5. Covariance and Correlation

Recall that by taking the expected value of various transformations of a random variable, we can measure many interesting characteristics of the distribution of the variable. In this section, we will study an expected value that measures a special type of relationship between two real-valued variables. This relationship is very important both in probability and statistics.

Basic Theory

Definitions

As usual, our starting point is a random experiment with probability measure $\mathbb{P}$ on an underlying sample space. Unless otherwise noted, we assume that all expected values mentioned in this section exist. Suppose now that X and Y are real-valued random variables for the experiment with means $\mathbb{E}(X)$, $\mathbb{E}(Y)$ and variances $\text{var}(X)$, $\text{var}(Y)$, respectively.

1. The *covariance* of (X, Y) is defined by

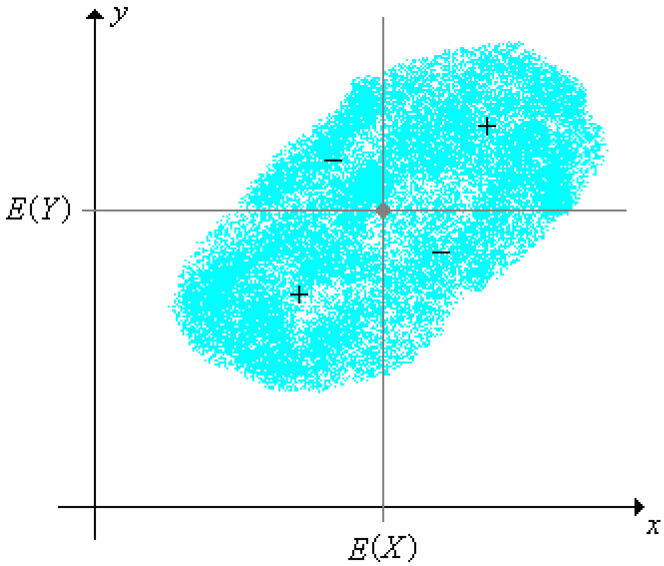
$$\text{cov}(X, Y) = \mathbb{E}\left([X - \mathbb{E}(X)][Y - \mathbb{E}(Y)]\right)$$

and, assuming the variances are positive, the *correlation* of (X, Y) is defined by

$$\text{cor}(X, Y) = \frac{\text{cov}(X, Y)}{\text{sd}(X)\text{sd}(Y)}$$

Correlation is a scaled version of covariance; note that the two parameters always have the same sign (positive, negative, or 0). When the sign is positive, the variables are said to be *positively correlated*; when the sign is negative, the variables are said to be *negatively correlated*; and when the sign is 0, the variables are said to be *uncorrelated*. Note also that correlation is dimensionless, since the numerator and denominator have the same physical units, namely the product of the units of X and Y.

As these terms suggest, covariance and correlation measure a certain kind of dependence between the variables. One of our goals is a deep understanding of this dependence. As a start, note that $(\mathbb{E}(X), \mathbb{E}(Y))$ is the center of the joint distribution of (X, Y), and the vertical and horizontal lines through this point separate $\mathbb{R}$ into four quadrants. The function $(x, y) \mapsto [x - \mathbb{E}(X)][y - \mathbb{E}(Y)]$ is positive on the first and third of these quadrants and negative on the second and fourth.



A joint distribution with $(\mathbb{E}(X), \mathbb{E}(Y))$ as the center of mass

Properties of Covariance

The following theorems give some basic properties of covariance. The main tool that we will need is the fact that expected value is a linear operation. Other important properties will be derived below, in the subsection on the best linear predictor. As usual, be sure to try the proofs yourself before reading the ones in the text.

Our first result is a formula that is better than the definition for computational purposes

2. $\text{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y)$.

▶ *Proof:*

By the previous result, we see that X and Y are uncorrelated if and only if $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. In particular, if X and Y are independent, then they are uncorrelated. However, the converse fails with a passion: an exercise below gives an example of two variables that are *functionally related* (the strongest form of dependence), yet uncorrelated. The computational exercises give other examples of dependent yet uncorrelated variables also. Note also that if one of the variables has mean 0, then the covariance is simply the expected product.

