aligned}\text{predicted son's height} &= 5'10'' + 2''Y \\ &= 5'10'' + 2'' \times 1.25 = 6'\,0.5''\end{aligned}$$

Discussion. Though the father is exceptionally tall (height $6'2''$), the son is not predicted to be $6'2''$, but only $6'0.5''$ tall. Galton called this phenomenon *regression to the mean.*

Problem 2. What is the chance that your prediction is off by more than 1 inch?

Solution. Since 1 inch is 0.5 times the SD of sons' heights, and we are given $X = 2.5$, the problem in standard units is to find

$$P(|Y - \rho X| > 0.5 \,|\, X = 2.5).$$

But since $Y - \rho X = \sqrt{1-\rho^2}Z$ is independent of X with normal $(0, 1-\rho^2)$ distribution, where

$$\sqrt{1-\rho^2} = \sqrt{0.75} \approx 0.87,$$

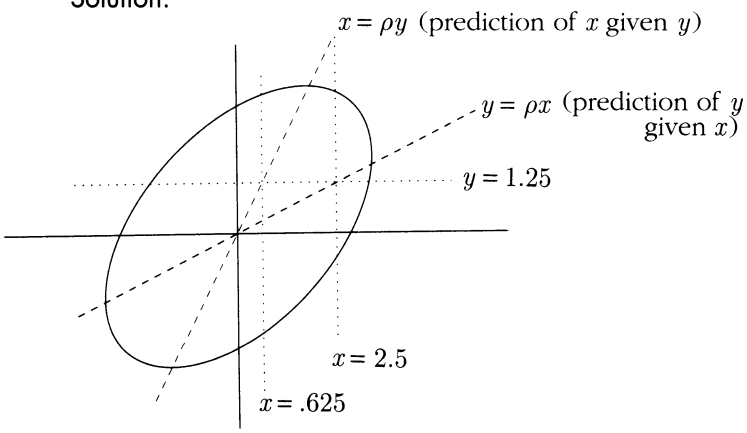
this is the same as

$$\begin{aligned}P(|Y - \rho X| > 0.5) &= P(\sqrt{1-\rho^2}|Z| > 0.5) \\ &= P(|Z| > 0.5/\sqrt{1-\rho^2}) \\ &= P(|Z| > 0.5/0.87) \\ &= 2[1 - \Phi(0.5/0.87)] \approx 2[1 - \Phi(0.57)] \approx 0.57\end{aligned}$$

So with about 57% chance, the prediction will be off by more than an inch.

Problem 3. Estimate the height of a father whose son is $6'0.5''$ tall.

Solution.



From above, $6'0.5''$ is the mean height of sons of $6'2''$ fathers. So you might guess that $6'2''$ was the mean height of fathers of $6'0.5''$ sons. But this is wrong, because a given father's height corresponds to a vertical slice through the scatter, whereas a given son's height corresponds to a horizontal slice, which is something quite different. See diagrams. The roles of X and Y must simply be switched in the calculation of Problem 1. The son's height of $6'0.5''$ is 1.25 in standard units. So

$$\begin{aligned}\text{estimated father's height} &= 0.5 \times 1.25 \quad \text{in standard units} \\ &= 0.625 \text{ in standard units} \\ &= 5'9'' + 0.625 \times 2'' = 5'10.25''\end{aligned}$$

Example 2. The probability that both variables are above average.

Problem 1. For the data in Example 1, what fraction of father–son pairs have both father and son of above average height?

Solution. Expressed in terms of the standardized variables X and Y, the problem is to find $P(X \geq 0, Y \geq 0)$. In principle, the answer can be computed as a double integral

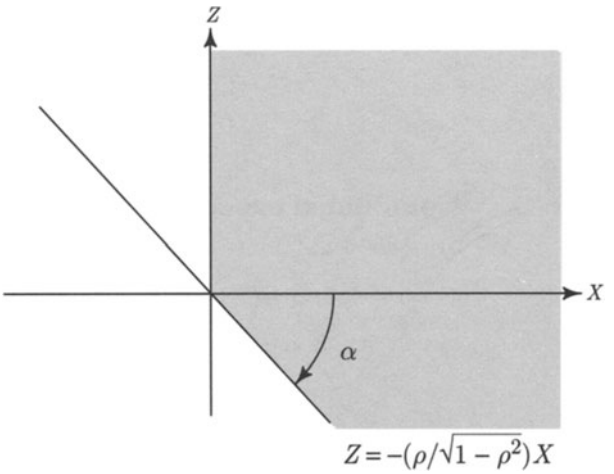
$$\iint_{\text{positive quadrant}} f(x, y)\, dx\, dy$$

where $f(x, y)$ is the standard bivariate normal density with $\rho = 0.5$. But, as usual, it is easier to first express X and Y in terms of independent standard normal variables X and Z:

$$Y = \rho X + \sqrt{1 - \rho^2} Z$$

Now the problem is to find

$$\begin{aligned} P(X \geq 0, Y \geq 0) &= P(X \geq 0, \rho X + \sqrt{1 - \rho^2} Z \geq 0) \\ &= P\left(X \geq 0, Z \geq \frac{-\rho}{\sqrt{1 - \rho^2}} X\right) \end{aligned}$$



The diagram shows the (X, Z) plane, with the line $Z = -\rho/\sqrt{1 - \rho^2} X$. The shaded region corresponds to the event above. The slope of the line is $-\rho/\sqrt{1 - \rho^2}$.

