

4.7: Conditional Expected Value

As usual, our starting point is a random experiment modeled by a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. So to review, Ω is the set of outcomes, \mathcal{F} the collection of events, and \mathbb{P} the probability measure on the sample space (Ω, \mathcal{F}) . Suppose next that X is a random variable taking values in a set S and that Y is a random variable taking values in $T \subseteq \mathbb{R}$. We assume that either Y has a discrete distribution, so that T is countable, or that Y has a continuous distribution so that T is an interval (or perhaps a union of intervals). In this section, we will study the conditional expected value of Y given X , a concept of fundamental importance in probability. As we will see, the expected value of Y given X is the function of X that best approximates Y in the mean square sense. Note that X is a general random variable, not necessarily real-valued, but as usual, we will assume that either X has a discrete distribution, so that S is countable or that X has a continuous distribution on $S \subseteq \mathbb{R}^n$ for some $n \in \mathbb{N}_+$. In the latter case, S is typically a region defined by inequalities involving elementary functions. We will also assume that all expected values that are mentioned exist (as real numbers).

Basic Theory

Definitions

Note that we can think of (X, Y) as a random variable that takes values in the Cartesian product set $S \times T$. We need recall some basic facts from our work with joint distributions and conditional distributions.

We assume that (X, Y) has joint probability density function f and we let g denote the (marginal) probability density function X . Recall that if Y has a discrete distribution then

$$g(x) = \sum_{y \in T} f(x, y), \quad x \in S \quad (4.7.1)$$

and if Y has a continuous distribution then

$$g(x) = \int_T f(x, y) dy, \quad x \in S \quad (4.7.2)$$

In either case, for $x \in S$, the conditional probability density function of Y given $X = x$ is defined by

$$h(y | x) = \frac{f(x, y)}{g(x)}, \quad y \in T \quad (4.7.3)$$

We are now ready for the basic definitions:

For $x \in S$, the *conditional expected value* of Y given $X = x \in S$ is simply the mean computed relative to the conditional distribution. So if Y has a discrete distribution then

$$E(Y | X = x) = \sum_{y \in T} yh(y | x), \quad x \in S \quad (4.7.4)$$

and if Y has a continuous distribution then

$$E(Y | X = x) = \int_T yh(y | x) dy, \quad x \in S \quad (4.7.5)$$

1. The function $v : S \rightarrow \mathbb{R}$ defined by $v(x) = E(Y | X = x)$ for $x \in S$ is the *regression function* of Y based on X .
2. The random variable $v(X)$ is called the *conditional expected value* of Y given X and is denoted $E(Y | X)$.

Intuitively, we treat X as known, and therefore not random, and we then average Y with respect to the probability distribution that remains. The advanced section on conditional expected value gives a much more general definition that unifies the definitions given here for the various distribution types.

Properties

The most important property of the random variable $E(Y | X)$ is given in the following theorem. In a sense, this result states that $E(Y | X)$ behaves just like Y in terms of other functions of X , and is essentially the only function of X with this property.

The fundamental property

1. $E[r(X)E(Y | X)] = E[r(X)Y]$ for every function $r : S \rightarrow \mathbb{R}$.
2. If $u : S \rightarrow \mathbb{R}$ satisfies $E[r(X)u(X)] = E[r(X)Y]$ for every $r : S \rightarrow \mathbb{R}$ then $\mathbb{P}[u(X) = E(Y | X)] = 1$.

Proof

We give the proof in the continuous case. The discrete case is analogous, with sums replacing integrals.

1. From the change of variables theorem for expected value,

$$E[r(X)E(Y | X)] = \int_S r(x)E(Y | X = x)g(x) dx = \int_S r(x) \left(\int_T yh(y | x) dy \right) g(x) dx \quad (4.7.6)$$

$$= \int_S \int_T r(x)yh(y | x)g(x) dy dx = \int_{S \times T} r(x)yf(x, y) d(x, y) = E[r(X)Y] \quad (4.7.7)$$

2. Suppose that $u_1 : S \rightarrow \mathbb{R}$ and $u_2 : S \rightarrow \mathbb{R}$ satisfy the condition in (b). Define $r : S \rightarrow \mathbb{R}$ by $r(x) = \mathbf{1}_{[u_1(x) > u_2(x)]}$. Then by assumption, $E[r(X)u_1(X)] = E[r(X)Y] = E[r(X)u_2(X)]$. But if $\mathbb{P}[u_1(X) > u_2(X)] > 0$ then $E[r(X)u_1(X)] > E[r(X)u_2(X)]$, a contradiction. Hence we must have $\mathbb{P}[u_1(X) > u_2(X)] = 0$ and by a symmetric argument, $\mathbb{P}[u_1(X) < u_2(X)] = 0$.

Two random variables that are equal with probability 1 are said to be equivalent. We often think of equivalent random variables as being essentially the same object, so the [fundamental property](#) above essentially characterizes $\mathbb{E}(Y | X)$. That is, we can think of $\mathbb{E}(Y | X)$ as any random variable that is a function of X and satisfies this property. Moreover the fundamental property can be used as a definition of conditional expected value, regardless of the type of the distribution of (X, Y) . If you are interested, read the more advanced treatment of conditional expected value.

Suppose that X is also real-valued. Recall that the best linear predictor of Y based on X was characterized by property (a), but with just two functions: $r(x) = 1$ and $r(x) = x$. Thus the characterization in the fundamental property is certainly reasonable, since (as we show below) $\mathbb{E}(Y | X)$ is the best predictor of Y among *all* functions of X , not just linear functions.

The [basic property](#) is also very useful for establishing other properties of conditional expected value. Our first consequence is the fact that Y and $\mathbb{E}(Y | X)$ have the same mean.

$$\mathbb{E}[\mathbb{E}(Y | X)] = \mathbb{E}(Y).$$

Proof

Let r be the constant function 1 in the [basic property](#).

Aside from the theoretical interest, [this theorem](#) is often a good way to compute $\mathbb{E}(Y)$ when we know the conditional distribution of Y given X . We say that we are computing the expected value of Y by *conditioning* on X .

For many basic properties of ordinary expected value, there are analogous results for conditional expected value. We start with two of the most important: every type of expected value must satisfy two critical properties: linearity and monotonicity. In the following two theorems, the random variables Y and Z are real-valued, and as before, X is a general random variable.

Linear Properties

1. $\mathbb{E}(Y + Z | X) = \mathbb{E}(Y | X) + \mathbb{E}(Z | X)$.
2. $\mathbb{E}(cY | X) = c\mathbb{E}(Y | X)$

Proof

1. Note that $\mathbb{E}(Y | X) + \mathbb{E}(Z | X)$ is a function of X . If $r : S \rightarrow \mathbb{R}$ then

$$\mathbb{E}(r(x)[\mathbb{E}(Y | X) + \mathbb{E}(Z | X)]) = \mathbb{E}[r(X)\mathbb{E}(Y | X)] + \mathbb{E}[r(X)\mathbb{E}(Z | X)] = \mathbb{E}[r(X)Y] + \mathbb{E}[r(X)Z] = \mathbb{E}[r(X)(Y + Z)] \quad (4.7.8)$$

Hence the result follows from the [basic property](#).