Trivially, covariance is a symmetric operation.

3. $\text{cov}(X, Y) = \text{cov}(Y, X)$.

As the name suggests, covariance generalizes variance.

4. $\text{cov}(X, X) = \text{var}(X)$.

▶ *Proof:*

Covariance is a linear operation in the first argument, if the second argument is fixed.

5. If X, Y, Z are real-valued random variables for the experiment, and c is a constant, then

a. $\text{cov}(X + Y, Z) = \text{cov}(X, Z) + \text{cov}(Y, Z)$
b. $\text{cov}(cX, Y) = c\,\text{cov}(X, Y)$

▶ *Proof:*

By symmetry, covariance is also a linear operation in the second argument, with the first argument fixed. Thus, the covariance operator is *bi-linear*. The general version of this property is given in the following theorem.

6. Suppose that $(X_1, X_2, \ldots, X_n)$ and $(Y_1, Y_2, \ldots, Y_m)$ are sequences of real-valued random variables for the experiment, and that $(a_1, a_2, \ldots, a_n)$ and $(b_1, b_2, \ldots, b_m)$ are constants. Then

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

The following result shows how covariance is changed under a linear transformation of one of the variables. This is simply a special case of the basic properties, but is worth stating.

7. If $a,\ b \in \mathbb{R}$ then $\operatorname{cov}(a + bX, Y) = b \operatorname{cov}(X, Y)$.

▶ *Proof:*

Of course, by symmetry, the same property holds in the second argument. Putting the two together we have that if $a,\ b,\ c,\ d \in \mathbb{R}$ then $\operatorname{cov}(a + bX, c + dY) = bd \operatorname{cov}(X, Y)$.

Properties of Correlation

Next we will establish some basic properties of correlation. Most of these follow easily from corresponding properties of covariance above. We assume that $\operatorname{var}(X) > 0$ and $\operatorname{var}(Y) > 0$, so that the random variable really are random and hence the correlation is well defined.

8. The correlation between X and Y is the covariance of the corresponding standard scores:

$$\operatorname{cor}(X, Y) = \operatorname{cov}\left(\frac{X - \mathbb{E}(X)}{\operatorname{sd}(X)}, \frac{Y - \mathbb{E}(Y)}{\operatorname{sd}(Y)}\right) = \mathbb{E}\left(\frac{X - \mathbb{E}(X)}{\operatorname{sd}(X)} \frac{Y - \mathbb{E}(Y)}{\operatorname{sd}(Y)}\right)$$

▶ *Proof:*

This shows again that correlation is dimensionless, since of course, the standard scores are dimensionless. Also, correlation is symmetric:

9. $\operatorname{cor}(X, Y) = \operatorname{cor}(Y, X)$.

Under a linear transformation of one of the variables, the correlation is unchanged if the slope is positve and changes sign if the slope is negative:

10. If $a,\ b \in \mathbb{R}$ and $b \ne 0$ then

a. $\operatorname{cor}(a + bX, Y) = \operatorname{cor}(X, Y)$ if $b > 0$
b. $\operatorname{cor}(a + bX, Y) = -\operatorname{cor}(X, Y)$ if $b < 0$

▶ *Proof:*

This result reinforces the fact that correlation is a standardized measure of association, since multiplying the variable by a positive constant is equivalent to a change of scale, and adding a contant to a variable is equivalent to a change of location. For example, in the Challenger data, the underlying variables are temperature at the time of launch (in degrees Fahrenheit) and O-ring erosion (in millimeters). The correlation between these two variables is of fundamental importance. If we decide to measure temperature in degrees

Celsius and O-ring erosion in inches, the correlation is unchanged. Of course, the same property holds in the second argument, so if $a, b, c, d \in \mathbb{R}$ with $b \neq 0$ and $d \neq 0$, then $\text{cor}(a + bX, c + dY) = \text{cor}(X, Y)$ if $bd > 0$ and $\text{cor}(a + bX, c + dY) = -\text{cor}(X, Y)$ if $bd < 0$.

