

Conditioning

1 Conditional Distributions

Let A and B be events, and suppose that $P(B) > 0$. We recall from Section 3 of the Introduction that the conditional probability of A given B is defined as $P(A | B) = P(A \cap B)/P(B)$ and that $P(A | B) = P(A)$ if A and B are independent.

Now, let (X, Y) be a two-dimensional random variable whose components are discrete.

Example 1.1. A symmetric die is thrown twice. Let U_1 be a random variable denoting the number of dots on the first throw, let U_2 be a random variable denoting the number of dots on the second throw, and set $X = U_1 + U_2$ and $Y = \min\{U_1, U_2\}$.

Suppose we wish to find the distribution of Y for some given value of X , for example, $P(Y = 2 | X = 7)$.

Set $A = \{Y = 2\}$ and $B = \{X = 7\}$. From the definition of conditional probabilities we obtain

$$P(Y = 2 | X = 7) = P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{2}{36}}{\frac{1}{6}} = \frac{1}{3}. \quad \square$$

With this method one may compute $P(Y = y | X = x)$ for any fixed value of x as y varies for arbitrary, discrete, jointly distributed random variables. This leads to the following definition.

Definition 1.1. Let X and Y be discrete, jointly distributed random variables. For $P(X = x) > 0$ the conditional probability function of Y given that $X = x$ is

$$p_{Y|X=x}(y) = P(Y = y | X = x) = \frac{p_{X,Y}(x, y)}{p_X(x)},$$

and the conditional distribution function of Y given that $X = x$ is

$$F_{Y|X=x}(y) = \sum_{z \leq y} p_{Y|X=x}(z). \quad \square$$

Exercise 1.1. Show that $p_{Y|X=x}(y)$ is a probability function of a true probability distribution. \square

It follows immediately (please check) that

$$p_{Y|X=x}(y) = \frac{p_{X,Y}(x, y)}{p_X(x)} = \frac{p_{X,Y}(x, y)}{\sum_z p_{X,Y}(x, z)}$$

and that

$$F_{Y|X=x}(y) = \frac{\sum_{z \leq y} p_{X,Y}(x, z)}{p_X(x)} = \frac{\sum_{z \leq y} p_{X,Y}(x, z)}{\sum_z p_{X,Y}(x, z)}.$$

Exercise 1.2. Compute the conditional probability function $p_{Y|X=x}(y)$ and the conditional distribution function $F_{Y|X=x}(y)$ in Example 1.1. \square

Now let X and Y have a joint continuous distribution. Expressions like $P(Y = y | X = x)$ have no meaning in this case, since the probability that a fixed value is assumed equals zero. However, an examination of how the preceding conditional probabilities are computed makes the following definition very natural.

Definition 1.2. Let X and Y have a joint continuous distribution. For $f_X(x) > 0$, the conditional density function of Y given that $X = x$ is

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x, y)}{f_X(x)},$$

and the conditional distribution function of Y given that $X = x$ is

$$F_{Y|X=x}(y) = \int_{-\infty}^y f_{Y|X=x}(z) dz. \quad \square$$

In analogy with the discrete case, we further have

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x, y)}{\int_{-\infty}^{\infty} f_{X,Y}(x, z) dz}$$

and

$$F_{Y|X=x}(y) = \frac{\int_{-\infty}^y f_{X,Y}(x, z) dz}{\int_{-\infty}^{\infty} f_{X,Y}(x, z) dz}.$$

Exercise 1.3. Show that $f_{Y|X=x}(y)$ is a density function of a true probability distribution.

Exercise 1.4. Find the conditional distribution of Y given that $X = x$ in Example 1.1.1 and Exercise 1.1.3.

Exercise 1.5. Prove that if X and Y are independent then the conditional distributions and the unconditional distributions are the same. Explain why this is reasonable. \square

Remark 1.1. Definitions 1.1 and 1.2 can be extended to situations with more than two random variables. How? \square

2 Conditional Expectation and Conditional Variance

In the same vein as the concepts of expected value and variance are introduced as convenient location and dispersion measures for (ordinary) random variables or distributions, it is natural to introduce analogs to these concepts for conditional distributions. The following example shows how such notions enter naturally.

Example 2.1. A stick of length one is broken at a random point, uniformly distributed over the stick. The remaining piece is broken once more. Find the expected value and variance of the piece that now remains.

In order to solve this problem we let $X \in U(0, 1)$ be the first remaining piece. The second remaining piece Y is uniformly distributed on the interval $(0, X)$. This is to be interpreted as follows: Given that $X = x$, the random variable Y is uniformly distributed on the interval $(0, x)$:

$$Y \mid X = x \in U(0, x),$$

that is, $f_{Y|X=x}(y) = 1/x$ for $0 < y < x$ and 0, otherwise. Clearly, $E X = 1/2$ and $\text{Var } X = 1/12$. Furthermore, intuition suggests that

$$E(Y \mid X = x) = \frac{x}{2} \quad \text{and} \quad \text{Var}(Y \mid X = x) = \frac{x^2}{12}. \quad (2.1)$$

We wish to determine $E Y$ and $\text{Var } Y$ somehow with the aid of the preceding relations. \square

We are now ready to state our first definition.

Definition 2.1. Let X and Y be jointly distributed random variables. The conditional expectation of Y given that $X = x$ is

$$E(Y \mid X = x) = \begin{cases} \sum y p_{Y|X=x}(y) & \text{in the discrete case,} \\ \int_{-\infty}^y y f_{Y|X=x}(y) dy & \text{in the continuous case,} \end{cases}$$

provided the relevant sum or integral is absolutely convergent. \square

Exercise 2.1. Let X , Y , Y_1 , and Y_2 be random variables, let g be a function, and c a constant. Show that

- (a) $E(c \mid X = x) = c$,
- (b) $E(Y_1 + Y_2 \mid X = x) = E(Y_1 \mid X = x) + E(Y_2 \mid X = x)$,
- (c) $E(cY \mid X = x) = c \cdot E(Y \mid X = x)$,
- (d) $E(g(X, Y) \mid X = x) = E(g(x, Y) \mid X = x)$,
- (e) $E(Y \mid X = x) = EY$ if X and Y are independent. □

The conditional distribution of Y given that $X = x$ depends on the value of x (unless X and Y are independent). This implies that the conditional expectation $E(Y \mid X = x)$ is a function of x , that is,

$$E(Y \mid X = x) = h(x) \quad (2.2)$$

for some function h . (If X and Y are independent, then check that $h(x) = EY$, a constant.)

An object of considerable interest and importance is the random variable $h(X)$, which we denote by

$$h(X) = E(Y \mid X). \quad (2.3)$$

This random variable is of interest not only in the context of probability theory (as we shall see later) but also in statistics in connection with estimation. Loosely speaking, it turns out that if Y is a “good” estimator and X is “suitably” chosen, then $E(Y \mid X)$ is a “better” estimator. Technically, given a so-called unbiased estimator U of a parameter θ , it is possible to construct another unbiased estimator V by considering the conditional expectation of U with respect to what is called a sufficient statistic T ; that is, $V = E(U \mid T)$. The point is that $EU = EV = \theta$ (unbiasedness) and that $\text{Var } V \leq \text{Var } U$ (this follows essentially from the sufficiency and Theorem 2.3 ahead). For details, we refer to the statistics literature provided in Appendix A.

A natural question at this point is: What is the expected value of the random variable $E(Y \mid X)$?