So for α as in the diagram, considered a negative angle,

$$\tan\alpha = \frac{-\rho}{\sqrt{1-\rho^2}}$$
$$= \frac{-0.5}{\sqrt{0.75}} = -1/\sqrt{3}$$

So $\alpha = -30°$. Thus the angle at the corner of the shaded region is $-\alpha + 90° = 120°$. By rotational symmetry, the chance that (X, Z) lies in the shaded region is the ratio of angles $120°/360° = 1/3$. So

$$P(X \geq 0, Y \geq 0) = 1/3$$

In other words, about one-third of the father–son pairs had both father and son above average height.

Problem 2. Suppose you have data on two variables with a bivariate normal distribution, and 3/8 of the data is above average in both variables. Estimate ρ.

Solution. Transform to standard units and use the same linear change of variable as in the solution of the previous problem. Now

$$\frac{3}{8} = \frac{135°}{360°}$$

so the angle of the corner at the origin is 135°. Thus α in the diagram is $-45°$, and by the previous solution

$$\frac{-\rho}{\sqrt{1-\rho^2}} = \tan\alpha = \tan(-45°) = -1$$

So $\rho = 1/\sqrt{2}$.

Example 3. Conditional expectation of Y given X in an interval.

Suppose (X, Y) has standard bivariate normal density with correlation ρ.

Problem. For $a < b$, find $E(Y \mid a < X < b)$.

Solution. Given that X has a particular value $x \in (a, b)$, the expected value of Y is

$$E(Y \mid X = x) = \rho x.$$

Given just $(a < X < b)$ the precise value of X is unknown. But by the rule of average conditional expectations, $E(Y \mid a < X < b)$ can be found by integration of the conditional expectation $E(Y \mid X = x) = \rho x$ with respect to the conditional density of X given $a < X < b$. This gives

$$E(Y \mid a < X < b) = \int_a^b \rho x f_X(x \mid a < X < b)\, dx$$

where for $a < x < b$

$$\begin{aligned} f_X(x \mid a < X < b)\, dx &= P(X \in dx \mid a < X < b) \\ &= \frac{P(X \in dx, a < X < b)}{P(a < X < b)} \\ &= \frac{P(X \in dx)}{P(a < X < b)} \\ &= \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}}{\Phi(b) - \Phi(a)}\, dx \end{aligned}$$

Substituting this expression gives

$$E(Y \mid a < X < b) = \int_a^b \rho x \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}}{\Phi(b) - \Phi(a)}\, dx = \frac{\frac{\rho}{\sqrt{2\pi}} \left[e^{-\frac{1}{2}a^2} - e^{-\frac{1}{2}b^2} \right]}{\Phi(b) - \Phi(a)}$$

Example 4. Midterm and final.

Midterm and final scores in a large class have an approximately bivariate normal distribution, with parameters

midterm scores:	mean: 65	SD: 18
final scores:	mean: 60	SD: 20
	correlation: 0.75	

Problem. Estimate the average final score of students who were above average on the midterm.

Solution. Let X and Y denote the midterm and final scores in standard units. The event "midterm score above average" is the same as the event $X > 0$. Take $a = 0$ and $b = \infty$ in the previous example to get

$$E(Y|X > 0) = \frac{\rho}{\sqrt{2\pi}} \left[\frac{1 - 0}{0.5} \right] = \frac{0.75 \times 2}{\sqrt{2\pi}} \approx 0.6$$

So the average final score of those who scored above average on the midterm is 0.6, in standard units. Thus the required score is

$$60 + 20 \times 0.6 = 72$$

Linear Combinations of Several Independent normal variables

The standard bivariate normal distribution was defined as the joint distribution of a particular pair of linear combinations of independent standard normal variables X and Z, namely, X and $\rho X + \sqrt{1-\rho^2}Z$. While this representation seems at first artificial, the examples show how it is the basis of all calculations involving the more general bivariate normal distribution, which is obtained by allowing arbitrary means and variances, but insisting that the two standardized variables are standard bivariate normal.

The rotational symmetry of the joint distribution of two independent standard normal variables Z_1 and Z_2 implies that the joint distribution of any two linear combinations of Z_1 and Z_2, say

$$V = a_1 Z_1 + a_2 Z_2 \qquad \text{and} \qquad W = a_1 Z_1 + a_2 Z_2$$

is bivariate normal. By reducing to this case by scaling, the same conclusion is obtained for any two independent normal variables Z_1 and Z_2 (not necessarily standard). It can be shown that this extends to linear combinations of any number of independent normal variables Z_i:

Two Linear Combinations of Independent Normal Variables

Let

$$V = \sum_i a_i Z_i \qquad \text{and} \qquad W = \sum_i b_i Z_i$$

be two linear combinations of independent normal (μ_i, σ_i^2) variables Z_i. Then the joint distribution of V and W is bivariate normal.

Granted this, the parameters of the bivariate normal distribution of V and W are easily computed:

$$\mu_V = \sum_i a_i \mu_i \quad \text{and} \quad \mu_W = \sum_i b_i \mu_i$$

$$\sigma_V^2 = \sum_i a_i^2 \sigma_i^2 \quad \text{and} \quad \sigma_W^2 = \sum_i b_i^2 \sigma_i^2$$

$$Cov(V, W) = \sum_i a_i b_i \sigma_i^2$$

$$\rho = Cov(V,W)/\sigma_V\sigma_W$$

Thus the bivariate normal distribution adequately describes the dependence between any two linear combinations of independent normal variables. In particular, this discussion implies the following result:

Independence of Linear Combinations

Two linear combinations $V = \sum_i a_i Z_i$ and $W = \sum_i b_i Z_i$ of independent normal(μ_i, σ_i^2) variables Z_i are independent if and only if they are uncorrelated, that is, if and only if $\sum_i a_i b_i {\sigma_i}^2 = 0$.