The most important properties of covariance and correlation will emerge from our study of the best linear predictor below.

The Variance of a Sum

We will now show that the variance of a sum of variables is the sum of the pairwise covariances. This result is very useful since many random variables with common distributions can be written as sums of simpler random variables (see in particular the binomial distribution and hypergeometric distribution below).

11. If $(X_1, X_2, \ldots, X_n)$ is a sequence of real-valued random variables for the experiment, then

$$\text{var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \sum_{j=1}^n \text{cov}(X_i, X_j) = \sum_{i=1}^n \text{var}(X_i) + 2 \sum_{\{\{i,j\}:\ i<j\}} \text{cov}(X_i, X_j)$$

▶ *Proof:*

Note that the variance of a sum can be greater, smaller, or equal to the sum of the variances, depending on the pure covariance terms. As a special case of the previous result, when $n = 2$, we have

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\,\text{cov}(X, Y)$$

12. If $(X_1, X_2, \ldots, X_n)$ is a sequence of pairwise uncorrelated, real-valued random variables then

$$\text{var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{var}(X_i)$$

▶ *Proof:*

Note that the last result holds, in particular, if the random variables are independent.

13. If X and Y are real-valued random variables then $\text{var}(X + Y) + \text{var}(X - Y) = 2\,[\text{var}(X) + \text{var}(Y)]$.

▶ *Proof:*

14. If X and Y are real-valued random variables with $\text{var}(X) = \text{var}(Y)$ then $X + Y$ and $X - Y$ are uncorrelated.

▶ *Proof:*

Random Samples

In the following exercises, suppose that $(X_1, X_2, \ldots)$ is a sequence of independent, real-valued random variables with a common distribution that has mean μ and standard deviation $\sigma > 0$. In statistical terms, the variables form a random sample from the common distribution.

15. Let $Y_n = \sum_{i=1}^n X_i$.

a. $\mathbb{E}(Y_n) = n\,\mu$
b. $\text{var}(Y_n) = n\,\sigma^2$

▶ *Proof:*

16. Let $M_n = Y_n \big/ n = \frac{1}{n} \sum_{i=1}^n X_i$. Thus, M_n is the sample mean.

a. $\mathbb{E}(M_n) = \mu$
b. $\text{var}(M_n) = \sigma^2/n$
c. $\text{var}(M_n) \to 0$ as $n \to \infty$
d. $\mathbb{P}(|M_n - \mu| > \epsilon) \to 0$ as $n \to \infty$ for every $\epsilon > 0$.

▶ *Proof:*

Part (c) of the last exercise means that $M_n \to \mu$ as $n \to \infty$ in mean square. Part (d) means that $M_n \to \mu$ as $n \to \infty$ in probability. These are both versions of the weak law of large numbers, one of the fundamental theorems of probability.

17. The standard score of the sum Y_n and the standard score of the sample mean M_n are the same:

$$Z_n = \frac{Y_n - n\,\mu}{\sqrt{n}\,\sigma} = \frac{M_n - \mu}{\sigma/\sqrt{n}}$$

a. $\mathbb{E}(Z_n) = 0$
b. $\text{var}(Z_n) = 1$

▶ *Proof:*

The central limit theorem, the other fundamental theorem of probability, states that the distribution of Z_n converges to the standard normal distribution as $n \to \infty$.

Events

Suppose that A and B are events in a random experiment. The covariance and correlation of A and B are defined to be the covariance and correlation, respectively, of their indicator random variables $\mathbf{1}_A$ and $\mathbf{1}_B$.

18. If A and B are events then

a. $\text{cov}(A, B) = \mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)$
b. $\text{cor}(A, B) = [\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)] \big/ \sqrt{\mathbb{P}(A)\,[1 - \mathbb{P}(A)]\,\mathbb{P}(B)\,[1 - \mathbb{P}(B)]}$

▶ *Proof:*

In particular, note that A and B are positively correlated, negatively correlated, or independent, respectively (as defined in the section on conditional probability) if and only if the indicator variables of A and B are positively correlated, negatively correlated, or uncorrelated, as defined in this section.