Theorem 2.1. *Suppose that $E|Y| < \infty$. Then*

$$E(E(Y \mid X)) = EY.$$

Proof. We prove the theorem for the continuous case and leave the (completely analogous) proof for the discrete case as an exercise.

$$\begin{aligned} E(E(Y \mid X)) &= E h(X) = \int_{-\infty}^{\infty} h(x) f_X(x) dx \\ &= \int_{-\infty}^{\infty} E(Y \mid X = x) f_X(x) dx \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} y f_{Y|X=x}(y) dy \right) f_X(x) dx \end{aligned}$$

$$\begin{aligned}
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y \frac{f_{X,Y}(x,y)}{f_X(x)} f_X(x) dy dx = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{X,Y}(x,y) dy dx \\
&= \int_{-\infty}^{\infty} y \left(\int_{-\infty}^{\infty} f_{X,Y}(x,y) dx \right) dy = \int_{-\infty}^{\infty} y f_Y(y) dy = EY. \quad \square
\end{aligned}$$

Remark 2.1. Theorem 2.1 can be interpreted as an “expectation version” of the law of total probability.

Remark 2.2. Clearly, EY must exist in order for Theorem 2.1 to make sense, that is, the corresponding sum or integral must be absolutely convergent. Now, given this assumption, one can show that $E(E(Y|X))$ exists and is finite and that the computations in the proof, such as reversing orders of integration, are permitted. We shall, in the sequel, permit ourselves at times to be somewhat sloppy about such verifications. Analogous remarks apply to further results ahead.

We close this remark by pointing out that the conclusion always holds in case Y is nonnegative, in the sense that if one of the members is infinite, then so is the other. \square

Exercise 2.2. The object of this exercise is to show that if we do not assume that $E|Y| < \infty$ in Theorem 2.1, then the conclusion does not necessarily hold. Namely, suppose that $X \in \Gamma(1/2, 2)$ ($= \chi^2(1)$) and that

$$f_{Y|X=x}(y) = \frac{1}{\sqrt{2\pi}} x^{\frac{1}{2}} e^{-\frac{1}{2}xy^2}, \quad -\infty < y < \infty.$$

- (a) Compute $E(Y|X=x)$, $E(Y|X)$, and, finally, $E(E(Y|X))$.
- (b) Show that $Y \in C(0, 1)$.
- (c) What about EY ? \square

We are now able to find EY in Example 2.1.

Example 2.1 (continued). It follows from the definition that the first part of (2.1) holds:

$$E(Y | X = x) = \frac{x}{2}, \quad \text{that is, } h(x) = \frac{x}{2}.$$

An application of Theorem 2.1 now yields

$$EY = E(E(Y | X)) = E h(X) = E\left(\frac{1}{2}X\right) = \frac{1}{2}EX = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.$$

We have thus determined EY without prior knowledge about the distribution of Y . \square

Exercise 2.3. Find the expectation of the remaining piece after it has been broken off n times. \square

Remark 2.3. That the result $EY = 1/4$ is reasonable can intuitively be seen from the fact that X on average equals $1/2$ and that Y on average equals half the value of X , that is $1/2$ of $1/2$. The proof of Theorem 2.1 consists, in fact, of a stringent version of this kind of argument. \square

Theorem 2.2. *Let X and Y be random variables and g be a function. We have*

- (a) $E(g(X)Y | X) = g(X) \cdot E(Y | X)$, and \square
- (b) $E(Y | X) = EY$ if X and Y are independent. \square

Exercise 2.4. Prove Theorem 2.2. \square

Remark 2.4. Conditioning with respect to X means that X should be interpreted as known, and, hence, $g(X)$ as a constant that thus may be moved in front of the expectation (recall Exercise 2.1(a)). This explains why Theorem 2.2(a) should hold. Part (b) follows from the fact that the conditional distribution and the unconditional distribution coincide if X and Y are independent; in particular, this should remain true for the conditional expectation and the unconditional expectation (recall Exercises 1.5 and 2.1(e)). \square

A natural problem is to find the variance of the remaining piece Y in Example 2.1, which, in turn, suggests the introduction of the concept of conditional variance.

Definition 2.2. *Let X and Y have a joint distribution. The conditional variance of Y given that $X = x$ is*

$$\text{Var}(Y | X = x) = E((Y - E(Y | X = x))^2 | X = x),$$

provided the corresponding sum or integral is absolutely convergent. \square

The conditional variance is (also) a function of x ; call it $v(x)$. The corresponding random variable is

$$v(X) = \text{Var}(Y | X). \quad (2.4)$$

The following result is fundamental.

Theorem 2.3. *Let X and Y be random variables and g a real-valued function. If $EY^2 < \infty$ and $E(g(X))^2 < \infty$, then*

$$E(Y - g(X))^2 = E \text{Var}(Y | X) + E(E(Y | X) - g(X))^2.$$

Proof. An expansion of the left-hand side yields

$$\begin{aligned} E(Y - g(X))^2 &= E(Y - E(Y | X) + E(Y | X) - g(X))^2 \\ &= E(Y - E(Y | X))^2 + 2E(Y - E(Y | X))(E(Y | X) - g(X)) \\ &\quad + E(E(Y | X) - g(X))^2. \end{aligned}$$

Using Theorem 2.1, the right-hand side becomes

$$\begin{aligned} & E E((Y - E(Y | X))^2 | X) + 2 E E((Y - E(Y | X)) \\ & \quad \times (E(Y | X) - g(X)) | X) + E(E(Y | X) - g(X))^2 \\ & = E \text{Var}(Y | X) + 2 E\{(E(Y | X) - g(X)) E(Y - E(Y | X) | X)\} \\ & \quad + E(E(Y | X) - g(X))^2 \end{aligned}$$

by Theorem 2.2(a). Finally, since $E(Y - E(Y | X) | X) = 0$, this equals

$$E \text{Var}(Y | X) + 2 E\{(E(Y | X) - g(X)) \cdot 0\} + E(E(Y | X) - g(X))^2,$$

which was to be proved. \square

The particular choice $g(X) = EY$, together with an application of Theorem 2.1, yields the following corollary:

Corollary 2.3.1. *Suppose that $EY^2 < \infty$. Then*

$$\text{Var } Y = E \text{Var}(Y | X) + \text{Var}(E(Y | X)). \quad \square$$

Example 2.1 (continued). Let us determine $\text{Var } Y$ with the aid of Corollary 2.3.1.

It follows from second part of formula (2.1) that

$$\text{Var}(Y | X = x) = \frac{1}{12}x^2, \quad \text{and hence,} \quad v(X) = \frac{1}{12}X^2,$$

so that

$$E \text{Var}(Y | X) = E v(X) = E\left(\frac{1}{12}X^2\right) = \frac{1}{12} \cdot \frac{1}{3} = \frac{1}{36}.$$

Furthermore,

$$\text{Var}(E(Y | X)) = \text{Var}(h(X)) = \text{Var}\left(\frac{1}{2}X\right) = \frac{1}{4}\text{Var}(X) = \frac{1}{4} \cdot \frac{1}{12} = \frac{1}{48}.$$

An application of Corollary 2.3.1 finally yields $\text{Var } Y = 1/36 + 1/48 = 7/144$.

