

3. 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](#). Suppose 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 (assumed finite). The **covariance** of X and Y is defined by

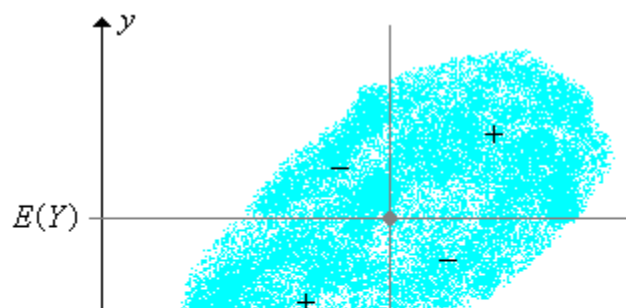
$$\text{cov}(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)))$$

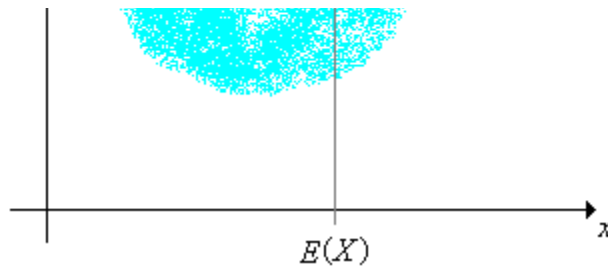
and (assuming the variances are positive) the **correlation** of X and Y is defined by

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

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.

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 line through this point separate \mathbb{R}^2 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.





Properties

The following exercises give some basic properties of covariance. The main tool that you 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](#).

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

By [Exercise 1](#), 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, as the following exercise shows. (Other examples of dependent yet uncorrelated variables occur in the computational exercises.)

2. Suppose that X is uniformly distributed on the interval $[-a, a]$, where $a > 0$, and $Y = X^2$. Show that X and Y are uncorrelated even though Y is a function of X (the strongest form of dependence).

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

4. Show that $\text{cov}(X, X) = \text{var}(X)$. Thus, covariance subsumes variance.

5. Show that $\text{cov}(aX + bY, Z) = a\text{cov}(X, Z) + b\text{cov}(Y, Z)$

6. Suppose that (X_1, X_2, \dots, X_n) and (Y_1, Y_2, \dots, Y_m) are sequences of real-valued random variables for an experiment. Prove the following property (known as **bi-linearity**).

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

7. Show that the correlation between X and Y is simply the covariance of the corresponding standard scores:

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

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).

8. Suppose that (X_1, X_2, \dots, X_n) is a sequence of real-valued random variables. Use [Exercise 3](#), [Exercise 4](#),

and [Exercise 5](#) to show that

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

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.

▣ 9. Suppose that (X_1, X_2, \dots, X_n) is a sequence of pairwise uncorrelated, real-valued random variables.

Show that

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

Note that the result in the previous exercise holds, in particular, if the random variables are mutually [independent](#).

▣ 10. Suppose that X and Y are real-valued random variables. Show that

$$\text{var}(X + Y) + \text{var}(X - Y) = 2 \text{var}(X) + 2 \text{var}(Y).$$

▣ 11. Suppose that X and Y are real-valued random variables with $\text{var}(X) = \text{var}(Y)$. Show that $X + Y$ and $X - Y$ are uncorrelated.

Random Samples

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

▣ 12. Let $Y_n = \sum_{i=1}^n X_i$. Show that

- $\mathbb{E}(Y_n) = n\mu$
- $\text{var}(Y_n) = n\sigma^2$

▣ 13. Let $M_n = \frac{Y_n}{n} = \frac{1}{n} \sum_{i=1}^n X_i$. Thus, M_n is the [sample mean](#). Show that

- $\mathbb{E}(M_n) = \mu$
- $\text{var}(M_n) = \mathbb{E}((M_n - \mu)^2) = \frac{\sigma^2}{n}$, so $\text{var}(M_n) \rightarrow 0$ as $n \rightarrow \infty$.
- $\mathbb{P}(|M_n - \mu| > \varepsilon) \rightarrow 0$ as $n \rightarrow \infty$ for any $\varepsilon > 0$. (*Hint:* Use [Chebyshev's inequality](#)).