19. If A and B are events then

a. $\text{cov}(A, B^c) = -\text{cov}(A, B)$
b. $\text{cov}(A^c, B^c) = \text{cov}(A, B)$

▶ *Proof:*

20. If $A \subseteq B$ then

a. $\text{cov}(A, B) = \mathbb{P}(A)[1 - \mathbb{P}(B)]$
b. $\text{cor}(A, B) = \sqrt{\mathbb{P}(A)\,[1 - \mathbb{P}(B)] \big/ \mathbb{P}(B)\,[1 - \mathbb{P}(A)]}$

▶ *Proof:*

The Best Linear Predictor

What linear function of X (that is, a function of the form $a + bX$ where $a, b \in \mathbb{R}$) is closest to Y in the sense of minimizing mean square error? The question is fundamentally important in the case where random variable X (the *predictor variable*) is observable and random variable Y (the *response variable*) is not. The linear function can be used to estimate Y from an observed value of X. Moreover, the solution will have the added benefit of showing that covariance and correlation measure the *linear* relationship between X and Y. To avoid trivial cases, let us assume that $\text{var}(X) > 0$ and $\text{var}(Y) > 0$, so that the random variables really are random.

The solution to our problem turns out to be the linear function of X with the same expected value as Y, and whose covariance with X is the same as that of Y. Define

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

21. $L(Y \mid X)$ is the only linear function of X that satisfies

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

▶ *Proof:*

Note that in the presence of part (a), part (b) is equivalent to $\mathbb{E}[XL(Y \mid X)] = \mathbb{E}(XY)$. Here is another minor variation, but one that will be very useful: $L(Y \mid X)$ is the only linear function of X with the same mean as Y and with the property that $Y - L(Y \mid X)$ is uncorrelated with every linear function of X.

22. $L(Y \mid X)$ is the only linear function of X that satisfies

a. $\mathbb{E}[L(Y \mid X)] = \mathbb{E}(Y)$
b. $\text{cov}[Y - L(Y \mid X), U] = 0$ for every linear function U of X.

▶ *Proof:*

The variance of $L(Y \mid X)$ and its covariance with Y turn out to be the same.

23. Additional properties of $L(Y \mid X)$:

a. $\text{var}[L(Y \mid X)] = \text{cov}^2(X, Y) \big/ \text{var}(X)$
b. $\text{cov}[L(Y \mid X), Y] = \text{cov}^2(X, Y) \big/ \text{var}(X)$

▶ *Proof:*

We can now prove the fundamental result that $L(Y \mid X)$ is the linear function of X that is closest to Y in the mean square sense. We give two proofs; the first is more straightforward, but the second is more interesting and elegant.

24. Suppose that U is a linear function of X. Then

a. $\mathbb{E}\left([Y - L(Y \mid X)]^2\right) \le \mathbb{E}\left[(Y - U)^2\right]$
b. Equality occurs in (a) if and only if $U = L(Y \mid X)$ with probability 1.

▶ *Proof from calculus:*

▶ *Proof using properties:*

25. The mean square error when $L(Y \mid X)$ is used as a predictor of Y is

$$\mathbb{E}\left([Y - L(Y \mid X)]^2\right) = \text{var}(Y)\left[1 - \text{cor}^2(X, Y)\right]$$

▶ *Proof:*

Our solution to the best linear perdictor problems yields important properties of covariance and correlation.

26. Additional properties of covariance and correlation:

a. $-1 \le \text{cor}(X, Y) \le 1$
b. $-\text{sd}(X)\text{sd}(Y) \le \text{cov}(X, Y) \le \text{sd}(X)\text{sd}(Y)$
c. $\text{cor}(X, Y) = 1$ if and only if Y is a linear function of X with positive sloper.
d. $\text{cor}(X, Y) = -1$ if and only if Y is a linear function of X with negative slope.