We have thus computed $\text{Var } Y$ without knowing the distribution of Y . \square

Exercise 2.5. Find the distribution of Y in Example 2.1, and verify the values of EY and $\text{Var } Y$ obtained above. \square

A discrete variant of Example 2.1 is the following: Let X be uniformly distributed over the numbers $1, 2, \dots, 6$ (that is, throw a symmetric die) and let Y be uniformly distributed over the numbers $1, 2, \dots, X$ (that is, then throw a symmetric die with X faces). In this case,

$$h(x) = E(Y | X = x) = \frac{1+x}{2},$$

from which it follows that

$$EY = E h(X) = E\left(\frac{1+X}{2}\right) = \frac{1}{2}(1 + EX) = \frac{1}{2}(1 + 3.5) = 2.25.$$

The computation of $\text{Var } Y$ is somewhat more elaborate. We leave the details to the reader. \square

3 Distributions with Random Parameters

We begin with two examples:

Example 3.1. Suppose that the density X of red blood corpuscles in humans follows a Poisson distribution whose parameter depends on the observed individual. This means that for Jürg we have $X \in \text{Po}(m_J)$, where m_J is Jürg's parameter value, while for Alice we have $X \in \text{Po}(m_A)$, where m_A is Alice's parameter value. For a person selected at random we may consider the parameter value M as a random variable such that, given that $M = m$, we have $X \in \text{Po}(m)$; namely,

$$P(X = k \mid M = m) = e^{-m} \cdot \frac{m^k}{k!}, \quad k = 0, 1, 2, \dots \quad (3.1)$$

Thus, if we *know* that Alice was chosen, then $P(X = k \mid M = m_A) = e^{-m_A} \cdot m_A^k / k!$, for $k = 0, 1, 2, \dots$, as before. We shall soon see that X itself (unconditioned) need not follow a Poisson distribution.

Example 3.2. A radioactive substance emits α -particles in such a way that the number of emitted particles during an hour, N , follows a $\text{Po}(\lambda)$ -distribution. The particle counter, however, is somewhat unreliable in the sense that an emitted particle is registered with probability p ($0 < p < 1$), whereas it remains unregistered with probability $q = 1 - p$. All particles are registered independently of each other. This means that if we *know* that n particles were emitted during a specific hour, then the number of registered particles $X \in \text{Bin}(n, p)$, that is,

$$P(X = k \mid N = n) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, \dots, n \quad (3.2)$$

(and $N \in \text{Po}(\lambda)$). If, however, we observe the process during an arbitrarily chosen hour, it follows, as will be seen below, that the number of registered particles does not follow a binomial distribution (but instead a Poisson distribution). \square

The common feature in these examples is that the random variable under consideration, X , has a known distribution but with a parameter that is a random variable. Somewhat imprecisely, we might say that in Example 3.1 we have $X \in \text{Po}(M)$, where M follows some distribution, and that in Example 3.2 we have $X \in \text{Bin}(N, p)$, where $N \in \text{Po}(\lambda)$. We prefer, however, to describe these cases as

$$X \mid M = m \in \text{Po}(m) \quad \text{with} \quad M \in F, \quad (3.3)$$

where F is some distribution, and

$$X \mid N = n \in \text{Bin}(n, p) \quad \text{with} \quad N \in \text{Po}(\lambda), \quad (3.4)$$

respectively.

Let us now determine the (unconditional) distributions of X in our examples, where, in Example 3.1, we assume that $M \in \text{Exp}(1)$.

Example 3.1 (continued). We thus have

$$X \mid M = m \in \text{Po}(m) \quad \text{with} \quad M \in \text{Exp}(1). \quad (3.5)$$

By (the continuous version of) the law of total probability, we obtain, for $k = 0, 1, 2, \dots$,

$$\begin{aligned} P(X = k) &= \int_0^\infty P(X = k \mid M = x) \cdot f_M(x) dx \\ &= \int_0^\infty e^{-x} \frac{x^k}{k!} \cdot e^{-x} dx = \int_0^\infty \frac{x^k}{k!} e^{-2x} dx \\ &= \frac{1}{2^{k+1}} \cdot \int_0^\infty \frac{1}{\Gamma(k+1)} 2^{k+1} x^{k+1-1} e^{-2x} dx \\ &= \frac{1}{2^{k+1}} \cdot 1 = \frac{1}{2} \cdot \left(\frac{1}{2}\right)^k, \end{aligned}$$

that is, $X \in \text{Ge}(1/2)$. The unconditional distribution in this case thus is not a Poisson distribution; it is a geometric distribution. \square

Exercise 3.1. Determine the distribution of X if M has

- (a) an $\text{Exp}(a)$ -distribution,
- (b) a $\Gamma(p, a)$ -distribution. \square

Note also that we may use the formulas from Section 2 to compute EX and $\text{Var } X$ without knowing the distribution of X . Namely, since $E(X \mid M = m) = m$ (i.e., $h(M) = E(X \mid M) = M$), Theorem 2.1 yields

$$EX = E(E(X \mid M)) = EM = 1,$$

and Corollary 2.3.1 yields

$$\text{Var } X = E \text{Var}(X \mid M) + \text{Var}(E(X \mid M)) = EM + \text{Var } M = 1 + 1 = 2.$$

If, however, the distribution has been determined (as above), the formulas from Section 2 may be used for checking.

If applied to Exercise 3.1(a), the latter formulas yield $EX = a$ and $\text{Var } X = a + a^2$. Since this situation differs from Example 3.1 only by a rescaling of M , one might perhaps guess that the solution is another geometric distribution. If this were true, we would have

$$EX = a = \frac{q}{p} = \frac{1-p}{p} = \frac{1}{p} - 1; \quad p = \frac{1}{a+1}.$$

This value of p inserted in the expression for the variance yields

$$\frac{q}{p^2} = \frac{1-p}{p^2} = \frac{1}{p^2} - \frac{1}{p} = (a+1)^2 - (a+1) = a^2 + a,$$

which coincides with our computations above and provides the guess that $X \in \text{Ge}(1/(a+1))$.

Remark 3.1. In Example 3.1 we used the results of Section 2.2 to confirm our *result*. In Exercise 3.1(a) they were used to confirm (provide) a *guess*. \square

We now turn to the α -particles.

Example 3.2 (continued). Intuitively, the deficiency of the particle counter implies that the radiation actually measured is, on average, a fraction p of the original Poisson stream of particles. We might therefore expect that the number of registered particles during one hour should be a $\text{Po}(\lambda p)$ -distributed random variable. That this is actually correct is verified next.

The model implies that

$$X \mid N = n \in \text{Bin}(n, p) \quad \text{with} \quad N \in \text{Po}(\lambda).$$

The law of total probability yields, for $k = 0, 1, 2, \dots$,

$$\begin{aligned} P(X = k) &= \sum_{n=0}^{\infty} P(X = k \mid N = n) \cdot P(N = n) \\ &= \sum_{n=k}^{\infty} \binom{n}{k} p^k q^{n-k} \cdot e^{-\lambda} \frac{\lambda^n}{n!} \\ &= \frac{p^k}{k!} e^{-\lambda} \sum_{n=k}^{\infty} \frac{\lambda^n}{(n-k)!} q^{n-k} = \frac{(\lambda p)^k}{k!} e^{-\lambda} \sum_{n=k}^{\infty} \frac{(\lambda q)^{n-k}}{(n-k)!} \\ &= \frac{(\lambda p)^k}{k!} e^{-\lambda} \sum_{j=0}^{\infty} \frac{(\lambda q)^j}{j!} = \frac{(\lambda p)^k}{k!} e^{-\lambda} \cdot e^{\lambda q} = e^{-\lambda p} \frac{(\lambda p)^k}{k!}, \end{aligned}$$

that is, $X \in \text{Po}(\lambda p)$. The unconditional distribution thus is not a binomial distribution; it is a Poisson distribution. \square

Remark 3.2. This is an example of a so-called thinned Poisson process. For more details, we refer to Section 8.6. \square

Exercise 3.2. Use Theorem 2.1 and Corollary 2.3.1 to check the values of $E X$ and $\text{Var } X$. \square

A family of distributions that is of special interest is the family of mixed normal, or mixed Gaussian, distributions. These are normal distributions with a random variance, namely,

$$X \mid \Sigma^2 = y \in N(\mu, y) \quad \text{with} \quad \Sigma^2 \in F, \tag{3.6}$$

where F is some distribution (on $(0, \infty)$).