2. Note that $c\mathbb{E}(Y | X)$ is a function of X . If $r : S \rightarrow \mathbb{R}$ then

$$\mathbb{E}[r(X)c\mathbb{E}(Y | X)] = c\mathbb{E}[r(X)\mathbb{E}(Y | X)] = c\mathbb{E}[r(X)Y] = \mathbb{E}[r(X)(cY)] \quad (4.7.9)$$

Hence the result follows from the [basic property](#)

Part (a) is the additive property and part (b) is the scaling property. The scaling property will be significantly generalized below in (8).

Positive and Increasing Properties

1. If $Y \geq 0$ then $\mathbb{E}(Y | X) \geq 0$.
2. If $Y \leq Z$ then $\mathbb{E}(Y | X) \leq \mathbb{E}(Z | X)$.
3. $|\mathbb{E}(Y | X)| \leq \mathbb{E}(|Y| | X)$

Proof

1. This follows directly from the definition.
2. Note that if $Y \leq Z$ then $Y - Z \leq 0$ so by (a) and linearity,

$$\mathbb{E}(Y - Z | X) = \mathbb{E}(Y | X) - \mathbb{E}(Z | X) \geq 0 \quad (4.7.10)$$

3. Note that $-|Y| \leq Y \leq |Y|$ and hence by (b) and linearity, $-\mathbb{E}(|Y| | X) \leq \mathbb{E}(Y | X) \leq \mathbb{E}(|Y| | X)$.

Our next few properties relate to the idea that $\mathbb{E}(Y | X)$ is the expected value of Y given X . The first property is essentially a restatement of the [fundamental property](#).

If $r : S \rightarrow \mathbb{R}$, then $Y - \mathbb{E}(Y | X)$ and $r(X)$ are uncorrelated.

Proof

Note that $Y - \mathbb{E}(Y | X)$ has mean 0 by the [mean property](#). Hence, by the [basic property](#),

$$\text{cov}[Y - \mathbb{E}(Y | X), r(X)] = \mathbb{E}\{[Y - \mathbb{E}(Y | X)]r(X)\} = \mathbb{E}[Yr(X)] - \mathbb{E}[\mathbb{E}(Y | X)r(X)] = 0 \quad (4.7.11)$$

The next result states that any (deterministic) function of X acts like a constant in terms of the conditional expected value with respect to X .

If $s : S \rightarrow \mathbb{R}$ then

$$\mathbb{E}[s(X)Y | X] = s(X)\mathbb{E}(Y | X) \quad (4.7.12)$$

Proof

Note that $s(X)\mathbb{E}(Y | X)$ is a function of X . If $r : S \rightarrow \mathbb{R}$ then

$$\mathbb{E}[r(X)s(X)\mathbb{E}(Y | X)] = \mathbb{E}[r(X)s(X)Y] \quad (4.7.13)$$

So the result now follows from the [basic property](#).

The following rule generalizes theorem (8) and is sometimes referred to as the *substitution rule* for conditional expected value.

If $s : S \times T \rightarrow \mathbb{R}$ then

$$\mathbb{E}[s(X, Y) \mid X = x] = \mathbb{E}[s(x, Y) \mid X = x] \quad (4.7.14)$$

In particular, it follows from (8) that $\mathbb{E}[s(X) \mid X] = s(X)$. At the opposite extreme, we have the next result: If X and Y are independent, then knowledge of X gives no information about Y and so the conditional expected value with respect to X reduces to the ordinary (unconditional) expected value of Y .

If X and Y are independent then

$$\mathbb{E}(Y \mid X) = \mathbb{E}(Y) \quad (4.7.15)$$

Proof

Trivially, $\mathbb{E}(Y)$ is a (constant) function of X . If $r : S \rightarrow \mathbb{R}$ then $\mathbb{E}[\mathbb{E}(Y)r(X)] = \mathbb{E}(Y)\mathbb{E}[r(X)] = \mathbb{E}[Yr(X)]$, the last equality by independence. Hence the result follows from the [basic property](#).

Suppose now that Z is real-valued and that X and Y are random variables (all defined on the same probability space, of course). The following theorem gives a consistency condition of sorts. Iterated conditional expected values reduce to a single conditional expected value with respect to the minimum amount of information. For simplicity, we write $\mathbb{E}[Z \mid X, Y]$ rather than $\mathbb{E}[Z \mid (X, Y)]$.

Consistency

1. $\mathbb{E}[\mathbb{E}(Z \mid X, Y) \mid X] = \mathbb{E}(Z \mid X)$
2. $\mathbb{E}[\mathbb{E}(Z \mid X) \mid X, Y] = \mathbb{E}(Z \mid X)$

Proof

1. Suppose that X takes values in S and Y takes values in T , so that (X, Y) takes values in $S \times T$. By definition, $\mathbb{E}(Z \mid X)$ is a function of X . If $r : S \rightarrow \mathbb{R}$ then trivially r can be thought of as a function on $S \times T$ as well. Hence

$$\mathbb{E}[r(X)\mathbb{E}(Z \mid X)] = \mathbb{E}[r(X)Z] = \mathbb{E}[r(X)\mathbb{E}(Z \mid X, Y)] \quad (4.7.16)$$

It follows from the [basic property](#) that $\mathbb{E}[\mathbb{E}(Z \mid X, Y) \mid X] = \mathbb{E}(Z \mid X)$.

2. Note that since $\mathbb{E}(Z \mid X)$ is a function of X , it is trivially a function of (X, Y) . Hence from (8), $\mathbb{E}[\mathbb{E}(Z \mid X) \mid X, Y] = \mathbb{E}(Z \mid X)$.

Finally we show that $\mathbb{E}(Y \mid X)$ has the same covariance with X as does Y , not surprising since again, $\mathbb{E}(Y \mid X)$ behaves just like Y in its relations with X .

$$\text{cov}[X, \mathbb{E}(Y \mid X)] = \text{cov}(X, Y).$$

Proof

$\text{cov}[X, \mathbb{E}(Y \mid X)] = \mathbb{E}[X\mathbb{E}(Y \mid X)] - \mathbb{E}(X)\mathbb{E}[\mathbb{E}(Y \mid X)]$. But $\mathbb{E}[X\mathbb{E}(Y \mid X)] = \mathbb{E}(XY)$ by [basic property](#), and $\mathbb{E}[\mathbb{E}(Y \mid X)] = \mathbb{E}(Y)$ by the [mean property](#). Hence $\text{cov}[X, \mathbb{E}(Y \mid X)] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \text{cov}(X, Y)$.

Conditional Probability

The conditional probability of an event A , given random variable X (as above), can be defined as a special case of the conditional expected value. As usual, let $\mathbf{1}_A$ denote the indicator random variable of A .