▶ *Proof:*

The last two results clearly show that $\text{cov}(X, Y)$ and $\text{cor}(X, Y)$ measure the *linear* association between X and Y. The equivalent inequalities (a) and (b) above are referred to as the *correlation inequality*. They are also versions of the *Cauchy Schwarz inequality*, named for Augustin Cauchy and Karl Schwarz

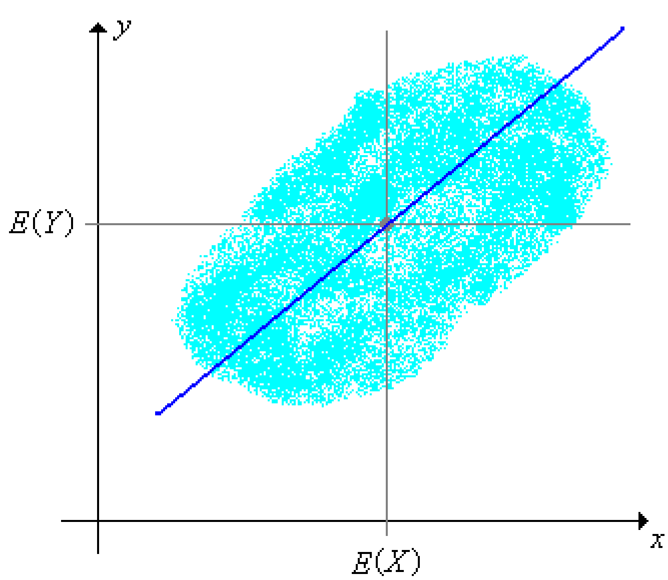
Recall from our previous discussion of variance that the best *constant* predictor of Y, in the sense of minimizing mean square error, is $\mathbb{E}(Y)$ and the minimum value of the mean square error for this predictor is $\text{var}(Y)$. Thus, the difference between the variance of Y and the mean square error above for $L(Y \mid X)$ is the reduction in the variance of Y when the linear term in X is added to the predictor:

$$\text{var}(Y) - \mathbb{E}\left([Y - L(Y \mid X)]^2\right) = \text{var}(Y)\,\text{cor}^2(X, Y)$$

Thus $\text{cor}^2(X, Y)$ is the *proportion* of reduction in $\text{var}(Y)$ when X is included as a predictor variable. This quantity is called the (distribution) *coefficient of determination*. Now let

$$L(Y \mid X = x) = \mathbb{E}(Y) + \frac{\text{cov}(X, Y)}{\text{var}(X)}[x - \mathbb{E}(X)], \quad x \in \mathbb{R}$$

The function $x \mapsto L(Y \mid X = x)$ is known as the *distribution regression function* for Y given X, and its graph is known as the *distribution regression line*. Note that the regression line passes through $(\mathbb{E}(X), \mathbb{E}(Y))$, the center of the joint distribution.



The distribution regression line

However, the choice of predictor variable and response variable is crucial.

27. The regression line for Y given X and the regression line for X given Y are not the same line, except in the trivial case where the variables are perfectly correlated. However, the coefficient of determination is the same, regardless of which variable is the predictor and which is the response.

▶ *Proof:*

28. Suppose that A and B are events in a random experiment with $0 < \mathbb{P}(A) < 1$ and $0 < \mathbb{P}(B) < 1$. Then

a. $\text{cor}(A, B) = 1$ if and only $\mathbb{P}(A \setminus B) + \mathbb{P}(B \setminus A) = 0$. (That is, A and B are equivalent events.)
b. $\text{cor}(A, B) = -1$ if and only $\mathbb{P}(A \setminus B^c) + \mathbb{P}(B^c \setminus A) = 0$. (That is, A and B^c are equivalent events.)

The concept of best linear predictor is more powerful than might first appear, because it can be applied to *transformations* of the variables. Specifically, suppose that X and Y are random variables for our experiment, taking values in general spaces S and T, respectively. Suppose also that g and h are real-valued functions defined on S and T, respectively. We can find $L[h(Y) \mid g(X)]$, the linear function of $g(X)$ that is closest to $h(Y)$ in the mean square sense. The results of this subsection apply, of course, with $g(X)$ replacing X and $h(Y)$ replacing Y. Of course, we must be able to compute the appropriate means, variances, and covariances.

We close this subsection with two additional properties of the best linear predictor, the *linearity properties*.