For simplicity we assume in the following that $\mu = 0$.

As an example, consider normally distributed observations with rare disturbances. More specifically, the observations might be $N(0, 1)$ -distributed with probability 0.99 and $N(0, 100)$ -distributed with probability 0.01. We may write this as

$$X \in N(0, \Sigma^2), \quad \text{where} \quad P(\Sigma^2 = 1) = 0.99 \quad \text{and} \quad P(\Sigma^2 = 100) = 0.01.$$

By Theorem 2.1 it follows immediately that $EX = 0$. As for the variance, Corollary 2.3.1 tells us that

$$\begin{aligned} \text{Var } X &= E \text{Var}(X | \Sigma^2) + \text{Var}(E(X | \Sigma^2)) \\ &= E \Sigma^2 = 0.99 \cdot 1 + 100 \cdot 0.01 = 1.99. \end{aligned}$$

If Σ^2 has a continuous distribution, computations such as those above yield

$$F_X(x) = \int_0^\infty \Phi\left(\frac{x}{\sqrt{y}}\right) f_{\Sigma^2}(y) dy,$$

from which the density function of X is obtained by differentiation:

$$f_X(x) = \int_0^\infty \frac{1}{\sqrt{y}} \phi\left(\frac{x}{\sqrt{y}}\right) f_{\Sigma^2}(y) dy = \int_0^\infty \frac{1}{\sqrt{2\pi y}} e^{-x^2/2y} f_{\Sigma^2}(y) dy. \quad (3.7)$$

Mean and variance can be found via the results of Section 2:

$$\begin{aligned} EX &= E(E(X | \Sigma^2)) = 0, \\ \text{Var } X &= E \text{Var}(X | \Sigma^2) + \text{Var}(E(X | \Sigma^2)) = E \Sigma^2. \end{aligned}$$

Next, we determine the distribution of X under the particular assumption that $\Sigma^2 \in \text{Exp}(1)$. We are thus faced with the situation

$$X | \Sigma^2 = y \in N(0, y) \quad \text{with} \quad \Sigma^2 \in \text{Exp}(1) \quad (3.8)$$

By (3.7),

$$\begin{aligned} f_X(x) &= \int_0^\infty \frac{1}{\sqrt{2\pi y}} e^{-x^2/2y} e^{-y} dy = [\text{set } y = u^2] \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2u^2} e^{-u^2} \cdot 2 du = \sqrt{\frac{2}{\pi}} \int_0^\infty \exp\left\{-\frac{x^2}{2u^2} - u^2\right\} du. \end{aligned}$$

In order to solve this integral, the following device may be of use: Let $x > 0$, set

$$I(x) = \int_0^\infty \exp\left\{-\frac{x^2}{2u^2} - u^2\right\} du,$$

differentiate (differentiation and integration may be interchanged), and make the change of variable $y = x/u\sqrt{2}$. This yields

$$I'(x) = \int_0^\infty \left(-\frac{x}{u^2}\right) \exp\left\{-\frac{x^2}{2u^2} - u^2\right\} du = -\sqrt{2} \int_0^\infty \exp\left\{-y^2 - \frac{x^2}{2y^2}\right\} dy.$$

It follows that I satisfies the differential equation

$$I'(x) = -\sqrt{2}I(x)$$

with the initial condition

$$I(0) = \int_0^\infty e^{-u^2} du = \frac{\sqrt{\pi}}{2},$$

the solution of which is

$$I(x) = \frac{\sqrt{\pi}}{2} e^{-x\sqrt{2}}, \quad x > 0. \quad (3.9)$$

By inserting (3.9) into the expression for $f_X(x)$, and noting that the density is symmetric around $x = 0$, we finally obtain

$$f_X(x) = \sqrt{\frac{2}{\pi}} \frac{\sqrt{\pi}}{2} e^{-|x|\sqrt{2}} = \frac{1}{\sqrt{2}} e^{-|x|\sqrt{2}} = \frac{1}{2} \sqrt{2} e^{-|x|\sqrt{2}}, \quad -\infty < x < \infty,$$

that is, $X \in L(\frac{1}{\sqrt{2}})$; a Laplace distribution.

An extra check yields $EX = 0$ and $\text{Var } X = E\Sigma^2 = 1$ ($= 2 \cdot (\frac{1}{\sqrt{2}})^2$), as desired.

Exercise 3.3. Show that if X has a normal distribution such that the mean is zero and the inverse of the variance is Γ -distributed, viz.,

$$X \mid \Sigma^2 = \lambda \in N(0, 1/\lambda) \quad \text{with} \quad \Sigma^2 \in \Gamma\left(\frac{n}{2}, \frac{2}{n}\right),$$

then $X \in t(n)$.

Exercise 3.4. Sheila has a coin with $P(\text{head}) = p_1$ and Betty has a coin with $P(\text{head}) = p_2$. Sheila tosses her coin m times. Each time she obtains “heads,” Betty tosses her coin (otherwise not). Find the distribution of the total number of heads obtained by Betty.

Further, check that mean and variance coincide with the values obtained by Theorem 2.1 and Corollary 2.3.1. Alternatively, find mean and variance first and try to guess the desired distribution (and check if your guess was correct).

As a hint, observe that the game can be modeled as follows: Let N be the number of heads obtained by Sheila and X be the number of heads obtained by Betty. We thus wish to find the distribution of X , where

$$X \mid N = n \in \text{Bin}(n, p_2) \quad \text{with} \quad N \in \text{Bin}(m, p_1), \quad 0 < p_1, p_2 < 1. \quad \square$$

We shall return to the topic of this section in Section 3.5.

4 The Bayesian Approach

A typical problem in probability theory begins with assumptions such as “let $X \in \text{Po}(m)$,” “let $Y \in N(\mu, \sigma^2)$,” “toss a symmetric coin 15 times,” and so forth. In the computations that follow, one tacitly assumes that all parameters are known, that the coin is *exactly* symmetric, and so on.

In statistics one assumes (certain) parameters to be unknown, for example, that the coin might be asymmetric, and one searches for methods, devices, and rules to decide whether or not one should believe in certain hypotheses. Two typical illustrations in the Gaussian approach are “ μ unknown and σ known” and “ μ and σ unknown.”

The Bayesian approach is a kind of compromise. One claims, for example, that parameters are never *completely* unknown; one always has *some* prior opinion or knowledge about them.