If A is an event, defined

$$\mathbb{P}(A \mid X) = \mathbb{E}(\mathbf{1}_A \mid X) \quad (4.7.17)$$

Here is the fundamental property for conditional probability:

The fundamental property

1. $\mathbb{E}[r(X)\mathbb{P}(A \mid X)] = \mathbb{E}[r(X)\mathbf{1}_A]$ for every function $r : S \rightarrow \mathbb{R}$.
2. If $u : S \rightarrow \mathbb{R}$ and $u(X)$ satisfies $\mathbb{E}[r(X)u(X)] = \mathbb{E}[r(X)\mathbf{1}_A]$ for every function $r : S \rightarrow \mathbb{R}$, then $\mathbb{P}[u(X) = \mathbb{P}(A \mid X)] = 1$.

For example, suppose that X has a discrete distribution on a countable set S with probability density function g . Then (a) becomes

$$\sum_{x \in S} r(x)\mathbb{P}(A \mid X = x)g(x) = \sum_{x \in S} r(x)\mathbb{P}(A, X = x) \quad (4.7.18)$$

But this is obvious since $\mathbb{P}(A \mid X = x) = \mathbb{P}(A, X = x) / \mathbb{P}(X = x)$ and $g(x) = \mathbb{P}(X = x)$. Similarly, if X has a continuous distribution on $S \subseteq \mathbb{R}^n$ then (a) states that

$$\mathbb{E}[r(X)\mathbf{1}_A] = \int_S r(x)\mathbb{P}(A \mid X = x)g(x) dx \quad (4.7.19)$$

The properties above for conditional expected value, of course, have special cases for conditional probability.

$$\mathbb{P}(A) = \mathbb{E}[\mathbb{P}(A \mid X)].$$

Proof

This is a direct result of the [mean property](#), since $\mathbb{E}(\mathbf{1}_A) = \mathbb{P}(A)$.

Again, the result in the previous exercise is often a good way to compute $\mathbb{P}(A)$ when we know the conditional probability of A given X . We say that we are computing the probability of A by *conditioning* on X . This is a very compact and elegant version of the conditioning result given first in the section on Conditional Probability in the

chapter on Probability Spaces and later in the section on Discrete Distributions in the Chapter on Distributions.

The following result gives the conditional version of the axioms of probability.

Axioms of probability

1. $\mathbb{P}(A | X) \geq 0$ for every event A .
2. $\mathbb{P}(\Omega | X) = 1$
3. If $\{A_i : i \in I\}$ is a countable collection of disjoint events then $\mathbb{P}(\bigcup_{i \in I} A_i | X) = \sum_{i \in I} \mathbb{P}(A_i | X)$.

Details

There are some technical issues involving the countable additivity property (c). The conditional probabilities are random variables, and so for a given collection $\{A_i : i \in I\}$, the left and right sides are the same with probability 1. We will return to this point in the more advanced section on conditional expected value

From the last result, it follows that other standard probability rules hold for conditional probability given X . These results include

- the complement rule
- the increasing property
- Boole's inequality
- Bonferroni's inequality
- the inclusion-exclusion laws

The Best Predictor

The next result shows that, of all functions of X , $\mathbb{E}(Y | X)$ is closest to Y , in the sense of mean square error. This is fundamentally important in statistical problems where the *predictor vector* X can be observed but not the *response variable* Y . In this subsection and the next, we assume that the real-valued random variables have finite variance.

If $u : S \rightarrow \mathbb{R}$, then

1. $\mathbb{E}([\mathbb{E}(Y | X) - Y]^2) \leq \mathbb{E}([u(X) - Y]^2)$
2. Equality holds in (a) if and only if $u(X) = \mathbb{E}(Y | X)$ with probability 1.

Proof

1. Note that

$$\mathbb{E}([Y - u(X)]^2) = \mathbb{E}([Y - \mathbb{E}(Y | X) + \mathbb{E}(Y | X) - u(X)]^2) \quad (4.7.20)$$

$$= \mathbb{E}([Y - \mathbb{E}(Y | X)]^2) + 2\mathbb{E}([Y - \mathbb{E}(Y | X)][\mathbb{E}(Y | X) - u(X)]) + \mathbb{E}([\mathbb{E}(Y | X) - u(X)]^2) \quad (4.7.21)$$

But $Y - \mathbb{E}(Y | X)$ has mean 0, so the middle term on the right is $2\text{cov}[Y - \mathbb{E}(Y | X), \mathbb{E}(Y | X) - u(X)]$. Moreover, $\mathbb{E}(Y | X) - u(X)$ is a function of X and hence is uncorrelated with $Y - \mathbb{E}(Y | X)$ by the general [uncorrelated property](#). Hence the middle term is 0, so

$$\mathbb{E}([Y - u(X)]^2) = \mathbb{E}([Y - \mathbb{E}(Y | X)]^2) + \mathbb{E}([\mathbb{E}(Y | X) - u(X)]^2) \quad (4.7.22)$$

and therefore $\mathbb{E}([Y - \mathbb{E}(Y | X)]^2) \leq \mathbb{E}([Y - u(X)]^2)$.

2. Equality holds if and only if $\mathbb{E}([\mathbb{E}(Y | X) - u(X)]^2) = 0$, if and only if $\mathbb{P}[u(X) = \mathbb{E}(Y | X)] = 1$.

Suppose now that X is real-valued. In the section on covariance and correlation, we found that the best *linear* predictor of Y given X is

$$L(Y | X) = \mathbb{E}(Y) + \frac{\text{cov}(X, Y)}{\text{var}(X)}[X - \mathbb{E}(X)] \quad (4.7.23)$$

On the other hand, $\mathbb{E}(Y | X)$ is the best predictor of Y among *all* functions of X . It follows that if $\mathbb{E}(Y | X)$ happens to be a linear function of X then it must be the case that $\mathbb{E}(Y | X) = L(Y | X)$. However, we will give a direct proof also:

If $\mathbb{E}(Y | X) = a + bX$ for constants a and b then $\mathbb{E}(Y | X) = L(Y | X)$; that is,

1. $b = \text{cov}(X, Y) / \text{var}(X)$
2. $a = \mathbb{E}(Y) - \mathbb{E}(X)\text{cov}(X, Y) / \text{var}(X)$

Proof

First, $\mathbb{E}(Y) = \mathbb{E}[\mathbb{E}(Y | X)] = a + b\mathbb{E}(X)$, so $a = \mathbb{E}(Y) - b\mathbb{E}(X)$. Next, $\text{cov}(X, Y) = \text{cov}[X\mathbb{E}(Y | X)] = \text{cov}(X, a + bX) = b\text{var}(X)$ and therefore $b = \text{cov}(X, Y) / \text{var}(X)$.

Conditional Variance

The conditional variance of Y given X is defined like the ordinary variance, but with all expected values conditioned on X .