Just as the bivariate normal distribution is the joint distribution of two linear combinations of independent normal variables, the *multivariate normal distribution* is the joint distribution of several linear combinations of independent normal variables. It can be shown that several linear combinations of independent normal variables are mutually independent if and only if the covariance between every pair of them is zero. This is a special and important property of normally distributed random variables. It makes covariance and correlation perfectly suited to the analysis of linear combinations of such variables. Keep in mind however, that in general uncorrelated random variables are not necessarily independent.

Exercises 6.5

1. Here is a summary of Pre-SAT and SAT scores of a large group of students.

PSAT scores:	average: 1200	SD: 100
SAT scores:	average: 1300	SD: 90
	correlation: 0.6	

Assume the data are approximately bivariate normal in distribution.

a) Of the students who scored 1000 on the PSAT, about what percentage scored above average on the SAT?

b) Of the students who scored below average on the PSAT, about what percentage scored above average on the SAT?

c) About what percentage of students got at least 50 points more on the SAT than on the PSAT?

2. Data from a large population indicate that the heights of mothers and daughters in this population follow the bivariate normal distribution with correlation 0.5. Both variables have mean 5 feet 4 inches, and standard deviation 2 inches. Among the daughters of above average height, what percent were shorter than their mothers?

3. Heights and weights of a large group of people follow a bivariate normal distribution, with correlation 0.75. Of the people in the 90th percentile of weights, about what percentage are above the 90th percentile of heights?

4. Suppose X and Y are standard normal variables. Find an expression for $P(X+2Y \leq 3)$ in terms of the standard normal distribution function Φ,

a) in case X and Y are independent;

b) in case X and Y have bivariate normal distribution with correlation $1/2$.

5. Let X and Y have bivariate normal distribution with parameters μ_X, μ_Y, ${\sigma_X}^2$, ${\sigma_Y}^2$, and ρ. Let $P(X > \mu_X, Y > \mu_Y) = q$. Find:

a) a formula for q in terms of ρ; b) a formula for ρ in terms of q.

6. Let X and Y be independent standard normal variables.

a) For a constant k, find $P(X > kY)$.

b) If $U = \sqrt{3}X + Y$, and $V = X - \sqrt{3}Y$, find $P(U > kV)$.

c) Find $P(U^2 + V^2 < 1)$.

d) Find the conditional distribution of X given $V = v$.

7. Let X and Y have bivariate normal distribution with parameters μ_X, μ_Y, ${\sigma_X}^2$, ${\sigma_Y}^2$, and ρ.

a) Show that X and Y are independent if and only if they are uncorrelated.

b) Find $E(Y|X = x)$. c) Find $Var(Y|X = x)$.

d) Show that for constants a, b, and c, $aX + bY + c$ has a normal distribution. Find its mean and variance in terms of the parameters of X and Y.

e) Show that if $\mu_X = \mu_Y = 0$, then $X\cos\theta + Y\sin\theta$ and $-X\sin\theta + Y\cos\theta$ are independent normal variables, where

$$\theta = \frac{1}{2}\cot^{-1}\left[\frac{{\sigma_X}^2 - {\sigma_Y}^2}{2\rho\sigma_X\sigma_Y}\right]$$

Explain the geometric significance of θ in terms of the axes of an ellipse of constant density for (X, Y).

8. Let X_1 and X_2 be two independent standard normal random variables. Define two new random variables as follows: $Y_1 = X_1 + X_2$ and $Y_2 = \alpha X_1 + 2X_2$. You are not given the constant α but it is known that $Cov\{Y_1, Y_2\} = 0$. Find

a) the density of Y_2; b) $Cov\{X_2, Y_2\}$.

9. Suppose that W has normal (μ, σ^2) distribution. Given that $W = w$, suppose that Z has normal $(a\,w + b, \tau^2)$ distribution.

a) Show the joint distribution of W and Z is bivariate normal, and find its parameters.

b) What is the distribution of Z?

c) What is the conditional distribution of W given $Z = z$?

10. Show that if V and W have a bivariate normal distribution then

a) every linear combination $aV + bW$ has a normal distribution;

b) every pair of linear combinations $(aV + bW,\ cV + dW)$ has a bivariate normal distribution.

c) Find the parameters of the distributions obtained in a) and b) in terms of the parameters of the joint distribution of V and W.

11. Show that for standard bivariate normal variables X and Y with correlation ρ,

$$E(\max(X, Y)) = \sqrt{\frac{1-\rho}{\pi}}$$

12. Suppose that the magnitude of a signal received from a satellite is

$$S = a + bV + W$$

where V is a voltage which the satellite is measuring, a and b are constants, and W is a noise term. Suppose V and W are independent and normally distributed with means 0 and variances ${\sigma_V}^2$ and ${\sigma_W}^2$.

a) Find $Corr(S, V)$.

b) Given that $S = s$, what is the distribution of V?

c) What is the best estimate of V given $S = s$?

d) If this estimate is used repeatedly for different values of S coming from a sequence of independent values of V and W with the given normal distributions, what is the long-run average absolute value of the error of estimation?

13. Find a formula in terms of ρ for the ratio of the lengths of the axes of an ellipse of constant density in the standard bivariate normal distribution with correlation ρ. (Let the ratio be the length of the axis at $+45°$ over the length of the axis at $-45°$.)

Check your answer by measurement with a ruler in Figure 3 in the case where $\rho = 1/2$.

[*Hint*: Let $\rho = \cos\theta$ and reason from Figure 3 that an ellipse of constant density is the image in the (X, Y) plane of the unit circle in the (X, Z) plane. Now consider the images of the points $(\cos\theta/2, \sin\theta/2)$ and $(\cos(\theta/2 + \pi/2), \sin(\theta/2 + \pi/2))$ in the (X, Y) plane which end up on the $\pm 45°$ lines in the (X, Z) plane, and use trigonometric identities.]