A probabilistic model describing this approach was given in Example 3.1. The opening statement there was that the density of red blood corpuscles follows a Poisson distribution. One interpretation of that statement could have been that whenever we are faced with a blood sample the density of red blood corpuscles in the sample is Poissonian. The Bayesian approach taken in Example 3.1 is that whenever we know from whom the blood sample has been taken, the density of red blood corpuscles in the sample is Poissonian, however, with a parameter depending on the individual. If we do not know from whom the sample has been taken, then the parameter is unknown; it is a random variable following some distribution. We also found that if this distribution is the standard exponential, then the density of red blood corpuscles is geometric (and hence not Poissonian).

The prior knowledge about the parameters in this approach is expressed in such a way that the parameters are assumed to follow some probability distribution, called the *prior* (or a priori) distribution. If one wishes to assume that a parameter is “completely unknown,” one might solve the situation by attributing some uniform distribution to the parameter.

In this terminology we may formulate our findings in Example 3.1 as follows: If the parameter in a Poisson distribution has a standard exponential prior distribution, then the random variable under consideration follows a $\text{Ge}(1/2)$ -distribution.

Frequently, one performs random experiments in order to estimate (unknown) parameters. The estimates are based on observations from some probability distribution. The Bayesian analog is to determine the conditional distribution of the parameter given the result of the random experiment. Such a distribution is called the *posterior* (or a posteriori) distribution.

Next we determine the posterior distribution in Example 3.1.

Example 4.1. The model in the example was

$$X \mid M = m \in \text{Po}(m) \quad \text{with} \quad M \in \text{Exp}(1). \quad (4.1)$$

We further had found that $X \in \text{Ge}(1/2)$. Now we wish to determine the conditional distribution of M given the value of X .

For $x > 0$, we have

$$\begin{aligned} F_{M|X=k}(x) &= P(M \leq x \mid X = k) = \frac{P(\{M \leq x\} \cap \{X = k\})}{P(X = k)} \\ &= \frac{\int_0^x P(X = k \mid M = y) \cdot f_M(y) dy}{P(X = k)} \\ &= \frac{\int_0^x e^{-y} \frac{y^k}{k!} \cdot e^{-y} dy}{(\frac{1}{2})^{k+1}} = \int_0^x \frac{1}{\Gamma(k+1)} y^k 2^{k+1} e^{-2y} dy, \end{aligned}$$

which, after differentiation, yields

$$f_{M|X=k}(x) = \frac{1}{\Gamma(k+1)} x^k 2^{k+1} e^{-2x}, \quad x > 0.$$

Thus, $M \mid X = k \in \Gamma(k+1, \frac{1}{2})$ or, in our new terminology, the posterior distribution of M given that X equals k is $\Gamma(k+1, \frac{1}{2})$. \square

Remark 4.1. Note that, starting from the distribution of X given M (and from that of M), we have determined the distribution of M given X and that the solution of the problem, in fact, amounted to applying a continuous version of Bayes' formula. \square

Exercise 4.1. Check that $E M$ and $\text{Var } M$ are what they are supposed to be by applying Theorem 2.1 and Corollary 2.3.1 to the posterior distribution. \square

We conclude this section by studying coin tossing from the Bayesian point of view under the assumption that nothing is known about $p = P(\text{heads})$.

Let X_n be the number of heads after n coin tosses. *One* possible model is

$$X_n \mid P = p \in \text{Bin}(n, p) \quad \text{with} \quad P \in U(0, 1). \quad (4.2)$$

The prior distribution of P , thus, is the $U(0, 1)$ -distribution. Models of this kind are called *mixed binomial models*.

For $k = 0, 1, 2, \dots, n$, we now obtain (via some facts about the beta distribution)

$$\begin{aligned} P(X_n = k) &= \int_0^1 \binom{n}{k} x^k (1-x)^{n-k} \cdot 1 dx \\ &= \binom{n}{k} \int_0^1 x^{(k+1)-1} (1-x)^{(n+1-k)-1} dx \\ &= \binom{n}{k} \frac{\Gamma(k+1)\Gamma(n+1-k)}{\Gamma(k+1+n+1-k)} \\ &= \frac{n! k! (n-k)!}{k! (n-k)! (n+1)!} = \frac{1}{n+1}. \end{aligned}$$

This means that X_n is uniformly distributed over the integers $0, 1, \dots, n$.

A second thought reveals that this is a very reasonable conclusion. Since *nothing* is known about the coin (in the sense of relation (4.2)), there is nothing that favors a specific outcome, that is, all outcomes should be equally probable.

If p is known, we know that the results in different tosses are independent and that the probability of heads given that we obtained 100 heads in a row (still) equals p . What about these facts in the Bayesian model?

$$\begin{aligned} P(X_{n+1} = n+1 \mid X_n = n) &= \frac{P(\{X_{n+1} = n+1\} \cap \{X_n = n\})}{P(X_n = n)} \\ &= \frac{P(X_{n+1} = n+1)}{P(X_n = n)} \\ &= \frac{\frac{1}{n+2}}{\frac{1}{n+1}} = \frac{n+1}{n+2} \rightarrow 1 \quad \text{as } n \rightarrow \infty. \end{aligned}$$

This means that if we know that there were many heads in a row then the (conditional) probability of another head is very large; the results in different tosses are not at all independent.

Why is this the case? Let us find the posterior distribution of P .

$$\begin{aligned} P(P \leq x \mid X_n = k) &= \frac{\int_0^x P(X_n = k \mid P = y) \cdot f_P(y) dy}{P(X_n = k)} \\ &= \frac{\int_0^x \binom{n}{k} y^k (1-y)^{n-k} \cdot 1 dy}{\frac{1}{n+1}} \\ &= (n+1) \binom{n}{k} \int_0^x y^k (1-y)^{n-k} dy. \end{aligned}$$

Differentiation yields

$$f_{P|X_n=k}(x) = \frac{\Gamma(n+2)}{\Gamma(k+1)\Gamma(n+1-k)} x^k (1-x)^{n-k}, \quad 0 < x < 1,$$

viz., a $\beta(k+1, n+1-k)$ -distribution.

For $k = n$ we obtain in particular (or, by direct computation)

$$f_{P|X_n=n}(x) = (n+1)x^n, \quad 0 < x < 1.$$

It follows that

$$P(P > 1 - \varepsilon \mid X_n = n) = 1 - (1 - \varepsilon)^{n+1} \rightarrow 1 \quad \text{as } n \rightarrow \infty$$

for all $\varepsilon > 0$. This means that if we know that there were many heads in a row then we also know that p is close to 1 and thus that it is very likely that the next toss will yield another head.

Remark 4.2. It is, of course, possible to consider the posterior distribution as a prior distribution for a further random experiment, and so on. \square

5 Regression and Prediction

A common statistics problem is to analyze how different (levels of) treatments or treatment combinations affect the outcome of an experiment. The yield of a crop, for example, may depend on variability in watering, fertilization, climate, and other factors in the various areas where the experiment is performed. One problem is that one cannot predict the outcome y exactly, meaning without error, even if the levels of the treatments x_1, x_2, \dots, x_n are known exactly. An important function for predicting the outcome is the conditional expectation of the (random) outcome Y given the (random) levels of treatment X_1, X_2, \dots, X_n .