The *conditional variance* of Y given X is defined as

$$\text{var}(Y | X) = \mathbb{E}([Y - \mathbb{E}(Y | X)]^2 | X) \quad (4.7.24)$$

Thus, $\text{var}(Y | X)$ is a function of X , and in particular, is a random variable. Our first result is a computational formula that is analogous to the one for standard variance —the variance is the mean of the square minus the square of the mean, but now with all expected values conditioned on X :

$$\text{var}(Y | X) = \mathbb{E}(Y^2 | X) - [\mathbb{E}(Y | X)]^2.$$

Proof

Expanding the square in the definition and using basic properties of conditional expectation, we have

$$\text{var}(Y | X) = \mathbb{E} \left(Y^2 - 2Y\mathbb{E}(Y | X) + [\mathbb{E}(Y | X)]^2 \mid X \right) = \mathbb{E}(Y^2 | X) - 2\mathbb{E}[Y\mathbb{E}(Y | X) | X] + \mathbb{E}([\mathbb{E}(Y | X)]^2 | X) \quad (4.7.25)$$

$$= \mathbb{E}(Y^2 | X) - 2\mathbb{E}(Y | X)\mathbb{E}(Y | X) + [\mathbb{E}(Y | X)]^2 = \mathbb{E}(Y^2 | X) - [\mathbb{E}(Y | X)]^2 \quad (4.7.26)$$

Our next result shows how to compute the ordinary variance of Y by *conditioning on X* .

$$\text{var}(Y) = \mathbb{E}[\text{var}(Y | X)] + \text{var}[\mathbb{E}(Y | X)].$$

Proof

From the previous theorem and properties of conditional expected value we have $\mathbb{E}[\text{var}(Y | X)] = \mathbb{E}(Y^2) - \mathbb{E}([\mathbb{E}(Y | X)]^2)$. But $\mathbb{E}(Y^2) = \text{var}(Y) + [\mathbb{E}(Y)]^2$ and similarly, $\mathbb{E}([\mathbb{E}(Y | X)]^2) = \text{var}[\mathbb{E}(Y | X)] + (\mathbb{E}[\mathbb{E}(Y | X)])^2$. But also, $\mathbb{E}[\mathbb{E}(Y | X)] = \mathbb{E}(Y)$ so substituting we get $\mathbb{E}[\text{var}(Y | X)] = \text{var}(Y) - \text{var}[\mathbb{E}(Y | X)]$.

Thus, the variance of Y is the expected conditional variance plus the variance of the conditional expected value. This result is often a good way to compute $\text{var}(Y)$ when we know the conditional distribution of Y given X . With the help of (21) we can give a formula for the mean square error when $\mathbb{E}(Y | X)$ is used a predictor of Y .

Mean square error

$$\mathbb{E}([Y - \mathbb{E}(Y | X)]^2) = \text{var}(Y) - \text{var}[\mathbb{E}(Y | X)] \quad (4.7.27)$$

Proof

From the definition of [conditional variance](#), and using [mean property](#) and [variance formula](#) we have

$$\mathbb{E}([Y - \mathbb{E}(Y | X)]^2) = \mathbb{E}[\text{var}(Y | X)] = \text{var}(Y) - \text{var}[\mathbb{E}(Y | X)] \quad (4.7.28)$$

Let us return to the study of predictors of the real-valued random variable Y , and compare the three predictors we have studied in terms of mean square error.

Suppose that Y is a real-valued random variable.

1. The best *constant predictor* of Y is $\mathbb{E}(Y)$ with mean square error $\text{var}(Y)$.
2. If X is another real-valued random variable, then the best *linear predictor* of Y given X is

$$L(Y | X) = \mathbb{E}(Y) + \frac{\text{cov}(X, Y)}{\text{var}(X)}[X - \mathbb{E}(X)] \quad (4.7.29)$$

with mean square error $\text{var}(Y)[1 - \text{cor}^2(X, Y)]$.

3. If X is a general random variable, then the best *overall predictor* of Y given X is $\mathbb{E}(Y | X)$ with mean square error $\text{var}(Y) - \text{var}[\mathbb{E}(Y | X)]$.

Conditional Covariance

Suppose that Y and Z are real-valued random variables, and that X is a general random variable, all defined on our underlying probability space. Analogous to variance, the conditional covariance of Y and Z given X is defined like the ordinary covariance, but with all expected values conditioned on X .

The *conditional covariance* of Y and Z given X is defined as

$$\text{cov}(Y, Z | X) = \mathbb{E}([Y - \mathbb{E}(Y | X)][Z - \mathbb{E}(Z | X)] | X) \quad (4.7.30)$$

Thus, $\text{cov}(Y, Z | X)$ is a function of X , and in particular, is a random variable. Our first result is a computational formula that is analogous to the one for standard covariance—the covariance is the mean of the product minus the product of the means, but now with all expected values conditioned on X :

$$\text{cov}(Y, Z | X) = \mathbb{E}(YZ | X) - \mathbb{E}(Y | X)\mathbb{E}(Z | X).$$

Proof

Expanding the product in the definition and using basic properties of conditional expectation, we have

$$\begin{aligned} \text{cov}(Y, Z | X) &= \mathbb{E} \left(YZ - Y\mathbb{E}(Z | X) - Z\mathbb{E}(Y | X) + \mathbb{E}(Y | X)\mathbb{E}(Z | X) \mid X \right) = \mathbb{E}(YZ | X) - \mathbb{E}[Y\mathbb{E}(Z | X) | X] - \mathbb{E}[Z\mathbb{E}(Y | X) | X] \\ &\quad + \mathbb{E}[\mathbb{E}(Y | X)\mathbb{E}(Z | X) | X] \end{aligned} \quad (4.7.31)$$

$$= \mathbb{E}(YZ | X) - \mathbb{E}(Y | X)\mathbb{E}(Z | X) - \mathbb{E}(Y | X)\mathbb{E}(Z | X) + \mathbb{E}(Y | X)\mathbb{E}(Z | X) = \mathbb{E}(YZ | X) - \mathbb{E}(Y | X)\mathbb{E}(Z | X) \quad (4.7.32)$$

Our next result shows how to compute the ordinary covariance of Y and Z by *conditioning on X* .

$$\text{cov}(Y, Z) = \mathbb{E}[\text{cov}(Y, Z | X)] + \text{cov}[\mathbb{E}(Y | X), \mathbb{E}(Z | X)].$$

Proof

From (25) and properties of conditional expected value we have

$$\mathbb{E}[\text{cov}(Y, Z | X)] = \mathbb{E}(YZ) - \mathbb{E}[\mathbb{E}(Y | X)\mathbb{E}(Z | X)] \quad (4.7.33)$$

But $\mathbb{E}(YZ) = \text{cov}(Y, Z) + \mathbb{E}(Y)\mathbb{E}(Z)$ and similarly,

$$\mathbb{E}[\mathbb{E}(Y | X)\mathbb{E}(Z | X)] = \text{cov}[\mathbb{E}(Y | X), \mathbb{E}(Z | X)] + \mathbb{E}[\mathbb{E}(Y | X)]\mathbb{E}[\mathbb{E}(Z | X)] \quad (4.7.34)$$

But also, $\mathbb{E}[\mathbb{E}(Y | X)] = \mathbb{E}(Y)$ and $\mathbb{E}[\mathbb{E}(Z | X)] = \mathbb{E}(Z)$ so substituting we get

$$\mathbb{E}[\text{cov}(Y, Z | X)] = \text{cov}(Y, Z) - \text{cov}[\mathbb{E}(Y | X), \mathbb{E}(Z | X)] \quad (4.7.35)$$

Thus, the covariance of Y and Z is the expected conditional covariance plus the covariance of the conditional expected values. This result is often a good way to compute $\text{cov}(Y, Z)$ when we know the conditional distribution of (Y, Z) given X .