Dependence: Summary

Conditional Distributions: Let X be a discrete random variable. The conditional probability of an event A given $X = x$ is

$$P(A|X = x) = \frac{P(A, X = x)}{P(X = x)}$$

by the division rule of Section 1.4.

For continuously distributed X, there is instead the *infinitesimal conditioning formula*

$$P(A|X = x) = \frac{P(A, X \in dx)}{P(X \in dx)}$$

Understand $P(A|X = x)$ as the chance of A given that X falls in a very small interval near x.

If X and Y are discrete random variables, the conditional probability of $Y = y$ given $X = x$ is

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)}$$

If X and Y are continuous random variables with joint density $f_{X,Y}$, the conditional density of Y at y given $X = x$ is $f_Y(y|X = x)$ where

$$f_Y(y|X = x)dy = P(Y \in dy|X \in dx) = \frac{f_{X,Y}(x, y)dx\,dy}{f_X(x)dx} = \frac{f_{X,Y}(x, y)}{f_X(x)}dy$$

Once you have conditioned on $X = x$, you can treat the random variable X as the constant x. Conditional distributions given $X = x$ behave exactly like ordinary distributions, with the constant x as a parameter.

Conditional expectation: The *conditional expectation of Y given $X = x$*, denoted $E(Y|X = x)$, is defined as the expectation of Y relative to the conditional distribution of Y given $X = x$.

The *conditional expectation of Y given X*, denoted $E(Y|X)$, is a random variable, whose value is $E(Y|X = x)$ if $(X = x)$. Thus the random variable $E(Y|X)$ is a function of the random variable X, namely, $f(X)$, where $f(x) = E(Y|X = x)$ for every x.

Expectation is the expectation of conditional expectation: $E(Y) = E[E(Y|X)]$.

See boxes on pages 424 and 425 for important properties of conditional distributions and expectations, and a comparison of the discrete and continuous cases.

Independence: Random variables X and Y are independent if and only if for all subsets B in the range of Y, and all x

$$P(Y \in B|X = x) = P(Y \in B)$$

That is, the conditional distribution of Y given $X = x$ does not depend on x.

Equivalently, X and Y are independent if the conditional distribution of X given $Y = y$ does not depend on y.

Covariance and correlation: $Cov(X,Y) = E\Big[[X-E(X)][Y-E(Y)]\Big] = E(XY)-E(X)E(Y)$

$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$

$$Corr(X,Y) = \frac{Cov(X,Y)}{SD(X)SD(Y)} \in [-1,1]$$

X and Y independent $\Longrightarrow Corr(X,Y) = 0$ *but not conversely.*

X and Y uncorrelated $\Longleftrightarrow Corr(X,Y) = 0$

$\Longleftrightarrow Cov(X,Y) = 0 \Longleftrightarrow E(XY) = E(X)E(Y).$

Bivariate normal: X and Y have standard bivariate normal distribution with correlation ρ if and only if

$$Y = \rho X + \sqrt{1-\rho^2}Z,$$

where X and Z are independent standard normal variables.

Marginals. Both X and Y have standard normal distribution.

Conditionals.

Given $X = x$, Y has normal $(\rho x, 1-\rho^2)$ distribution.

Given $Y = y$, X has normal $(\rho y, 1-\rho^2)$ distribution.

X and Y have bivariate normal distribution with parameters $\mu_X, \mu_Y, {\sigma_X}^2, {\sigma_Y}^2$, and ρ if and only if the standardized variables $X^* = (X-\mu_X)/\sigma_X$ and $Y^* = (Y-\mu_Y)/\sigma_Y$ have standard bivariate normal distribution with correlation ρ. Conditional distributions in this case are derived from the standardized case by a linear change of variable. All probabilities and expectations for bivariate normal variables are found by a linear change of variable to independent standard normal variables.

Independence. X and Y with bivariate normal distribution are independent if and only if they are uncorrelated.

Review Exercises

1. Let X and Y be independent random variables. Suppose X has Poisson distribution with parameter λ_1, and Y has Poisson distribution with parameter λ_2.

a) Given that $X + Y = 100$, what are the possible values of X?

b) For each possible value k, find $P(X = k \mid X + Y = 100)$.

c) Take $\lambda_1 = 1$ and $\lambda_2 = 99$. Given $X + Y = 100$, estimate the chance that X is 4 or 5 or 6.

2. Let N denote the number of children in a randomly picked family. Suppose N has geometric distribution:

$$P(N = n) = (1/3)(2/3)^{n-1} \qquad (n = 1, 2, 3, \ldots)$$

and suppose each child is equally likely to be male or female. Let X be the number of male children and Y the number of female children, in a randomly picked family:

a) Find the joint distribution of (X, Y).

b) Given $Y = 0$, what is the most likely value of X?

c) What is the conditional expectation of X given $Y = 0$?

3. A list of $2n$ numbers has mean μ and variance σ^2. Suppose that n numbers are picked at random from the list. Let A_n be the average of these n numbers, B_n the average of the other numbers. Find: a) $E(A_n - B_n)$; b) $SD(A_n - B_n)$.

4. Suppose X and Y have joint density function

$$f(x, y) = \begin{cases} c/x^3 & x > y > 1 \\ 0 & \text{otherwise} \end{cases}$$

where c is a constant.

a) Find c. b) Find the marginal density of X.

c) What is the conditional distribution of Y given $X = x$?

5. Suppose X and Y are random variables with joint density in the plane $f(x, y) = ce^{-(x^2+xy+y^2)}$ where c is a constant. a) Find c. b) Find $Corr(X, Y)$.

6. Let X and Y be independent exponential random variables each with mean 1. Find
a) the joint density of $X + Y$ and $X - Y$;
b) $Corr(X + Y, X - Y)$.