Let X_1, X_2, \dots, X_n and Y be jointly distributed random variables, and set

$$h(\mathbf{x}) = h(x_1, \dots, x_n) = E(Y \mid X_1 = x_1, \dots, X_n = x_n) = E(Y \mid \mathbf{X} = \mathbf{x}).$$

Definition 5.1. *The function h is called the regression function Y on \mathbf{X} . \square*

Remark 5.1. For $n = 1$ we have $h(x) = E(Y \mid X = x)$, which is the ordinary conditional expectation. \square

Definition 5.2. *A predictor (for Y) based on \mathbf{X} is a function, $d(\mathbf{X})$. The predictor is called linear if d is linear, that is, if $d(\mathbf{X}) = a_0 + a_1X_1 + \dots + a_nX_n$, where a_0, a_1, \dots, a_n are constants. \square*

Predictors are used to predict (as the name suggests). The prediction error is given by the random variable

$$Y - d(\mathbf{X}). \tag{5.1}$$

There are several ways to compare different predictors. One suitable measure is defined as follows:

Definition 5.3. *The expected quadratic prediction error is*

$$E(Y - d(\mathbf{X}))^2.$$

Moreover, if d_1 and d_2 are predictors, we say that d_1 is better than d_2 if $E(Y - d_1(\mathbf{X}))^2 \leq E(Y - d_2(\mathbf{X}))^2$. \square

In the following we confine ourselves to considering the case $n = 1$. A predictor is thus a function of X , $d(X)$, and the expected quadratic prediction error is $E(Y - d(X))^2$. If the predictor is linear, that is, if $d(X) = a + bX$, where a and b are constants, the expected quadratic prediction error is $E(Y - (a + bX))^2$.

Example 5.1. Pick a point uniformly distributed in the triangle $x, y \geq 0, x + y \leq 1$. We wish to determine the regression functions $E(Y | X = x)$ and $E(X | Y = y)$.

To solve this problem we first note that the joint density of X and Y is

$$f_{X,Y}(x, y) = \begin{cases} c, & \text{for } x, y \geq 0, x + y \leq 1, \\ 0, & \text{otherwise,} \end{cases}$$

where c is some constant, which is found by noticing that the total mass equals 1. We thus have

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx dy = \int_0^1 \left(\int_0^{1-x} c dy \right) dx \\ &= c \int_0^1 (1-x) dx = c \left[-\frac{(1-x)^2}{2} \right]_0^1 = \frac{c}{2}, \end{aligned}$$

from which it follows that $c = 2$.

In order to determine the conditional densities we first compute the marginal ones:

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy = \int_0^{1-x} 2 dy = 2(1-x), \quad 0 < x < 1, \\ f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx = \int_0^{1-y} 2 dx = 2(1-y), \quad 0 < y < 1. \end{aligned}$$

Incidentally, X and Y have the same distribution for reasons of symmetry. Finally,

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x, y)}{f_X(x)} = \frac{2}{2(1-x)} = \frac{1}{1-x}, \quad 0 < y < 1-x,$$

and so

$$E(Y | X = x) = \int_0^{1-x} y \cdot \frac{1}{1-x} dy = \frac{1}{1-x} \left[\frac{y^2}{2} \right]_0^{1-x} = \frac{(1-x)^2}{2(1-x)} = \frac{1-x}{2}$$

and, by symmetry,

$$E(X | Y = y) = \frac{1-y}{2}. \quad \square$$

Remark 5.2. Note also, for example, that $Y | X = x \in U(0, 1-x)$ in the example, that is, the density is, for x fixed, a constant (which is the inverse of the length of the interval $(0, 1-x)$). This implies that $E(Y | X = x) = (1-x)/2$, which agrees with the previous results. It also provides an alternative solution to the last part of the problem. In this case the gain is marginal, but in a more technically complicated situation it might be more substantial. \square

Exercise 5.1. Solve the same problem when

$$f_{X,Y}(x, y) = \begin{cases} cx, & \text{for } 0 < x, y < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Exercise 5.2. Solve the same problem when

$$f_{X,Y}(x, y) = \begin{cases} e^{-y}, & \text{for } 0 < x < y, \\ 0, & \text{otherwise.} \end{cases} \quad \square$$

Theorem 5.1. Suppose that $EY^2 < \infty$. Then $h(X) = E(Y | X)$ (i.e., the regression function Y on X) is the best predictor of Y based on X .

Proof. By Theorem 2.3 we know that for an arbitrary predictor $d(X)$,

$$E(Y - d(X))^2 = E \text{Var}(Y | X) + E(h(X) - d(X))^2 \geq E \text{Var}(Y | X),$$

where equality holds iff $d(X) = h(X)$ (more precisely, iff $P(d(X) = h(X)) = 1$). The choice $d(x) = h(x)$ thus yields minimal expected quadratic prediction error. \square

Example 5.2. In Example 5.1 we found the regression function of Y based on X to be $(1 - X)/2$. By Theorem 5.1 it is the best predictor of Y based on X . A simple calculation shows that the expected quadratic prediction error is $E(Y - (1 - X)/2)^2 = 1/48$.

We also noted that X and Y have the same marginal distribution. A (very) naive suggestion for another predictor therefore might be X itself. The expected quadratic prediction error for this predictor is $E(Y - X)^2 = 1/4 > 1/48$, which shows that the regression function is indeed a better predictor. \square

Sometimes it is difficult to determine regression functions explicitly. In such cases one might be satisfied with the best *linear* predictor. This means that one wishes to minimize $E(Y - (a + bX))^2$ as a function of a and b , which leads to the well-known method of least squares. The solution of this problem is given in the following result.

Theorem 5.2. Suppose that $EX^2 < \infty$ and $EY^2 < \infty$. Set $\mu_x = EX$, $\mu_y = EY$, $\sigma_x^2 = \text{Var } X$, $\sigma_y^2 = \text{Var } Y$, $\sigma_{xy} = \text{Cov}(X, Y)$, and $\rho = \sigma_{xy}/\sigma_x\sigma_y$. The best linear predictor of Y based on X is

$$L(X) = \alpha + \beta X,$$

where

$$\alpha = \mu_y - \frac{\sigma_{xy}}{\sigma_x^2} \mu_x = \mu_y - \rho \frac{\sigma_y}{\sigma_x} \mu_x \quad \text{and} \quad \beta = \frac{\sigma_{xy}}{\sigma_x^2} = \rho \frac{\sigma_y}{\sigma_x}. \quad \square$$

The best linear predictor thus is

$$\mu_y + \rho \frac{\sigma_y}{\sigma_x} (X - \mu_x). \quad (5.2)$$

Definition 5.4. The line $y = \mu_y + \rho \frac{\sigma_y}{\sigma_x} (x - \mu_x)$ is called the regression line Y on X . The slope, $\rho \frac{\sigma_y}{\sigma_x}$, of the line is called the regression coefficient. \square

Remark 5.3. Note that $y = L(x)$, where $L(X)$ is the best linear predictor of Y based on X .

Remark 5.4. If, in particular, (X, Y) has a joint Gaussian distribution, it turns out that the regression function is linear, that is, for this very important case the best linear predictor is, in fact, the best predictor. For details, we refer the reader to Section 5.6. \square

Example 5.1 (continued). The regression function Y on X turned out to be linear in this example; $y = (1-x)/2$. It follows in particular that the regression function coincides with the regression line Y on X . The regression coefficient equals $-1/2$. \square

The expected quadratic prediction error of the best linear predictor of Y based on X is obtained as follows:

Theorem 5.3. $E(Y - L(X))^2 = \sigma_y^2(1 - \rho^2)$.