Examples and Applications

As always, be sure to try the proofs and computations yourself before reading the ones in the text.

Simple Continuous Distributions

Suppose that (X, Y) has probability density function f defined by $f(x, y) = x + y$ for $0 \leq x \leq 1, 0 \leq y \leq 1$.

1. Find $L(Y | X)$.
2. Find $\mathbb{E}(Y | X)$.
3. Graph $L(Y | X = x)$ and $\mathbb{E}(Y | X = x)$ as functions of x , on the same axes.
4. Find $\text{var}(Y)$.
5. Find $\text{var}(Y) [1 - \text{cor}^2(X, Y)]$.
6. Find $\text{var}(Y) - \text{var}[\mathbb{E}(Y | X)]$.

Answer

1. $\frac{7}{11} - \frac{1}{11}X$
2. $\frac{3X+2}{6X+3}$
4. $\frac{11}{144} = 0.0764$
5. $\frac{5}{66} = 0.0758$
6. $\frac{1}{12} - \frac{1}{144} \ln 3 = 0.0757$

Suppose that (X, Y) has probability density function f defined by $f(x, y) = 2(x + y)$ for $0 \leq x \leq y \leq 1$.

1. Find $L(Y | X)$.
2. Find $\mathbb{E}(Y | X)$.
3. Graph $L(Y | X = x)$ and $\mathbb{E}(Y | X = x)$ as functions of x , on the same axes.
4. Find $\text{var}(Y)$.
5. Find $\text{var}(Y) [1 - \text{cor}^2(X, Y)]$.
6. Find $\text{var}(Y) - \text{var}[\mathbb{E}(Y | X)]$.

Answer

1. $\frac{26}{43} + \frac{15}{43}X$
2. $\frac{5X^2+5X+2}{9X+3}$
4. $\frac{3}{80} = 0.0375$
5. $\frac{13}{430} = 0.0302$
6. $\frac{1837}{21870} - \frac{512}{6561} \ln(2) = 0.0299$

Suppose that (X, Y) has probability density function f defined by $f(x, y) = 6x^2y$ for $0 \leq x \leq 1, 0 \leq y \leq 1$.

1. Find $L(Y | X)$.
2. Find $\mathbb{E}(Y | X)$.
3. Graph $L(Y | X = x)$ and $\mathbb{E}(Y | X = x)$ as functions of x , on the same axes.
4. Find $\text{var}(Y)$.
5. Find $\text{var}(Y) [1 - \text{cor}^2(X, Y)]$.
6. Find $\text{var}(Y) - \text{var}[\mathbb{E}(Y | X)]$.

Answer

Note that X and Y are independent.

1. $\frac{2}{3}$
2. $\frac{2}{3}$
4. $\frac{1}{18}$
5. $\frac{1}{18}$
6. $\frac{1}{18}$

Suppose that (X, Y) has probability density function f defined by $f(x, y) = 15x^2y$ for $0 \leq x \leq y \leq 1$.

1. Find $L(Y | X)$.
2. Find $\mathbb{E}(Y | X)$.

3. Graph $L(Y | X = x)$ and $\mathbb{E}(Y | X = x)$ as functions of x , on the same axes.
4. Find $\text{var}(Y)$.
5. Find $\text{var}(Y) [1 - \text{cor}^2(X, Y)]$.
6. Find $\text{var}(Y) - \text{var}[\mathbb{E}(Y | X)]$.

Answer

1. $\frac{30}{51} + \frac{20}{51}X$
2. $\frac{2(X^2+X+1)}{3(X+1)}$
4. $\frac{5}{252} = 0.0198$
5. $\frac{5}{357} = 0.0140$
6. $\frac{292}{63} - \frac{20}{3}\ln(2) = 0.0139$

Exercises on Basic Properties

Suppose that X , Y , and Z are real-valued random variables with $\mathbb{E}(Y | X) = X^3$ and $\mathbb{E}(Z | X) = \frac{1}{1+X^2}$. Find $\mathbb{E}(Y e^X - Z \sin X | X)$.

Answer

$$X^3 e^X - \frac{\sin X}{1+X^2}$$

Uniform Distributions

As usual, continuous uniform distributions can give us some geometric insight.

Recall first that for $n \in \mathbb{N}_+$, the standard measure on \mathbb{R}^n is

$$\lambda_n(A) = \int_A 1 dx, \quad A \subseteq \mathbb{R}^n \quad (4.7.36)$$

In particular, $\lambda_1(A)$ is the length of $A \subseteq \mathbb{R}$, $\lambda_2(A)$ is the area of $A \subseteq \mathbb{R}^2$, and $\lambda_3(A)$ is the volume of $A \subseteq \mathbb{R}^3$.

Details

Technically λ_n is Lebesgue measure on the measurable subsets of \mathbb{R}^n . The integral representation is valid for the types of sets that occur in applications. In the discussion below, all subsets are assumed to be measurable.

With our usual setup, suppose that X takes values in $S \subseteq \mathbb{R}^n$, Y takes values in $T \subseteq \mathbb{R}$, and that (X, Y) is uniformly distributed on $R \subseteq S \times T \subseteq \mathbb{R}^{n+1}$. So $0 < \lambda_{n+1}(R) < \infty$, and the joint probability density function f of (X, Y) is given by $f(x, y) = 1/\lambda_{n+1}(R)$ for $(x, y) \in R$. Recall that uniform distributions, whether discrete or continuous, always have constant densities. Finally, recall that the cross section of R at $x \in S$ is $T_x = \{y \in T : (x, y) \in R\}$.

In the setting above, suppose that T_x is a bounded interval with midpoint $m(x)$ and length $l(x)$ for each $x \in S$. Then

1. $\mathbb{E}(Y | X) = m(X)$
2. $\text{var}(Y | X) = \frac{1}{12}l^2(X)$

Proof

This follows immediately from the fact that the conditional distribution of Y given $X = x$ is uniformly distributed on T_x for each $x \in S$.

So in particular, the regression curve $x \mapsto \mathbb{E}(Y | X = x)$ follows the midpoints of the cross-sectional intervals.