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

▣ 14. Let $Z_n = \frac{Y_n - n\mu}{\sqrt{n}\sigma}$. Thus, Z_n is the standard score associated with Y_n . Show that

- a. $Z_n = \frac{M_n - \mu}{\sigma/\sqrt{n}}$ so that Z_n is also the standard score associated with M_n .
- b. $\mathbb{E}(Z_n) = 0$
- c. $\text{var}(Z_n) = 1$

The [central limit theorem](#), the other fundamental theorem of probability, states that the distribution of Z_n converges to the standard normal distribution as $n \rightarrow \infty$

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 $\mathbf{1}(B)$.

15. Show that

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

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.

16. Show that

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

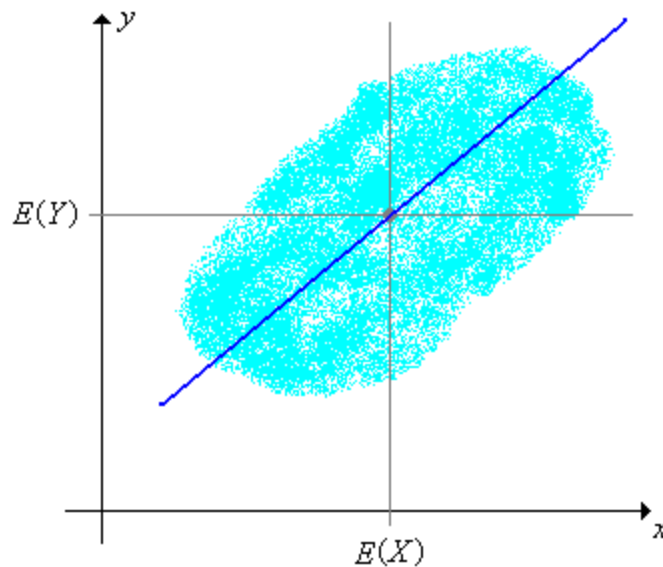
17. Suppose that $A \subseteq B$ Show that

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

The Best Linear Predictor

What linear function of X 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 show 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.



Let $\text{MSE}(a, b)$ denote the mean square error when $aX + b$ is used as an estimator of Y (as a function of the parameters a and b):

$$\text{MSE}(a, b) = \mathbb{E}((Y - (aX + b))^2)$$

18. Show that

$$\text{MSE}(a, b) = \mathbb{E}(Y^2) - 2a \mathbb{E}(XY) - 2b \mathbb{E}(Y) + a^2 \mathbb{E}(X^2) + 2ab \mathbb{E}(X) + b^2$$

19. Use basic calculus to show that $\text{MSE}(a, b)$ is minimized when

$$a = \frac{\text{cov}(X, Y)}{\text{var}(X)}, \quad b = \mathbb{E}(Y) - \frac{\text{cov}(X, Y)}{\text{var}(X)} \mathbb{E}(X)$$

Thus, the **best linear predictor** of Y given X is the random variable $L(Y|X)$ given by

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

20. Show that the minimum value of the mean square error function MSE , is

$$\mathbb{E}((Y - L(Y|X))^2) = \text{var}(Y) (1 - \text{cor}(X, Y)^2)$$

21. From the last exercise, verify the following important properties:

- $-1 \leq \text{cor}(X, Y) \leq 1$
- $-\text{sd}(X) \text{sd}(Y) \leq \text{cov}(X, Y) \leq \text{sd}(X) \text{sd}(Y)$
- $\text{cor}(X, Y) = 1$ if and only if $Y = aX + b$ with probability 1 for some constants $a > 0$ and b .
- $\text{cor}(X, Y) = -1$ if and only if $Y = aX + b$ with probability 1 for some constants $a < 0$ and b .