7. Suppose that a point (X, Y) is chosen according to the uniform distribution on the triangle with vertices $(0, 0), (0, 1), (1, 0)$. Calculate:

a) the mean and variance of X;

b) the conditional mean and variance of X given that $Y = 1/3$;

c) the mean and variance of $\max(X, Y)$;

d) the mean and variance of $\min(X, Y)$.

8. Let Y have exponential distribution with mean 0.5. Let X be such that, conditional on $Y = y$, X has exponential distribution with mean y. Find:

(a) the joint density of (X, Y); b) $E(X)$; c) $Corr(X, Y)$.

9. Let X, Y, and Z be independent uniform $(0, 1)$ variables. Find $P[(X/Y) > (Y/Z)]$.

10. Let T_A, T_B, and T_C be the failure times of components A, B, and C. Assume these are independent exponential random variables with rates α, β, and γ, respectively.

a) What is the distribution of the time until the first failure?

b) What is the probability that the first component to fail is component C?

c) Given that the first component to fail is component C, what is the distribution of the time between the first and second failures?

d) Write a formula for the (unconditional) c.d.f. of the time between the first and second failures.

11. Insurance claims arrive at an insurance company according to a Poisson process with rate λ. The amount of each claim has exponential distribution with rate μ, independently of times and amounts of all other claims. Let X_t denote the accumulated total of claims between time 0 and time t. Find simple formulae for

a) $E(X_t)$; $E(X_t^2)$; c) $SD(X_t)$; d) $Corr(X_s, X_t)$ for $s \leq t$.

12. An elevator has an occupancy warning of no more than 26 people and of total weight no more than 4000 pounds. For the population of users, suppose weights are approximately normal with mean 150 pounds and standard deviation 30 pounds.

a) What is the probability that the total weight of a random sample of 26 people from the population exceeds 4000 pounds?

b) Suppose next that the people are carrying things and that the weight of these for an individual of weight X pounds, is approximately normal with mean $0.05X$ pounds and standard deviation 2 pounds. What is the probability that the total weight in the elevator now exceeds 4000 pounds?

c) The dimensions of the floor of the elevator are 54 inches by 92 inches. Suppose the amount of floor space needed by users is normally distributed with mean μ square inches and standard deviation 0.1μ. Find μ such that the probability 20 people can be accommodated is 0.99.

13. a) Let X and Y be two random variables with finite and nonzero variances. Show that $X - Y$ and $X + Y$ are uncorrelated if and only if $Var(X) = Var(Y)$.

b) Let X and Y have standard bivariate normal distribution with correlation 0.6. Find $P(X - Y < 1, X + Y > 2)$.

14. Heights. A population consists of 50% men and 50% women. The empirical distribution of heights over the population yields the following statistics:

	Average	Standard deviation
Men's heights	67 inches	3 inches
Women's heights	63 inches	3 inches

a) What is the average height over the whole population?

b) What is the standard deviation of heights over the whole population?

c) Suppose that men's heights are approximately normally distributed, and that women's heights are as well. Calculate the approximate proportion of individuals in the whole population with heights between 63 and 67 inches.

d) Repeat c), assuming instead that heights are normally distributed over the whole population. Explain why the answers are slightly different.

e) Suppose that a man and a woman are picked at random from this population. Making assumptions as in c), what is the probability that the man is taller than the woman? [*Hint:* No integration required!]

15. Sums of normals in the positive quadrant. Let X and Y be two independent standard normal variables.

a) Calculate $P(X \geq 0, Y \geq 0, X + Y \leq 1)$.

b) Find the conditional density of $X + Y$ given $X \geq 0$ and $Y \geq 0$, and sketch its graph.

c) Find, approximately, the median and the mode of this distribution.

16. Rainfall. Suppose that the distribution of annual rainfall in a particular place, measured in inches, is approximately gamma with shape parameter $r = 3$. If the mean annual rainfall is 20 inches, find approximations to the following:

a) the probability of more than 35 inches of rain in any particular year;

b) the probability that in ten consecutive years, it never rains more than 35 inches, assuming different years are independent;

c) still assuming independence of different years, the probability that the record rainfall over the last 20 years is exceeded in at least one of the next ten years, assuming the record rainfall over the last 20 years, R_{20} say, is known;

d) same as c), but assuming the value of R_{20} is unknown.

17. Symmetry under rotations.

a) Suppose the joint distribution of X and Y is symmetric under rotations. Are X and Y necessarily independent? Are they necessarily uncorrelated? Explain by arguments or examples.

b) Suppose (X, Y) is a point picked at random from the unit circle $X^2 + Y^2 = 1$. Calculate $E(X^2)$, $E(Y^2)$, and $E(XY)$.

c) Suppose U is uniformly distributed on $(0, 1)$, $X = \cos 2\pi U, Y = \sin 2\pi U$. Are X and Y uncorrelated? Are X and Y independent? Explain carefully the connection betweeen b) and c).

18. Maxima and minima of normal variables.
Calculate the expected values of $\max(X, Y)$ and $\min(X, Y)$:

a) if X and Y are independent standard normal variables;

b) if X and Y are independent normal (μ, σ^2);

c) if X and Y are standard bivariate normal with correlation ρ.

19. Suppose you sample with replacement n times from a population of n elements.

a) What fraction of the n elements should you expect to see in the sample?

b) For example, what fraction of all $\binom{52}{5}$ poker hands should you expect to see in $\binom{52}{5}$ independent deals?

c) Compute the variance of the fraction in a), and show that it is less than $1/4n$.

d) Evaluate for the example in b), and estimate the chance that your prediction in b) is off by more than 1%.

20. Craps. Find the expectated total number of times Y the pair of dice must be rolled in a craps game (see Exercise 3.4.8) by conditioning on the result of the first roll.