In each case below, suppose that (X, Y) is uniformly distributed on the give region. Find $\mathbb{E}(Y | X)$ and $\text{var}(Y | X)$

1. The rectangular region $R = [a, b] \times [c, d]$ where $a < b$ and $c < d$.
2. The triangular region $T = \{(x, y) \in \mathbb{R}^2 : -a \leq x \leq y \leq a\}$ where $a > 0$.
3. The circular region $C = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq r\}$ where $r > 0$.

Answer

1. $\mathbb{E}(Y | X) = \frac{1}{2}(c+d)$, $\text{var}(Y | X) = \frac{1}{12}(d-c)^2$. Note that X and Y are independent.
2. $\mathbb{E}(Y | X) = \frac{1}{2}(a+X)$, $\text{var}(Y | X) = \frac{1}{12}(a-X)^2$
3. $\mathbb{E}(Y | X) = 0$, $\text{var}(Y | X) = 4(r^2 - X^2)$

In the bivariate uniform experiment, select each of the following regions. In each case, run the simulation 2000 times and note the relationship between the cloud of points and the graph of the regression function.

1. square
2. triangle
3. circle

Suppose that X is uniformly distributed on the interval $(0, 1)$, and that given X , random variable Y is uniformly distributed on $(0, X)$. Find each of the following:

1. $\mathbb{E}(Y | X)$
2. $\mathbb{E}(Y)$
3. $\text{var}(Y | X)$
4. $\text{var}(Y)$

Answer

1. $\frac{1}{2}X$
2. $\frac{1}{4}$
3. $\frac{1}{12}X^2$
4. $\frac{7}{144}$

The Hypergeometric Distribution

Suppose that a population consists of m objects, and that each object is one of three types. There are a objects of type 1, b objects of type 2, and $m - a - b$ objects of type 0. The parameters a and b are positive integers with $a + b < m$. We sample n objects from the population at random, and without replacement, where $n \in \{0, 1, \dots, m\}$. Denote the number of type 1 and 2 objects in the sample by X and Y , so that the number of type 0 objects in the sample is $n - X - Y$. In the chapter on Distributions, we showed that the joint, marginal, and conditional distributions of X and Y are all hypergeometric—only the parameters change. Here is the relevant result for this section:

In the setting above,

1. $\mathbb{E}(Y | X) = \frac{b}{m-a}(n - X)$
2. $\text{var}(Y | X) = \frac{b(m-a-b)}{(m-a)^2(m-a-1)}(n - X)(m - a - n + X)$
3. $\mathbb{E}([Y - \mathbb{E}(Y | X)]^2) = \frac{n(m-n)b(m-a-b)}{m(m-1)(m-a)}$

Proof

Recall that (X, Y) has the (multivariate) hypergeometric distribution with parameters m, a, b , and n . Marginally, X has the hypergeometric distribution with parameters m, a , and n , and Y has the hypergeometric distribution with parameters m, b , and n . Given $X = x \in \{0, 1, \dots, n\}$, the remaining $n - x$ objects are chosen at random from a population of $m - a$ objects, of which b are type 2 and $m - a - b$ are type 0. Hence, the conditional distribution of Y given $X = x$ is hypergeometric with parameters $m - a, b$, and $n - x$. Parts (a) and (b) then follow from the standard formulas for the mean and variance of the hypergeometric distribution, as functions of the parameters. Part (c) is the mean square error, and in this case can be computed most easily as

$$\text{var}(Y) - \text{var}[\mathbb{E}(Y | X)] = \text{var}(Y) - \left(\frac{b}{m-a}\right)^2 \text{var}(X) = n \frac{b}{m} \frac{m-b}{m} \frac{m-n}{m-1} - \left(\frac{b}{m-a}\right)^2 n \frac{a}{m} \frac{m-a}{m} \frac{m-n}{m-1} \quad (4.7.37)$$

Simplifying gives the result.

Note that $\mathbb{E}(Y | X)$ is a linear function of X and hence $\mathbb{E}(Y | X) = L(Y | X)$.

In a collection of 120 objects, 50 are classified as *good*, 40 as *fair* and 30 as *poor*. A sample of 20 objects is selected at random and without replacement. Let X denote the number of good objects in the sample and Y the number of poor objects in the sample. Find each of the following:

1. $\mathbb{E}(Y | X)$
2. $\text{var}(Y | X)$
3. The predicted value of Y given $X = 8$

Answer

1. $\mathbb{E}(Y | X) = \frac{80}{7} - \frac{4}{7}X$
2. $\text{var}(Y | X) = \frac{4}{1127}(20 - X)(50 + X)$
3. $\frac{48}{7}$

The Multinomial Trials Model

Suppose that we have a sequence of n independent trials, and that each trial results in one of three outcomes, denoted 0, 1, and 2. On each trial, the probability of outcome 1 is p , the probability of outcome 2 is q , so that the probability of outcome 0 is $1 - p - q$. The parameters $p, q \in (0, 1)$ with $p + q < 1$, and of course $n \in \mathbb{N}_+$. Let X denote the number of trials that resulted in outcome 1, Y the number of trials that resulted in outcome 2, so that $n - X - Y$ is the number of trials that resulted in outcome 0. In the chapter on Distributions, we showed that the joint, marginal, and conditional distributions of X and Y are all multinomial—only the parameters change. Here is the relevant result for this section:

In the setting above,

1. $\mathbb{E}(Y | X) = \frac{q}{1-p}(n - X)$
2. $\text{var}(Y | X) = \frac{q(1-p-q)}{(1-p)^2}(n - X)$
3. $\mathbb{E}([Y - \mathbb{E}(Y | X)]^2) = \frac{q(1-p-q)}{1-p}n$

Proof

Recall that (X, Y) has the multinomial distribution with parameters n, p , and q . Marginally, X has the binomial distribution with parameters n and p , and Y has the binomial distribution with parameters n and q . Given $X = x \in \{0, 1, \dots, n\}$, the remaining $n - x$ trials are independent, but with just two outcomes: outcome 2 occurs with probability $q/(1 - p)$ and outcome 0 occurs with probability $1 - q/(1 - p)$. (These are the conditional probabilities of outcomes 2 and 0, respectively, given that outcome 1 did not occur.) Hence the conditional distribution of Y given $X = x$ is binomial with parameters $n - x$ and $q/(1 - p)$. Parts (a) and (b) then follow from the standard formulas for the mean and variance of the binomial distribution, as functions of the parameters. Part (c) is the mean square error and in this case can be computed most easily from

$$\mathbb{E}[\text{var}(Y | X)] = \frac{q(1-p-q)}{(1-p)^2}[n - \mathbb{E}(X)] = \frac{q(1-p-q)}{(1-p)^2}(n - np) = \frac{q(1-p-q)}{1-p}n \quad (4.7.38)$$

Note again that $\mathbb{E}(Y | X)$ is a linear function of X and hence $\mathbb{E}(Y | X) = L(Y | X)$.