These exercises show clearly that $\text{cov}(X, Y)$ and $\text{cor}(X, Y)$ measure the *linear* association between X and Y .

Recall 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 in [Exercise 20](#) is the reduction in the variance of Y when the linear term in X is added to the predictor.

22. Show that $\text{var}(Y) - \mathbb{E}((Y - L(Y|X))^2) = \text{var}(Y) \text{cor}(X, Y)^2$. The fraction of the reduction is $\text{cor}(X, Y)^2$, and hence this quantity is called the (distribution) **coefficient of determination**.

Now let

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

The function $x \mapsto L(Y|X = x)$ 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.

23. Show that $\mathbb{E}(L(Y|X)) = \mathbb{E}(Y)$.

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

24. Show that 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.

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

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

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 $L(h(Y)|g(X))$, 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 .

26. Suppose that Z is another real-valued random variable for the experiment and that c is a constant. Show that

- $L(Y + Z|X) = L(Y|X) + L(Z|X)$

$$b. L(cY|X) = cL(Y|X)$$

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](#).

Examples and Applications

Uniform Distributions

27. 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:

- $S = [a, b] \times [c, d]$ where $a < b$ and $c < d$
- $S = \{(x, y) \in \mathbb{R}^2 : -a \leq y \leq x \leq a\}$ where $a > 0$.
- $S = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq r^2\}$ where $r > 0$.



28. In the **bivariate uniform experiment**, select each of the regions below in turn. For each region, run the simulation 2000 times, updating every 10 runs. 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.

- Square
- Triangle
- Circle

29. Suppose that X is uniformly distributed on the interval $(0, 1)$ and that given $X = x$, Y is uniformly distributed on the interval $(0, x)$.

- Find $\text{cov}(X, Y)$
- Find $\text{cor}(X, Y)$
- Find $L(Y|X)$.
- Find $L(X|Y)$



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.

30. 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:

- (X_1, X_2)
- (X_1, Y)
- (X_1, U)
- (U, V)
- (U, Y)



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

- The sum of the scores.
- The average of the scores.



32. In the **dice experiment**, select the following random variables. In each case, increase the number of dice and observe the size and location of the density function and the mean-standard deviation bar. With $n = 20$ dice, run the experiment 1000 times, updating every 10 runs, and note the apparent convergence of the empirical moments to the distribution moments.

- The sum of the scores.
- The average of the scores.

33. Repeat [Exercise 31](#) for ace-six flat dice.



34. Repeat [Exercise 32](#) for ace-six flat dice.

35. 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:

- $L(Y|X_1)$.
- $L(U|X_1)$.
- $L(V|X_1)$.



Bernoulli Trials

Recall that a **Bernoulli trials process** is a sequence (X_1, X_2, \dots) 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 **James Bernoulli**. A separate chapter on the **Bernoulli Trials** explores this process in detail.

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

$$\mathbb{P}(Y_n = k) = \binom{n}{k} p^k (1-p)^{n-k}, k \in \{0, 1, \dots, n\}$$

36. Show that

- $\mathbb{E}(Y_n) = n p$
- $\text{var}(Y_n) = n p (1 - p)$

37. In the **binomial coin experiment**, select the number of heads. Vary n and p and note the shape of the density function and the size and location of the mean-standard deviation bar. For selected values of n and p , run the experiment 1000 times, updating every 10 runs, and note the apparent convergence of the sample mean and standard deviation to the distribution mean and standard deviation.

The proportion of successes in the first n trials is $M_n = \frac{Y_n}{n}$. This random variable is sometimes used as a **statistical estimator** of the parameter p , when the parameter is unknown.

38. Show that

- $\mathbb{E}(M_n) = p$
- $\text{var}(M_n) = \frac{p(1-p)}{n}$

39. In the **binomial coin experiment**, select the proportion of heads. Vary n and p and note the shape of the density function and the size and location of the mean-standard deviation bar. For selected values of n and p , run the experiment 1000 times, updating every 10 runs, and note the apparent convergence of 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 i^{th} object selected. Recall that (X_1, X_2, \dots, X_n) is a sequence of identically distributed (but *not* independent) indicator random variables. In fact the sequence is [exchangeable](#).