Proof.

$$\begin{aligned} E(Y - L(X))^2 &= E\left(Y - \mu_y - \rho \frac{\sigma_y}{\sigma_x} (X - \mu_x)\right)^2 = E(Y - \mu_y)^2 \\ &\quad + \rho^2 \frac{\sigma_y^2}{\sigma_x^2} E(X - \mu_x)^2 - 2\rho \frac{\sigma_y}{\sigma_x} E(Y - \mu_y)(X - \mu_x) \\ &= \sigma_y^2 + \rho^2 \cdot \sigma_y^2 - 2\rho \frac{\sigma_y}{\sigma_x} \sigma_{xy} = \sigma_y^2(1 - \rho^2). \end{aligned} \quad \square$$

Definition 5.5. The quantity $\sigma_y^2(1 - \rho^2)$ is called residual variance. \square

Exercise 5.3. Check via Theorem 5.3 that the residual variance in Example 5.1 equals $1/48$ as was claimed in Example 5.2. \square

The regression line X on Y is determined similarly. It is

$$x = \mu_x + \rho \frac{\sigma_x}{\sigma_y} (y - \mu_y),$$

which can be rewritten as

$$y = \mu_y + \frac{1}{\rho} \cdot \frac{\sigma_y}{\sigma_x} (x - \mu_x)$$

if $\rho \neq 0$. The regression lines Y on X and X on Y are thus, in general, different. They coincide iff they have the same slope—iff

$$\rho \cdot \frac{\sigma_y}{\sigma_x} = \frac{1}{\rho} \cdot \frac{\sigma_y}{\sigma_x} \iff |\rho| = 1,$$

that is, iff there exists a linear relation between X and Y . \square

Example 5.1 (continued). The regression function X on Y was also linear (and coincides with the regression line X on Y). The line has the form $x = (1-y)/2$, that is, $y = 1 - 2x$. In particular, we note that the slopes of the regression lines are $-1/2$ and -2 , respectively. \square

6 Problems

- Let X and Y be independent $\text{Exp}(1)$ -distributed random variables. Find the conditional distribution of X given that $X + Y = c$ (c is a positive constant).
- Let X and Y be independent $\Gamma(2, a)$ -distributed random variables. Find the conditional distribution of X given that $X + Y = 2$.
- The life of a repairing device is $\text{Exp}(1/a)$ -distributed. Peter wishes to use it on n different, independent, $\text{Exp}(1/na)$ -distributed occasions.
 - Compute the probability P_n that this is possible.
 - Determine the limit of P_n as $n \rightarrow \infty$.
- The life T (hours) of the lightbulb in an overhead projector follows an $\text{Exp}(10)$ -distribution. During a normal week it is used a $\text{Po}(12)$ -distributed number of lectures lasting exactly one hour each. Find the probability that a projector with a newly installed lightbulb functions throughout a normal week (without replacing the lightbulb).
- The random variables N, X_1, X_2, \dots are independent, $N \in \text{Po}(\lambda)$, and $X_k \in \text{Be}(1/2)$, $k \geq 1$. Set

$$Y_1 = \sum_{k=1}^N X_k \quad \text{and} \quad Y_2 = N - Y_1$$

($Y_1 = 0$ for $N = 0$). Show that Y_1 and Y_2 are independent, and determine their distributions.

- Suppose that $X \in N(0, 1)$ and $Y \in \text{Exp}(1)$ are independent random variables. Prove that $X\sqrt{2Y}$ has a standard Laplace distribution.
- Let $N \in \text{Ge}(p)$ and set $X = (-1)^N$. Compute
 - EX and $\text{Var } X$,
 - the distribution (probability function) of X .
- The density function of the two-dimensional random variable (X, Y) is

$$f_{X,Y}(x, y) = \begin{cases} \frac{x^2}{2 \cdot y^3} \cdot e^{-\frac{x}{y}}, & \text{for } 0 < x < \infty, \quad 0 < y < 1, \\ 0, & \text{otherwise.} \end{cases}$$

- (a) Determine the distribution of Y .
 (b) Find the conditional distribution of X given that $Y = y$.
 (c) Use the results from (a) and (b) to compute EX and $\text{Var } X$.
 9. The density of the random vector $(X, Y)'$ is

$$f_{X,Y}(x, y) = \begin{cases} cx, & \text{for } x \geq 0, \quad y \geq 0, \quad x + y \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

Compute

- (a) c ,
 (b) the conditional expectations $E(Y | X = x)$ and $E(X | Y = y)$.
 10. Suppose X and Y have a joint density function given by

$$f(x, y) = \begin{cases} cx^2, & \text{for } 0 < x < y < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Find c , the marginal density functions, EX , EY , and the conditional expectations $E(Y | X = x)$ and $E(X | Y = y)$.

11. Suppose X and Y have a joint density function given by

$$f(x, y) = \begin{cases} c \cdot x^2 y, & \text{for } 0 < y < x < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Compute c , the marginal densities, EX , EY , and the conditional expectations $E(Y | X = x)$ and $E(X | Y = y)$.

12. Let X and Y have joint density

$$f(x, y) = \begin{cases} cxy, & \text{when } 0 < y < x < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Compute the conditional expectations $E(Y | X = x)$ and $E(X | Y = y)$.

13. Let X and Y have joint density

$$f(x, y) = \begin{cases} cy, & \text{when } 0 < y < x < 2, \\ 0, & \text{otherwise.} \end{cases}$$

Compute the conditional expectations $E(Y | X = x)$ and $E(X | Y = y)$.

14. Suppose that X and Y are random variables with joint density

$$f(x, y) = \begin{cases} c(x + 2y), & \text{when } 0 < x < y < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Compute the regression functions $E(Y | X = x)$ and $E(X | Y = y)$.

15. Suppose that X and Y are random variables with a joint density

$$f(x, y) = \begin{cases} \frac{2}{5}(2x + 3y), & \text{when } 0 < x, y < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Compute the conditional expectations $E(Y | X = x)$ and $E(X | Y = y)$.

16. Let X and Y be random variables with a joint density

$$f(x, y) = \begin{cases} \frac{4}{5}(x + 3y)e^{-x-2y}, & \text{when } x, y > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Compute the regression functions $E(Y | X = x)$ and $E(X | Y = y)$.

17. Suppose that the joint density of X and Y is given by

$$f(x, y) = \begin{cases} xe^{-x-xy}, & \text{when } x > 0, y > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Determine the regression functions $E(Y | X = x)$ and $E(X | Y = y)$.

18. Let the joint density function of X and Y be given by

$$f(x, y) = \begin{cases} c(x + y), & \text{for } 0 < x < y < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Determine c , the marginal densities, EX , EY , and the conditional expectations $E(Y | X = x)$ and $E(X | Y = y)$.

19. Let the joint density of X and Y be given by

$$f_{X,Y}(x, y) = \begin{cases} c, & \text{for } 0 \leq x \leq 1, \quad x^2 \leq y \leq x, \\ 0, & \text{otherwise.} \end{cases}$$

Compute c , the marginal densities, and the conditional expectations $E(Y | X = x)$ and $E(X | Y = y)$.

20. Suppose that X and Y are random variables with joint density

$$f(x, y) = \begin{cases} cx, & \text{when } 0 < x < 1, \quad x^3 < y < x^{1/3}, \\ 0, & \text{otherwise.} \end{cases}$$

Compute the conditional expectations $E(Y | X = x)$ and $E(X | Y = y)$.