Suppose that a fair, 12-sided die is thrown 50 times. Let X denote the number of throws that resulted in a number from 1 to 5, and Y the number of throws that resulted in a number from 6 to 9. Find each of the following:

1. $\mathbb{E}(Y | X)$
2. $\text{var}(Y | X)$
3. The predicted value of Y given $X = 20$

Answer

1. $\mathbb{E}(Y | X) = \frac{4}{7}(50 - X)$
2. $\text{var}(Y | X) = \frac{12}{49}(50 - X)$
3. $\frac{120}{7}$

The Poisson Distribution

Recall that the *Poisson distribution*, named for Simeon Poisson, is widely used to model the number of “random points” in a region of time or space, under certain ideal conditions. The Poisson distribution is studied in more detail in the chapter on the Poisson Process. The Poisson distribution with parameter $r \in (0, \infty)$ has probability density function f defined by

$$f(x) = e^{-r} \frac{r^x}{x!}, \quad x \in \mathbb{N} \quad (4.7.39)$$

The parameter r is the mean and variance of the distribution.

Suppose that X and Y are independent random variables, and that X has the Poisson distribution with parameter $a \in (0, \infty)$ and Y has the Poisson distribution with parameter $b \in (0, \infty)$. Let $N = X + Y$. Then

1. $\mathbb{E}(X | N) = \frac{a}{a+b} N$
2. $\text{var}(X | N) = \frac{ab}{(a+b)^2} N$
3. $\mathbb{E}([X - \mathbb{E}(X | N)]^2) = \frac{ab}{a+b}$

Proof

We have shown before that the distribution of N is also Poisson, with parameter $a + b$, and that the conditional distribution of X given $N = n \in \mathbb{N}$ is binomial with parameters n and $a/(a + b)$. Hence parts (a) and (b) follow from the standard formulas for the mean and variance of the binomial distribution, as functions of the parameters. Part (c) is the mean square error, and in this case can be computed most easily as

$$\mathbb{E}[\text{var}(X | N)] = \frac{ab}{(a+b)^2} \mathbb{E}(N) = \frac{ab}{(a+b)^2} (a+b) = \frac{ab}{a+b} \quad (4.7.40)$$

Once again, $\mathbb{E}(X | N)$ is a linear function of N and so $\mathbb{E}(X | N) = L(X | N)$. If we reverse the roles of the variables, the conditional expected value is trivial from our basic properties:

$$\mathbb{E}(N | X) = \mathbb{E}(X + Y | X) = X + b \quad (4.7.41)$$

Coins and Dice

A pair of fair dice are thrown, and the scores (X_1, X_2) recorded. Let $Y = X_1 + X_2$ denote the sum of the scores and $U = \min\{X_1, X_2\}$ the minimum score. Find each of the following:

1. $\mathbb{E}(Y | X_1)$
2. $\mathbb{E}(U | X_1)$
3. $\mathbb{E}(Y | U)$
4. $\mathbb{E}(X_2 | X_1)$

Answer

1. $\frac{7}{2} + X_1$

2. x	1	2	3	4	5	6
$\mathbb{E}(U X_1 = x)$	1	$\frac{11}{6}$	$\frac{5}{2}$	3	$\frac{10}{3}$	$\frac{7}{2}$

3. u	1	2	3	4	5	6
$\mathbb{E}(Y U = u)$	$\frac{52}{11}$	$\frac{56}{9}$	$\frac{54}{7}$	$\frac{46}{5}$	$\frac{32}{3}$	12

4. $\frac{7}{2}$

A box contains 10 coins, labeled 0 to 9. The probability of heads for coin i is $\frac{i}{9}$. A coin is chosen at random from the box and tossed. Find the probability of heads.

Answer

$$\frac{1}{2}$$

This problem is an example of Laplace's rule of succession, named for Pierre Simon Laplace.

Random Sums of Random Variables

Suppose that $\mathbf{X} = (X_1, X_2, \dots)$ is a sequence of independent and identically distributed real-valued random variables. We will denote the common mean, variance, and moment generating function, respectively, by $\mu = \mathbb{E}(X_i)$, $\sigma^2 = \text{var}(X_i)$, and $G(t) = \mathbb{E}(e^{tX_i})$. Let

$$Y_n = \sum_{i=1}^n X_i, \quad n \in \mathbb{N} \quad (4.7.42)$$

so that (Y_0, Y_1, \dots) is the partial sum process associated with \mathbf{X} . Suppose now that N is a random variable taking values in \mathbb{N} , independent of \mathbf{X} . Then

$$Y_N = \sum_{i=1}^N X_i \quad (4.7.43)$$

is a random sum of random variables; the terms in the sum are random, and the number of terms is random. This type of variable occurs in many different contexts. For example, N might represent the number of customers who enter a store in a given period of time, and X_i the amount spent by the customer i , so that Y_N is the total revenue of the store during the period.

The conditional and ordinary expected value of Y_N are

1. $\mathbb{E}(Y_N | N) = N\mu$
2. $\mathbb{E}(Y_N) = \mathbb{E}(N)\mu$

Proof

1. Using the substitution rule and the independence of N and \mathbf{X} we have

$$\mathbb{E}(Y_N | N = n) = \mathbb{E}(Y_n | N = n) = \mathbb{E}(Y_n) = \sum_{i=1}^n \mathbb{E}(X_i) = n\mu \quad (4.7.44)$$

so $\mathbb{E}(Y_N | N) = N\mu$.

2. From (a) and conditioning, $\mathbb{E}(Y_N) = \mathbb{E}[\mathbb{E}(Y_N | N)] = \mathbb{E}(N\mu) = \mathbb{E}(N)\mu$.

Wald's equation, named for Abraham Wald, is a generalization of the previous result to the case where N is not necessarily independent of \mathbf{X} , but rather is a *stopping time* for \mathbf{X} . Roughly, this means that the event $N = n$ depends only (X_1, X_2, \dots, X_n) . Wald's equation is discussed in the chapter on Random Samples. An elegant proof of and Wald's equation is given in the chapter on Martingales. The advanced section on stopping times is in the chapter on Probability Measures.

The conditional and ordinary variance of Y_N are

1. $\text{var}(Y_N | N) = N\sigma^2$
2. $\text{var}(Y_N) = \mathbb{E}(N)\sigma^2 + \text{var}(N)\mu^2$

Proof

1. Using the substitution rule, the independence of N and \mathbf{X} , and the fact that \mathbf{X} is an IID sequence, we have

$$\text{var}(Y_N | N = n) = \text{var}(Y_n | N = n) = \text{var}(Y_n) = \sum_{i=1}^n \text{var}(X_i) = n\sigma^2 \quad (4.7.45)$$

so $\text{var}(Y_N | N) = N\sigma^2$.