29. Suppose that X, Y, and Z are real-valued random variables for the experiment, and that c is a constant. Then

a. $L(Y + Z \mid X) = L(Y \mid X) + L(Z \mid X)$
b. $L(c\,Y \mid X) = c\,L(Y \mid X)$

▶ *Proof from the definitions:*

▶ *Proof by characterizing properties:*

There are several extensions and generalizations of the ideas in the subsection:

- The corresponding statistical problem of estimating a and b, when these distribution parameters are unknown, is considered in the section on Sample Covariance and Correlation.
- The problem finding the function of X (using *all* reasonable functions, not just linear ones) that is closest to Y in the mean square error sense is considered in the section on Conditional Expected Value.
- The best linear prediction problem when the predictor and response variables are random vectors is considered in the section on Expected Value and Covariance Matrices.

The use of characterizing properties will play a crucial role in these extensions.

Examples and Applications

Uniform Distributions

30. Suppose that X is uniformly distributed on the interval $[-1, 1]$ and $Y = X^2$. Then X and Y are uncorrelated even though *Y is a function of X* (the strongest form of dependence).

▶ *Proof:*

31. Suppose that (X, Y) is uniformly distributed on the region $S \subseteq \mathbb{R}^2$. Find $\text{cov}(X, Y)$ and $\text{cor}(X, Y)$ and determine whether the variables are independent in each of the following cases:

a. $S = [a, b] \times [c, d]$ where $a < b$ and $c < d$, so S is a rectangle.

b. $S = \{(x, y) \in \mathbb{R}^2 : -a \le y \le x \le a\}$ where $a > 0$, so S is a triangle
c. $S = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \le r^2\}$ where $r > 0$, so S is a circle

▶ *Answer:*

32. In the bivariate uniform experiment, select each of the regions below in turn. For each region, run the simulation 2000 times and note the value of the correlation and the shape of the cloud of points in the scatterplot. Compare with the results in the last exercise.

a. Square
b. Triangle
c. Circle

33. Suppose that X is uniformly distributed on the interval $(0, 1)$ and that given $X = x \in (0, 1)$, Y is uniformly distributed on the interval $(0, x)$. Find each of the following:

a. $\text{cov}(X, Y)$
b. $\text{cor}(X, Y)$
c. $L(Y \mid X)$
d. $L(X \mid Y)$

▶ *Answer:*

Dice

Recall that a *standard die* is a six-sided die. A *fair* die is one in which the faces are equally likely. An *ace-six flat die* is a standard die in which faces 1 and 6 have probability $\frac{1}{4}$ each, and faces 2, 3, 4, and 5 have probability $\frac{1}{8}$ each.

34. A pair of standard, fair dice are thrown and the scores (X_1, X_2) recorded. Let $Y = X_1 + X_2$ denote the sum of the scores, $U = \min\{X_1, X_2\}$ the minimum scores, and $V = \max\{X_1, X_2\}$ the maximum score. Find the covariance and correlation of each of the following pairs of variables:

a. (X_1, X_2)
b. (X_1, Y)
c. (X_1, U)
d. (U, V)
e. (U, Y)

▶ *Answer:*

35. Suppose that n fair dice are thrown. Find the mean and variance of each of the following variables:

a. Y_n, the sum of the scores.
b. M_n, the average of the scores.

▶ *Answer:*

36. In the dice experiment, select fair dice, and select the following random variables. In each case, increase the number of dice and observe the size and location of the probability density function and the mean $\pm$ standard deviation bar. With $n = 20$ dice, run the experiment 1000 times and compare the sample mean and standard deviation to the distribution mean and standard deviation.

a. The sum of the scores.

b. The average of the scores.

37. Suppose that n ace-six flat dice are thrown. Find the mean and variance of each of the following variables:

a. Y_n, the sum of the scores.
b. M_n, the average of the scores.

▶ *Answer:*

38. In the dice experiment, select ace-six flat dice, and select the following random variables. In each case, increase the number of dice and observe the size and location of the probability density function and the mean $\pm$ standard deviation bar. With $n = 20$ dice, run the experiment 1000 times and compare the sample mean and standard deviation to the distribution mean and standard deviation.

a. The sum of the scores.
b. The average of the scores.