21. I toss a coin which lands heads with probability p. Let W_H be the number of tosses till I get a head, W_{HH} the number of tosses till I get two heads in a row, and W_{HHH} the number of tosses till I get three heads in a row. Find:

a) $E(W_H)$; b) $E(W_{HH})$ [*Hint*: condition on whether the first toss was heads or tails]; c) $E(W_{HHH})$ [*Hint*: condition on W_T].

d) Generalize to find the expected number of tosses to obtain m heads in a row.

22. Long runs of heads. In the play *Rosencrantz and Guildenstern are dead* by Tom Stoppard, the results of 101 apparently fair coin tosses are recorded: 100 heads in a row, followed by a tail. Suppose a fair coin is tossed independently once every second. About how many years do you expect it would take before 100 heads in a row came up? How long for it to be 99% sure that such a run will have appeared?

23. Suppose an insect lays a Poisson (λ) number of eggs. Suppose each egg hatches with probability p and dies with probability q, independently of each other egg. Show that the number of eggs that hatch and the number of eggs that die are independent Poisson random variables, and find their parameters.

24. I roll a random number of dice. If the number of dice rolled has the Poisson (12) distribution, find (and justify your answers)

a) the expectation of the total number of spots showing;

b) the standard deviation of the total number of spots showing.

25. Suppose the number of accidents in an interval of time has Poisson (λ) distribution. Suppose that in each accident there are k persons injured with probability p_k, independently of all other accidents. Let N_k be the number of accidents in which k persons are injured.

a) What is the joint distribution of N_1 and N_2?

b) Let M be the total number of persons injured. Find formulae for $E(M)$ and $SD(M)$ in terms of $p_1, p_2, \ldots$ and λ.

26. Distinguishing points in a Poisson scatter. In practical situations, if two points in a scatter are closer than some distance δ, it may not be possible to distinguish them. Suppose that this is the case, and that there is a Poisson scatter over the unit square, with intensity λ. Show that the probability of the event D, that all points in the scatter can be distinguished, is at least $1 - \frac{\pi}{2}\lambda^2\delta^2$.

[*Hint.* Show that $P(D|N=2) \geq 1-\pi\delta^2$ and $P(D|N=3) \geq (1-\pi\delta^2)(1-2\pi\delta^2)$ and so on. Use the inequality

$$(1-\alpha)(1-\beta) \geq 1-(\alpha+\beta) \qquad (\alpha>0,\ \beta>0)$$

repeatedly, to obtain

$$P(D|N) \geq 1-\frac{1}{2}N(N-1)\pi\delta^2]$$

27. Inhomogeneous Poisson scatter. Let Q be a probability distribution over a set S, $\lambda>0$. Consider a random scatter of points over the set S, where a Poisson (λ) number N of points are distributed independently at random according to Q. More formally, for B a subset of S, let $N(B)=0$ if $N=0$, and

$$N(B)=\sum_{i=1}^{n} I(X_i \in B) \qquad \text{if} \quad (N=n), \quad n=1,2,\ldots$$

where $X_1,\ldots,X_n$ are conditionally independent with common distribution Q given $(N=n)$, and N has Poisson (λ) distribution. Prove that
for disjoint $B_1,\ldots,B_j$, the $N(B_1),\ldots,N(B_j)$ are mutually independent Poisson random variables with parameters $\lambda(B_1),\ldots,\lambda(B_j)$ where $\lambda(B)=\lambda Q(B)$.
[*Hint*: Start by considering the case of B_1 and B_2 with $B_1+B_2=S$, and calculate $P(N(B_1)=n_1, N(B_2)=n_2)$ by conditioning on $N=n_1+n_2$. Argue that, in general, it suffices to consider a partition $B_1,\ldots,B_j$ of S, and proceed similarly. The multinomial coefficients $n!/(n_1!n_2!\cdots n_j!)$ should appear.]
Note. Such a collection of random variables $N(B)$ is called a *Poisson process with intensity measure $\lambda(B)$ on S*. For S the unit square and $\lambda(B)=\lambda\times Area(B)$ this is a construction of the Poisson scatter over the unit square considered in Section 3.5. Such a scatter is called *homogeneous*. If $Q(B)$ is not the uniform distribution, the scatter is called *inhomogeneous*. Note that if $Q\{s\}>0$ for a point $s\in S$, there may be more than one "hit" counted at s. In particular, if Q is a discrete measure with probabilities $q_1,\ldots,q_n$ at points $s_1,\ldots,s_n$, then $N(s_1),\ldots,N(s_n)$ are independent Poisson random variables with parameters $\lambda q_1,\ldots,\lambda q_n$.
Illustration. Suppose you roll a Poisson (λ) number N of dice. Then the number of times each of the six faces appears is an independent Poisson $(\lambda/6)$ random variable. And the number of odd faces and the number of even faces are two independent Poisson $(\lambda/2)$ random variables. But if you throw a fixed number n of dice these numbers are dependent, because they must add up to n.

28. You and I both toss a fair coin N times. You get X heads and I get Y heads.

a) If $P(X=Y)$ is approximately 10%, then approximately how large must N be?

b) The normal approximation says $P(\left|X-\frac{1}{2}N\right| \leq \frac{1}{2}\sqrt{N}) \approx 68\%$.

Given $X=Y$, is the conditional probability that $\left|X-\frac{1}{2}N\right| \leq \frac{1}{2}\sqrt{N}$ still about 68%, somewhat larger than 68%, or somewhat smaller than 68%? Explain which, *without* doing detailed calculations.