2. From (a) and the previous result,

$$\text{var}(Y_N) = \mathbb{E}[\text{var}(Y_N | N)] + \text{var}[\mathbb{E}(Y_N | N)] = \mathbb{E}(\sigma^2 N) + \text{var}(\mu N) = \mathbb{E}(N)\sigma^2 + \mu^2 \text{var}(N) \quad (4.7.46)$$

Let H denote the probability generating function of N . The conditional and ordinary moment generating function of Y_N are

1. $\mathbb{E}(e^{tY_N} | N) = [G(t)]^N$
2. $\mathbb{E}(e^{tY_N}) = H(G(t))$

Proof

1. Using the substitution rule, the independence of N and \mathbf{X} , and the fact that \mathbf{X} is an IID sequence, we have

$$\mathbb{E}(e^{tY_N} | N = n) = \mathbb{E}(e^{tY_n} | N = n) = \mathbb{E}(e^{tY_n}) = [G(t)]^n \quad (4.7.47)$$

(Recall that the MGF of the sum of independent variables is the product of the individual MGFs.)

2. From (a) and conditioning, $\mathbb{E}(e^{tY_N}) = \mathbb{E}[\mathbb{E}(e^{tY_N} | N)] = \mathbb{E}(G(t)^N) = H(G(t))$.

Thus the moment generating function of Y_N is $H \circ G$, the composition of the probability generating function of N with the common moment generating function of \mathbf{X} , a simple and elegant result.

In the die-coin experiment, a fair die is rolled and then a fair coin is tossed the number of times showing on the die. Let N denote the die score and Y the number of heads. Find each of the following:

1. The conditional distribution of Y given N .
2. $\mathbb{E}(Y | N)$
3. $\text{var}(Y | N)$
4. $\mathbb{E}(Y_i)$
5. $\text{var}(Y)$

Answer

1. Binomial with parameters N and $p = \frac{1}{2}$
2. $\frac{1}{2}N$
3. $\frac{1}{4}N$
4. $\frac{7}{4}$
5. $\frac{7}{3}$

Run the die-coin experiment 1000 times and compare the empirical mean and standard deviation to the distribution mean and standard deviation.

The number of customers entering a store in a given hour is a random variable with mean 20 and standard deviation 3. Each customer, independently of the others, spends a random amount of money with mean \$50 and standard deviation \$5. Find the mean and standard deviation of the amount of money spent during the hour.

Answer

1. \$1000
2. \$30.82

A coin has a random probability of heads V and is tossed a random number of times N . Suppose that V is uniformly distributed on $[0, 1]$; N has the Poisson distribution with parameter $a > 0$; and V and N are independent. Let Y denote the number of heads. Compute the following:

1. $\mathbb{E}(Y \mid N, V)$
2. $\mathbb{E}(Y \mid N)$
3. $\mathbb{E}(Y \mid V)$
4. $\mathbb{E}(Y)$
5. $\text{var}(Y \mid N, V)$
6. $\text{var}(Y)$

Answer

1. NV
2. $\frac{1}{2}N$
3. aV
4. $\frac{1}{2}a$
5. $NV(1 - V)$
6. $\frac{1}{12}a^2 + \frac{1}{2}a$

Mixtures of Distributions

Suppose that $\mathbf{X} = (X_1, X_2, \dots)$ is a sequence of real-valued random variables. Denote the mean, variance, and moment generating function of X_i by $\mu_i = \mathbb{E}(X_i)$, $\sigma_i^2 = \text{var}(X_i)$, and $M_i(t) = \mathbb{E}(e^{tX_i})$, for $i \in \mathbb{N}_+$. Suppose also that N is a random variable taking values in \mathbb{N}_+ , independent of \mathbf{X} . Denote the probability density function of N by $p_n = \mathbb{P}(N = n)$ for $n \in \mathbb{N}_+$. The distribution of the random variable X_N is a mixture of the distributions of $\mathbf{X} = (X_1, X_2, \dots)$, with the distribution of N as the mixing distribution.

The conditional and ordinary expected value of X_N are

1. $\mathbb{E}(X_N \mid N) = \mu_N$
2. $\mathbb{E}(X_N) = \sum_{n=1}^{\infty} p_n \mu_n$

Proof

1. Using the substitution rule and the independence of N and \mathbf{X} , we have $\mathbb{E}(X_N \mid N = n) = \mathbb{E}(X_n \mid N = n) = \mathbb{E}(X_n) = \mu_n$
2. From (a) and the conditioning rule,

$$\mathbb{E}(X_N) = \mathbb{E}[\mathbb{E}(X_N)] = \mathbb{E}(\mu_N) = \sum_{n=1}^{\infty} p_n \mu_n \quad (4.7.48)$$

The conditional and ordinary variance of X_N are

1. $\text{var}(X_N \mid N) = \sigma_N^2$
2. $\text{var}(X_N) = \sum_{n=1}^{\infty} p_n (\sigma_n^2 + \mu_n^2) - (\sum_{n=1}^{\infty} p_n \mu_n)^2$.

Proof

1. Using the substitution rule and the independence of N and \mathbf{X} , we have $\text{var}(X_N \mid N = n) = \text{var}(X_n \mid N = n) = \text{var}(X_n) = \sigma_n^2$
2. From (a) we have

$$\text{var}(X_N) = \mathbb{E}[\text{var}(X_N \mid N)] + \text{var}[\mathbb{E}(X_N \mid N)] = \mathbb{E}(\sigma_N^2) + \text{var}(\mu_N) = \mathbb{E}(\sigma_N^2) + \mathbb{E}(\mu_N^2) - [\mathbb{E}(\mu_N)]^2 \quad (4.7.49)$$

$$= \sum_{n=1}^{\infty} p_n \sigma_n^2 + \sum_{n=1}^{\infty} p_n \mu_n^2 - \left(\sum_{n=1}^{\infty} p_n \mu_n \right)^2 \quad (4.7.50)$$

The conditional and ordinary moment generating function of X_N are

1. $\mathbb{E}(e^{tX_N} \mid N) = M_N(t)$
2. $\mathbb{E}(e^{tX_N}) = \sum_{i=1}^{\infty} p_i M_i(t)$.

Proof

1. Using the substitution rule and the independence of N and \mathbf{X} , we have $\mathbb{E}(e^{tX_N} \mid N = n) = \mathbb{E}(e^{tX_n} \mid N = n) = \mathbb{E}(e^{tX_n}) = M_n(t)$

2. From (a) and the conditioning rule, $\mathbb{E}(e^{tX_N}) = \mathbb{E}[\mathbb{E}(e^{tX_N} | N)] = \mathbb{E}[M_N(t)] = \sum_{n=1}^{\infty} p_n M_n(t)$

In the coin-die experiment, a biased coin is tossed with probability of heads $\frac{1}{3}$. If the coin lands tails, a fair die is rolled; if the coin lands heads, an ace-six flat die is rolled (faces 1 and 6 have probability $\frac{1}{4}$ each, and faces 2, 3, 4, 5 have probability $\frac{1}{8}$ each). Find the mean and standard deviation of the die score.

Answer

1. $\frac{7}{2}$
2. 1.8634

Run the coin-die experiment 1000 times and note the apparent convergence of the empirical mean and standard deviation to the distribution mean and standard deviation.

This page titled [4.7: Conditional Expected Value](#) is shared under a [CC BY 2.0](#) license and was authored, remixed, and/or curated by [Kyle Siegrist \(Random Services\)](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.