39. 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, $U = \min\{X_1, X_2\}$ the minimum score, and $V = \max\{X_1, X_2\}$ the maximum score. Find each of the following:

a. $L(Y \mid X_1)$
b. $L(U \mid X_1)$
c. $L(V \mid X_1)$

▶ *Answer:*

Bernoulli Trials

Recall that a *Bernoulli trials process* is a sequence $\boldsymbol{X} = (X_1, X_2, \ldots)$ of independent, identically distributed indicator random variables. In the usual language of reliability, X_i denotes the outcome of trial i, where 1 denotes success and 0 denotes failure. The probability of success $p = \mathbb{P}(X_i = 1)$ is the basic parameter of the process. The process is named for Jacob Bernoulli. A separate chapter on the Bernoulli Trials explores this process in detail.

The number of successes in the first n trials is $Y = \sum_{i=1}^n X_i$. Recall that this random variable has the binomial distribution with parameters n and p, which has probability density function

$$f(y) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y \in \{0, 1, \ldots, n\}$$

40. The mean and variance of Y are

a. $\mathbb{E}(Y) = np$
b. $\operatorname{var}(Y) = np(1-p)$

▶ *Proof:*

41. In the binomial coin experiment, select the number of heads. Vary n and p and note the shape of the probability density function and the size and location of the mean $\pm$ standard deviation bar. For selected values of the parameters, run the experiment 1000 times and compare the sample mean and standard deviation to the distribution mean and standard deviation.

The proportion of successes in the first n trials is $M = Y/n$. This random variable is sometimes used as a statistical estimator of the parameter p, when the parameter is unknown.

42. The mean and variance of M_n are

a. $\mathbb{E}(M) = p$
b. $\text{var}(M) = p(1-p)/n$

▶ *Proof:*

43. In the binomial coin experiment, select the proportion of heads. Vary n and p and note the shape of the probability density function and the size and location of the mean $\pm$ standard deviation bar. For selected values of the parameters, run the experiment 1000 times and compare the sample mean and standard deviation to the distribution mean and standard deviation.

The Hypergeometric Distribution

Suppose that a population consists of m objects; r of the objects are type 1 and $m-r$ are type 0. A sample of n objects is chosen at random, without replacement. Let X_i denote the type of the ith object selected. Recall that $(X_1, X_2, \ldots, X_n)$ is a sequence of identically distributed (but *not* independent) indicator random variables.

Let Y denote the number of type 1 objects in the sample, so that $Y = \sum_{i=1}^n X_i$. Recall that this random variable has the hypergeometric distribution, which has probability density function.

$$f(y) = \frac{\binom{r}{y}\binom{m-r}{n-y}}{\binom{m}{n}}, \quad y \in \{0, 1, \ldots, n\}$$

44. For distinct i and j,

a. $\mathbb{E}(X_i) = \frac{r}{m}$
b. $\text{var}(X_i) = \frac{r}{m}\left(1 - \frac{r}{m}\right)$
c. $\text{cov}(X_i, X_j) = -\frac{r}{m}\left(1 - \frac{r}{m}\right)\frac{1}{m-1}$
d. $\text{cor}(X_i, X_j) = -\frac{1}{m-1}$

▶ *Proof:*

Note that the event of a type 1 object on draw i and the event of a type 1 object on draw j are negatively correlated, but the correlation depends only on the population size and not on the number of type 1 objects. Note also that the correlation is perfect if $m = 2$. Think about these result intuitively.

45. The mean and variance of Y are

a. $\mathbb{E}(Y) = n\frac{r}{m}$
b. $\text{var}(Y) = n\frac{r}{m}(1 - \frac{r}{m})\frac{m-n}{m-1}$

▶ *Proof:*

46. In the ball and urn experiment, select sampling without replacement. Vary m, r, and n and note the shape of the probability density function and the size and location of the mean $\pm$ standard deviation bar. For selected values of the parameters, run the experiment 1000 times and compare the sample mean and standard deviation to the distribution mean and standard deviation.