29. Variance of discrete order statistics. Let T_i be the place at which the ith good element appears in a random ordering of k good and $N-k$ bad elements. From Exercise 3.6.13, the mean of T_i is $E(T_i)=i(N+1)/(k+1)$. Calculate $SD(T_i)$ by the following steps.

a) Let $\alpha(k,N) = E\left(T_1(T_1-1)\right)$, $1 \le k \le N$. Show by conditioning on whether the first element is good or bad that

$$\alpha(k,N) = (N-k)\left[\frac{2}{k+1} + \frac{\alpha(k,N-1)}{N}\right]$$

b) Deduce that

$$\alpha(k,N) = \frac{2(N+1)(N-k)}{(k+1)(k+2)}$$

c) Deduce that

$$Var(T_1) = \frac{(N+1)(N-k)k}{(k+1)^2(k+2)}$$

d) Check the case $k=1$ by calculating $Var(T_1)$ directly from the distribution of T_1.

e) Let $W_i = T_{i+1} - T_i$, $\quad i = 1,\ldots,k+1$, $\quad$ where $T_0 = 0$ and $T_{k+1} = N+1$. Use the exchangeability of $W_1,\ldots,W_{k+1}$ to show that for each $i = 1,\ldots,k+1$

$$Var(T_i) = i\,Var(T_1) + i(i-1)\,Cov(W_1,W_2)$$

Deduce that

$$Cov(W_1,W_2) = -Var(T_1)/k$$

and hence that

$$Var(T_i) = \frac{i(k+1-i)(N+1)(N-k)}{(k+1)^2(k+2)}$$

f) Give an intuitive explanation of why $SD(T_i) = SD(T_{k+1-i})$.

g) Suppose that $T_1,\ldots,T_4$ are the places at which the aces appear in a well-shuffled deck of 52 cards. Find numerical values of $E(T_i)$ and $SD(T_i)$ for $i = 1,\ldots,4$.

30. Let $V_1,\ldots,V_n$ be the order statistics of n independent uniform $(0,1)$ variables. Let

$A_{\text{all}} = (V_1 + \cdots + V_n)/n$, $\quad$ average of all the order statistics,
$A_{\text{ext}} = (V_1 + V_n)/2$, $\quad$ average of the extremes,
$A_{\text{mid}} = V_{(n+1)/2}$, $\quad$ the middle value, where you can assume n is odd.

a) Show that for large n, each of the A's is most likely very close to $1/2$.

b) For large n, one of the A's can be expected to be very much closer to $1/2$ than the two others. Which one, and why?

c) For $n = 101$ find for each of the A's a good approximation to the probability that it is between .49 and .51.

31. From discrete to continuous spacings. Let $U_{(1)} < U_{(2)} < \ldots < U_{(n)}$ be the order statistics of n independent uniform $(0,1)$ variables $U_1,\ldots,U_n$. Let $V_1 = U_{(1)}$, $V_i = U_{(i)} - U_{(i-1)}$ for $1 \le i \le n$, and let $V_{n+1} = 1 - U_{(n)}$. Imagine the unit interval is cut into subintervals at each of the n random points U_i for $1 \le i \le n$. Then $V_1, V_2, \ldots, V_{n+1}$ are the lengths of the $n+1$ subintervals so obtained, in order from left to right. This model for cutting an interval at random is of interest in genetics. The V_i could represent the relative lengths of strands obtained by random cutting of a long molecule such as DNA. For a positive integer $N > n$ let $U'_1,\ldots,U'_{N-n}$ denote $N-n$ more uniform $(0,1)$ variables, independent of each other and of the cut points $U_1,\ldots,U_n$. For $1 \le i \le n+1$ let N_i denote the number of U'_i that fall in the interval $(U_{(i-1)}, U_{(i)})$ of length V_i (where $U_{(0)} = 0$ and $U_{(n+1)} = 1$ to make the definition work for N_1 and N_{n+1}).

a) Show that the joint distribution of $N_1, \ldots, N_{n+1}$ is identical to the joint distribution of the discrete spacings $W_1, \ldots, W_{n+1}$ derived from a random ordering of n aces and $N-n$ nonaces as in Exercise 3.6.13. That is to say, $(N_1, \ldots, N_{n+1})$ has uniform distribution over the set of all $(n+1)-$tuples of non-negative integers $(n_1, \ldots, n_{n+1})$ with $n_1 + \cdots + n_{n+1} = N - n$. In particular, $N_1, \ldots, N_{n+1}$ are exchangeable.

b) Conditionally given the continuous spacings $(V_1, \ldots, V_{n+1})$, the sequence $(N_1, \ldots, N_{n+1})$ is distributed like the number of results in each of $n+1$ categories in a sequence of $N-n$ independent trials with probability V_i of a result in category i on each trial. Explain why this is so. Deduce that for large N, N_i/N is almost equal to V_i for each i with overwhelmingly high probability. It follows that in the limit as $N \to \infty$ for fixed n, as discussed at the end of Exercise 3.6.12, the joint distribution of the normalized discrete spacings $(N_1/N, N_2/N, \ldots, N_{n+1}/N)$ converges to the joint distribution of the continuous spacings $V_1, V_2, \ldots, V_{n+1}$.

(Keep in mind that the distribution of N_i depends on N, so N_i/N does not just tend to zero: the sum over i of the N_i/N is identically equal to 1.) Since the N_i/N are exchangeable for every N, it follows that the V_i are exchangeable, something that is not obvious in the continuous model.