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

$$\mathbb{P}(Y_n = k) = \frac{\binom{r}{k} \binom{m-r}{n-k}}{\binom{m}{n}}, \quad k \in \{0, 1, \dots, n\}$$

40. Show that for distinct i and j ,

- $\text{cov}(X_i, X_j) = -\frac{r}{m} \left(1 - \frac{r}{m}\right) \frac{1}{m-1}$
- $\text{cor}(X_i, X_j) = -\frac{1}{m-1}$

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.

41. Show that

- $\mathbb{E}(Y_n) = n \frac{r}{m}$
- $\text{var}(Y_n) = n \frac{r}{m} \left(1 - \frac{r}{m}\right) \frac{m-n}{m-1}$

42. In the [ball and urn experiment](#), select sampling without replacement. Vary m , r , and n and note the shape of the density function and the size and location of the mean-standard deviation bar. For selected values of the parameters, run the experiment 1000 times, updating every 10 runs, and note the apparent convergence of the sample mean and standard deviation to the distribution mean and standard deviation.

Miscellaneous Exercises

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



44. 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 $\text{var}(2X + 3Y - 7)$.



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

46. 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 the covariance and correlation between A and B .

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

- Find $\text{cov}(X, Y)$
- Find $\text{cor}(X, Y)$
- Find $L(Y|X)$.
- Find $L(X|Y)$.

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

- Find $\text{cov}(X, Y)$
- Find $\text{cor}(X, Y)$
- Find $L(Y|X)$.
- Find $L(X|Y)$.

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

- Find $\text{cov}(X^2, Y)$
- Find $\text{cor}(X^2, Y)$
- Find $L(Y|X^2)$.
- Which predictor of Y is better, the one based on X or the one based on X^2 ?

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

- Find $\text{cov}(X, Y)$
- Find $\text{cor}(X, Y)$
- Find $L(Y|X)$.
- Find $L(X|Y)$.

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

- Find $\text{cov}(X, Y)$

- b. Find $\text{cor}(X, Y)$
- c. Find $L(Y|X)$.
- d. Find $L(X|Y)$.



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

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



Vector Space Concepts

Covariance is closely related to the concept of inner product in the theory of vector spaces. This connection can help illustrate many of the properties of covariance from a different point of view.

In this section, our **vector space** \mathcal{V}_2 consists of all real-valued random variables defined on a fixed probability space $(\Omega, \mathcal{F}, \mathbb{P})$ (that is, relative to the same random experiment) that have finite second moment. Recall that two random variables are **equivalent** if they are equal with probability 1. As usual, we consider two such random variables as the same vector, so that technically, our vector space consists of **equivalence classes** under this **equivalence relation**. The **addition operator** corresponds to the usual addition of two real-valued random variables, and the operation of **scalar multiplication** corresponds to the usual multiplication of a real-valued random variable by a real (non-random) number.

Inner Product

If X and Y are random variables in \mathcal{V}_2 , we define the **inner product** of X and Y by

$$\langle X, Y \rangle = \mathbb{E}(XY)$$

The following exercise gives results that are analogs of the **basic properties of covariance** given above, and show that this definition really does give an inner product on the vector space

53. Show that

- $\langle X, Y \rangle = \langle Y, X \rangle$
- $\langle X, X \rangle \geq 0$
- $\langle X, X \rangle = 0$ if and only if $\mathbb{P}(X = 0) = 1$ (so that X is equivalent to 0).
- $\langle aX, Y \rangle = a \langle X, Y \rangle$ for any constant a .
- $\langle X + Y, Z \rangle = \langle X, Z \rangle + \langle Y, Z \rangle$

Covariance and correlation can easily be expressed in terms of this inner product. The covariance of two random variables is the inner product of the corresponding centered variables. The correlation is the inner product of the corresponding **standard scores**.