Exercises on Basic Properties

47. Suppose that X and Y are real-valued random variables with $\text{cov}(X, Y) = 3$. Find $\text{cov}(2X - 5, 4Y + 2)$.

▶ *Answer:*

48. Suppose X and Y are real-valued random variables with $\text{var}(X) = 5$, $\text{var}(Y) = 9$, and $\text{cov}(X, Y) = -3$. Find

a. $\text{cor}(X, Y)$
b. $\text{var}(2X + 3Y - 7)$
c. $\text{cov}(5X + 2Y - 3, 3X - 4Y + 2)$
d. $\text{cor}(5X + 2Y - 3, 3X - 4Y + 2)$

▶ *Answer:*

49. Suppose that X and Y are independent, real-valued random variables with $\text{var}(X) = 6$ and $\text{var}(Y) = 8$. Find $\text{var}(3X - 4Y + 5)$.

▶ *Answer:*

50. Suppose that A and B are events in an experiment with $\mathbb{P}(A) = \frac{1}{2}$, $\mathbb{P}(B) = \frac{1}{3}$, and $\mathbb{P}(A \cap B) = \frac{1}{8}$. Find each of the following:

a. $\text{cov}(A, B)$
b. $\text{cor}(A, B)$

▶ *Answer:*

51. Suppose that X, Y, and Z are real-valued random variables for an experiment, and that $L(Y \mid X) = 2 - 3X$ and $L(Z \mid X) = 5 + 4X$. Find $L(6Y - 2Z \mid X)$.

▶ *Answer:*

52. Suppose that X and Y are real-valued random variables for an experiment, and that $\mathbb{E}(X) = 3$, $\text{var}(X) = 4$, and $L(Y \mid X) = 5 - 2X$. Find each of the following:

a. $\mathbb{E}(Y)$
b. $\text{cov}(X, Y)$

▶ *Answer:*

Simple Continuous Distributions

53. Suppose that (X, Y) has probability density function $f(x, y) = x + y$ for $0 \le x \le 1$, $0 \le y \le 1$. Find each of the following

a. $\text{cov}(X, Y)$
b. $\text{cor}(X, Y)$
c. $L(Y \mid X)$
d. $L(X \mid Y)$

▶ *Answer:*

54. Suppose that (X, Y) has probability density function $f(x, y) = 2(x + y)$ for $0 \le x \le y \le 1$. Find each of the following:

a. $\text{cov}(X, Y)$
b. $\text{cor}(X, Y)$
c. $L(Y \mid X)$
d. $L(X \mid Y)$

▶ *Answer:*

55. Suppose again that (X, Y) has probability density function $f(x, y) = 2(x + y)$ for $0 \le x \le y \le 1$.

a. Find $\text{cov}\left(X^2, Y\right)$.
b. Find $\text{cor}\left(X^2, Y\right)$.
c. Find $L\left(Y \mid X^2\right)$.
d. Which predictor of Y is better, the one based on X or the one based on X^2?

▶ *Answer:*

56. Suppose that (X, Y) has probability density function $f(x, y) = 6x^2 y$ for $0 \le x \le 1$, $0 \le y \le 1$. Find each of the following:

a. $\text{cov}(X, Y)$
b. $\text{cor}(X, Y)$
c. $L(Y \mid X)$
d. $L(X \mid Y)$

▶ *Answer:*

57. Suppose that (X, Y) has probability density function $f(x, y) = 15x^2 y$ for $0 \le x \le y \le 1$. Find each of the following:

a. $\text{cov}(X, Y)$
b. $\text{cor}(X, Y)$
c. $L(Y \mid X)$
d. $L(X \mid Y)$

▶ *Answer:*

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

a. Find $\text{cov}\left(\sqrt{X}, Y\right)$.
b. Find $\text{cor}\left(\sqrt{X}, Y\right)$.
c. Find $L\left(Y \mid \sqrt{X}\right)$.
d. Which of the predictors of Y is better, the one based on X of the one based on $\sqrt{X}$?

▶ *Answer:*