32. Joint distribution of continuous spacings. Continuing with the same notation as in Exercise 31,

a) Show that for $v_i \geq 0$ with $v_1 + \cdots + v_{n+1} = v \leq 1$

$$\lim_{N \to \infty} P(N_i/N \geq v_i \text{ for every } 1 \leq i \leq n+1) = (1-v)^n$$

by explicit evaluation of the limit, using Exercise 3.6.15 and the fact that $(N)_k \sim N^k$ as $N \to \infty$ for every $k = 1, 2, \ldots$. This yields the corresponding probabilty for the continuous model: for $v_i \geq 0$ with $v_1 + \cdots + v_{n+1} = v \leq 1$

$$P(V_i \geq v_i \text{ for every } 1 \leq i \leq n+1) = (1-v)^n$$

b) Show that the V_i have identical distribution with

$$P(V_i \geq v) = (1-v)^n \qquad (0 \leq v \leq 1)$$

c) Deduce that V_i has beta $(1, n)$ distribution.

33. Maximum and minimum spacings. Continuing with the notation of the preceding exercises, let $V_{\min} = \min_i V_i$ where the min is over $1 \leq i \leq n+1$.

a) Show that $V_{\min}$ has the same distribution as $V_1/(n+1)$. Deduce the mean and variance of $V_{\min}$ from the mean and variance of the beta $(1, n)$ distribution.

b) Let $V_{\max} = \max_i V_i$. Parallel to the discrete formula of Exercise 3.6.16, show that for $0 \leq v \leq 1$

$$P(V_{\max} \geq v) = \sum_{i=1}^{n+1} (-1)^{i-1} \binom{n+1}{i} (1-iv)_+^n$$

where $(1-iv)_+^n$ equals $(1-iv)^n$ if $iv \leq 1$, and equals 0 otherwise.

c) Deduce by integration of this tail probability from 0 to 1 that

$$E(V_{\max}) = \sum_{i=1}^{n+1} (-1)^{i-1} \binom{n+1}{i} \frac{1}{i(n+1)}$$

It is intuitively clear, and can be verified analytically, that as the number of cuts $n \to \infty$, $V_{\max} \to 0$, which forces the distribution of $V_{\max}$ to pile up around zero. But the rate of convergence is rather slow.

d) Find the numerical values of $E(V_{\min})$ and $E(V_{\max})$ for $n = 1, \ldots, 10$.

34. Dirichlet distribution. A sequence of random variables $(Q_1, \ldots, Q_m)$ has *Dirichlet distribution* with parameters $(r_1, \ldots, r_m)$ if $Q_i \geq 0$, $Q_1 + \cdots + Q_m = 1$, and

$$\frac{P(Q_i \in dq_i, 1 \leq i \leq m-1)}{dq_1 dq_2 \cdots dq_{m-1}} = \frac{\Gamma(r_1 + \cdots + r_m)}{\Gamma(r_1) \cdots \Gamma(r_m)} \prod_{i=1}^{m} q_i^{r_i - 1} \qquad (q_i \geq 0,\ q_1 + \cdots + q_m = 1)$$

For $m = 2$, (Q_1, Q_2) has Dirichlet distribution with parameters r and s if and only if $Q_2 = 1 - Q_1$ for Q_1 with beta (r, s) distribution. So the Dirichlet distribution is a multivariate extension of the beta distribution. There is a multivariate version of the result of Exercise 5.4.19: If $Y_i, 1 \leq i \leq m$ are independent with gamma (r_i, λ) distributions, $\sum = \sum_i Y_i$ and $Q_i = Y_i / \sum$, then $(Q_1, \ldots, Q_m)$ has Dirichlet distribution with parameters $(r_1, \ldots, r_m)$, independently of $\sum$, which has gamma (r, λ) distribution for $r = r_1 + \cdots + r_m$. Assuming this result, deduce the following properties of this Dirichlet distribution of $(Q_1, \ldots, Q_m)$:

a) The marginal distribution of Q_i is beta $(r_i, r - r_i)$.

b) For $i \neq j$ the distribution of $Q_i + Q_j$ is beta $(r_i + r_j, r - r_i - r_j)$. Similarly for any finite sum of at most $m - 1$ different Q_i.

c) The joint distribution of the continuous spacings $V_1, \cdots, V_{n+1}$ derived from n independent uniform$(0, 1)$ random variables as in Exercises 31 and 32 is Dirichlet with parameters $r_i = 1$ for $1 \leq i \leq m = n + 1$.

35. Dirichlet–multinomial. Suppose that $X_1, X_2, \ldots$ is a sequence of independent trials with m possible values $\{1, \ldots, m\}$, with probability q_i for value i on each trial. The parameters $(q_1, \ldots, q_m)$ are unknown, and regarded as the values of random variables $(Q_1, \ldots, Q_m)$. Suppose the prior distribution of $(Q_1, \ldots, Q_m)$ is Dirichlet with parameters $(r_1, \ldots, r_m)$, as in Exercise 34. After n trials, let N_i be the number of results i, that is the number of times that $X_j = i$ for $1 \leq j \leq n$. So the conditional distribution of $(N_1, \ldots, N_m)$ given $(Q_1, \ldots, Q_m)$ is multinomial with parameters n and $(Q_1, \ldots, Q_m)$.

a) Show that the posterior distribution of $(Q_1, \ldots, Q_m)$ given the results $(N_1, \ldots, N_m)$ of n trials is Dirichlet with parameters $(r_1 + N_1, \ldots, r_m + N_m)$.

b) Find a formula for the unconditional probability $P(N_i = n_i \text{ for } 1 \leq i \leq m)$ for any sequence of m non-negative integers n_i with $n_1 + \cdots + n_m = n$.
[*Hint*: Use the fact that the total integral of the Dirichlet joint density with parameters $(r_1 + n_1, \ldots, r_m + n_m)$ is 1].

c) Deduce in particular that if $r_i = 1$ for $1 \leq i \leq m$ then the unconditional distribution of $(N_1, \ldots, N_m)$ is uniform over its range of possible values.

d) Explain the result of part c) without integration by reference to Exercise 31.