54. Show that

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

The **norm** associated with the inner product is the 2-norm studied in the last section, and corresponds to the **root mean square** operation on a random variable. This fact is a fundamental reason why the 2-norm plays such a special, honored role; of all the k -norms, only the 2-norm corresponds to an inner product. In turn, this is one of the reasons that **root mean square difference** is of fundamental importance in probability and statistics.

55. Show that $\langle X, X \rangle = \|X\|_2^2 = \mathbb{E}(X^2)$.

Projection

Let X and Y be random variables in \mathcal{V}_2

56. Show that the following set is a subspace of \mathcal{V}_2 . In fact, it is the subspace **generated** by X and 1.

$$\mathcal{W} = \{aX + b : (a \in \mathbb{R}) \text{ and } (b \in \mathbb{R})\}$$

57. Show that the best linear predictor of Y given X can be characterized as the **projection** of Y onto the subspace \mathcal{W} . That is, show that $L(Y|X)$ is the only random variable $W \in \mathcal{W}$ with the property that $Y - W$ is perpendicular to \mathcal{W} . Specifically, find W such that satisfies the following two conditions:

- $\langle Y - W, X \rangle = 0$
- $\langle Y - W, 1 \rangle = 0$

Hölder's Inequality

The next exercise gives **Hölder's inequality**, named for **Otto Hölder**.

58. Suppose that $j > 1$, $k > 1$, and $\frac{1}{j} + \frac{1}{k} = 1$. Show that $\langle |X|, |Y| \rangle \leq \|X\|_j \|Y\|_k$ using the steps below:
- Show that $S = \{(x, y) \in \mathbb{R}^2 : (x \geq 0) \text{ and } (y \geq 0)\}$ is a convex set and $g(x, y) = x^{1/j} y^{1/k}$ is concave on S .
 - Use (a) and Jensen's inequality to show that if U and V are nonnegative random variables then

$$\mathbb{E}(U^{1/j} V^{1/k}) \leq \mathbb{E}(U)^{1/j} \mathbb{E}(V)^{1/k}$$
 - In (b), let $U = |X|^j$ and $V = |Y|^k$

In the context of the last exercise, j and k are called **conjugate exponents**. If we let $j = k = 2$ in Hölder's inequality, then we get the **Cauchy-Schwarz inequality**, named for **Augustin Cauchy** and **Karl Schwarz**. In turn, this is equivalent to the inequalities in [Exercise 21](#).

$$\mathbb{E}(|X Y|) \leq \sqrt{\mathbb{E}(X^2)} \sqrt{\mathbb{E}(Y^2)}$$

59. Suppose that (X, Y) has probability density function $f(x, y) = x + y$, $0 \leq x \leq 1$, $0 \leq y \leq 1$. Verify Hölder's inequality in the following cases:
- $j = k = 2$
 - $j = 3, k = \frac{3}{2}$



60. Suppose that j and k are conjugate exponents.
- Show that $k = \frac{j}{j-1}$.
 - Show that $k \downarrow 1$ as $j \uparrow \infty$

Theorems Revisited

The following exercise is an analog of the result in [Exercise 10](#).

61. Prove the **parallelogram rule**:

$$\|X + Y\|_2^2 + \|X - Y\|_2^2 = 2 \|X\|_2^2 + 2 \|Y\|_2^2$$

The following exercise is an analog of the result in [Exercise 9](#).

62. Prove the **Pythagorean theorem**, named for **Pythagoras** of course: if (X_1, X_2, \dots, X_n) is a sequence of real-valued random variables with $\langle X_i, X_j \rangle = 0$ for $i \neq j$ then

$$\|\sum_{i=1}^n X_i\|_2^2 = \sum_{i=1}^n \|X_i\|_2^2$$

[Virtual Laboratories](#) > [4. Expected Value](#) > [1](#) [2](#) [3](#) [4](#) [5](#) [6](#)

[Contents](#) | [Applets](#) | [Data Sets](#) | [Biographies](#) | [External Resources](#) | [Key words](#) | [Feedback](#) | [©](#)