21. Suppose that X and Y are random variables with joint density

$$f(x, y) = \begin{cases} cy, & \text{when } 0 < x < 1, \quad x^4 < y < x^{1/4}, \\ 0, & \text{otherwise.} \end{cases}$$

Compute the conditional expectations $E(Y | X = x)$ and $E(X | Y = y)$.

22. Let the joint density function of X and Y be given by

$$f(x, y) = \begin{cases} c \cdot x^3 y, & \text{for } x, y > 0, \quad x^2 + y^2 \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

Compute c , the marginal densities, and the conditional expectations $E(Y | X = x)$ and $E(X | Y = y)$.

23. The joint density function of X and Y is given by

$$f(x, y) = \begin{cases} c \cdot xy, & \text{for } x, y > 0, \quad 4x^2 + y^2 \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

Compute c , the marginal densities, and the conditional expectations $E(Y | X = x)$ and $E(X | Y = y)$.

24. Let X and Y have joint density

$$f(x, y) = \begin{cases} \frac{c}{x^3 y}, & \text{when } 1 < y < x, \\ 0, & \text{otherwise.} \end{cases}$$

Compute the conditional expectations $E(Y | X = x)$ and $E(X | Y = y)$.

25. Let X and Y have joint density

$$f(x, y) = \begin{cases} \frac{c}{x^4 y}, & \text{when } 1 < y < x, \\ 0, & \text{otherwise.} \end{cases}$$

Compute the conditional expectations $E(Y | X = x)$ and $E(X | Y = y)$.

26. Suppose that X and Y are random variables with a joint density

$$f(x, y) = \begin{cases} \frac{c}{(1 + x - y)^2}, & \text{when } 0 < y < x < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Compute the conditional expectations $E(Y | X = x)$ and $E(X | Y = y)$.

27. Suppose that X and Y are random variables with a joint density

$$f(x, y) = \begin{cases} c \cdot \cos x, & \text{when } 0 < y < x < \frac{\pi}{2}, \\ 0, & \text{otherwise.} \end{cases}$$

Compute the conditional expectations $E(Y | X = x)$ and $E(X | Y = y)$.

28. Let X and Y have joint density

$$f(x, y) = \begin{cases} c \log y, & \text{when } 0 < y < x < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Compute the conditional expectations $E(Y | X = x)$ and $E(X | Y = y)$.

29. The random vector $(X, Y)'$ has the following joint distribution:

$$P(X = m, Y = n) = \binom{m}{n} \frac{1}{2^m} \frac{m}{15},$$

where $m = 1, 2, \dots, 5$ and $n = 0, 1, \dots, m$. Compute $E(Y | X = m)$.

30. Show that a suitable power of a Weibull-distributed random variable whose parameter is gamma-distributed is Pareto-distributed. More precisely, show that if

$$X | A = a \in W\left(\frac{1}{a}, \frac{1}{b}\right) \quad \text{with} \quad A \in \Gamma(p, \theta),$$

then X^b has a (translated) Pareto distribution.

31. Show that an exponential random variable such that the inverse of the parameter is gamma-distributed is Pareto-distributed. More precisely, show that if

$$X | M = m \in \text{Exp}(m) \quad \text{with} \quad M^{-1} \in \Gamma(p, a),$$

then X has a (translated) Pareto distribution.

32. Let X and Y be random variables such that

$$Y | X = x \in \text{Exp}(1/x) \quad \text{with} \quad X \in \Gamma(2, 1).$$

(a) Show that Y has a translated Pareto distribution.

(b) Compute EY .

(c) Check the value in (b) by recomputing it via our favorite formula for conditional means.

33. Suppose that the random variable X is uniformly distributed symmetrically around zero, but in such a way that the parameter is uniform on $(0, 1)$; that is, suppose that

$$X | A = a \in U(-a, a) \quad \text{with} \quad A \in U(0, 1).$$

Find the distribution of X , EX , and $\text{Var } X$.

34. In Section 4 we studied the situation when a coin, such that $p = P(\text{head})$ is considered to be a $U(0, 1)$ -distributed random variable, is tossed, and found (i.a.) that if $X_n = \#$ heads after n tosses, then X_n is uniformly distributed over the integers $0, 1, \dots, n$.

Suppose instead that p is considered to be $\beta(2, 2)$ -distributed. What then? More precisely, consider the following model:

$$X_n | Y = y \in \text{Bin}(n, y) \quad \text{with} \quad f_Y(y) = 6y(1 - y), \quad 0 < y < 1.$$

(a) Compute EX_n and $\text{Var } X_n$.

(b) Determine the distribution of X_n .

35. Let X and Y be jointly distributed random variables such that

$$Y | X = x \in \text{Bin}(n, x) \quad \text{with} \quad X \in U(0, 1).$$

Compute EY , $\text{Var } Y$, and $\text{Cov}(X, Y)$ (without using what is known from Section 4 about the distribution of Y).

36. Let X and Y be jointly distributed random variables such that

$$Y \mid X = x \in \text{Fs}(x) \quad \text{with} \quad f_X(x) = 3x^2, \quad 0 \leq x \leq 1.$$

Compute EY , $\text{Var } Y$, $\text{Cov}(X, Y)$, and the distribution of Y .

37. Let X be the number of coin tosses until heads is obtained. Suppose that the probability of heads is unknown in the sense that we consider it to be a random variable $Y \in U(0, 1)$.
- (a) Find the distribution of X (cf. Problem 3.8.48).
 - (b) The expected value of an Fs-distributed random variable exists, as is well known. What about EX ?
 - (c) Suppose that the value $X = n$ has been observed. Find the posterior distribution of Y , that is, the distribution of $Y \mid X = n$.
38. Let p be the probability that the tip points downward after a person throws a drawing pin once. Annika throws a drawing pin until it points downward for the first time. Let X be the number of throws for this to happen. She then throws the drawing pin another X times. Let Y be the number of times the drawing pin points downward in the latter series of throws. Find the distribution of Y (cf. Problem 3.8.31).
39. A point P is chosen uniformly in an n -dimensional sphere of radius 1. Next, a point Q is chosen uniformly within the concentric sphere, centered at the origin, going through P . Let X and Y be the distances of P and Q , respectively, to the common center. Find the joint density function of X and Y and the conditional expectations $E(Y \mid X = x)$ and $E(X \mid Y = y)$.
- Hint 1.* Begin by trying the case $n = 2$.
- Hint 2.* The volume of an n -dimensional sphere of radius r is equal to $c_n r^n$, where c_n is some constant (which is of no interest for the problem).
- Remark.* For $n = 1$ we rediscover the stick from Example 2.1.
40. Let X and Y be independent random variables. The conditional distribution of Y given that $X = x$ then does not depend on x . Moreover, $E(Y \mid X = x)$ is independent of x ; recall Theorem 2.2(b) and Remark 2.4. Now, suppose instead that $E(Y \mid X = x)$ is independent of x (i.e., that $E(Y \mid X) = EY$). We say that Y has *constant regression with respect to X* . However, it does not necessarily follow that X and Y are independent. Namely, let the joint density of X and Y be given by

$$f(x, y) = \begin{cases} \frac{1}{2}, & \text{for } |x| + |y| \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

Show that Y has constant regression with respect to X and/but that X and Y are not independent.



<http://www.springer.com/978-1-4419-0161-3>

An Intermediate Course in Probability

Gut, A.

2009, XV, 303 p., Hardcover

ISBN: 978-1-4419-0161-3