

Virtual Laboratories > 6. Random Samples > 1 2 3 4 5 6 7

## 7. Sample Covariance and Correlation

### The Bivariate Model

Suppose again that we have a basic [random experiment](#), and that  $X$  and  $Y$  are real-valued [random variables](#) for the experiment. Equivalently,  $(X, Y)$  is a random vector taking values in  $\mathbb{R}^2$ . Please recall the basic properties of the [means](#),  $\mathbb{E}(X)$  and  $\mathbb{E}(Y)$ , the [variances](#),  $\text{var}(X)$  and  $\text{var}(Y)$  and the [covariance](#)  $\text{cov}(X, Y)$ . In particular, recall that the correlation is

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

We will also need a higher order bivariate moment. Let

$$d(X, Y) = \mathbb{E}(((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)))^2)$$

Now suppose that we run the basic experiment  $n$  times. This creates a compound experiment with a sequence of [independent](#) random vectors  $((X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n))$  each with the same distribution as  $(X, Y)$ . In statistical terms, this is a random sample of size  $n$  from the distribution of  $(X, Y)$ . As usual, we will let  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  denote the sequence of first coordinates; this is a random sample of size  $n$  from the distribution of  $X$ . Similarly, we will let  $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$  denote the sequence of second coordinates; this is a random sample of size  $n$  from the distribution of  $Y$ .

Recall that the [sample means](#) and [sample variances](#) for  $\mathbf{X}$  are defined as follows (and of course analogous definitions hold for  $\mathbf{Y}$ ):

$$M(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n X_i, \quad W^2(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}(X))^2, \quad S^2(\mathbf{X}) = \frac{1}{n-1} \sum_{i=1}^n (X_i - M(\mathbf{X}))^2$$

In this section, we will define and study statistics that are natural estimators of the distribution covariance and correlation. These statistics will be measures of the linear relationship of the sample points in the plane. As usual, the definitions depend on what other parameters are known and unknown.

### A Special Sample Covariance

Suppose first that the distribution means  $\mathbb{E}(X)$  and  $\mathbb{E}(Y)$  are known. This is usually an unrealistic assumption, of course, but is still a good place to start because the analysis is very simple and the results we obtain will be useful below. A natural estimator of  $\text{cov}(X, Y)$  in this case is

$$W(\mathbf{X}, \mathbf{Y}) = \frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}(X))(Y_i - \mathbb{E}(Y))$$

1. Show that  $W(\mathbf{X}, \mathbf{Y})$  is the sample mean for a random sample of size  $n$  from the distribution of  $(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))$ .

2. Use the result of Exercise 1 to show that

- $\mathbb{E}(W(\mathbf{X}, \mathbf{Y})) = \text{cov}(X, Y)$
- $\text{var}(W(\mathbf{X}, \mathbf{Y})) = \frac{1}{n} (d(X, Y) - \text{cov}^2(X, Y))$
- $W(\mathbf{X}, \mathbf{Y}) \rightarrow \text{cov}(X, Y)$  as  $n \rightarrow \infty$  with probability 1.

In particular,  $W(\mathbf{X}, \mathbf{Y})$  is an **unbiased** and **consistent** estimator of  $\text{cov}(X, Y)$ .

## Properties

The formula in the following exercise is sometimes better than the definition for computational purposes.

3. With  $\mathbf{X}\mathbf{Y}$  defined to be the sequence  $(X_1 Y_1, X_2 Y_2, \dots, X_n Y_n)$ , show that

$$W(\mathbf{X}, \mathbf{Y}) = M(\mathbf{X}\mathbf{Y}) - M(\mathbf{X})\mathbb{E}(Y) - M(\mathbf{Y})\mathbb{E}(X) + \mathbb{E}(X)\mathbb{E}(Y)$$

The properties established in the following exercises are analogies of properties for the distribution covariance

4. Show that  $W(\mathbf{X}, \mathbf{X}) = W^2(\mathbf{X})$

5. Show that  $W(\mathbf{X}, \mathbf{Y}) = W(\mathbf{Y}, \mathbf{X})$

6. Show that if  $a$  is a constant then  $W(a\mathbf{X}, \mathbf{Y}) = a W(\mathbf{X}, \mathbf{Y})$

7. Show that  $W(\mathbf{X} + \mathbf{Y}, \mathbf{Z}) = W(\mathbf{X}, \mathbf{Z}) + W(\mathbf{Y}, \mathbf{Z})$

The following exercise gives a formula for the sample variance of a sum. The result extends naturally to larger sums.

8. Show that  $W^2(\mathbf{X} + \mathbf{Y}) = W^2(\mathbf{X}) + W^2(\mathbf{Y}) + 2 W(\mathbf{X}, \mathbf{Y})$

## The Standard Sample Covariance

Consider now the more realistic assumption that the distribution means  $\mathbb{E}(X)$  and  $\mathbb{E}(Y)$  are unknown. A natural approach in this case is to average  $(X_i - M(\mathbf{X}))(Y_i - M(\mathbf{Y}))$  over  $i \in \{1, 2, \dots, n\}$ . But rather than dividing by  $n$  in our average, we should divide by whatever constant gives an unbiased estimator of  $\text{cov}(X, Y)$ .

9. Interpret the sign of  $(X_i - M(\mathbf{X}))(Y_i - M(\mathbf{Y}))$  geometrically, in terms of the scatterplot of points and its center.

**Derivation**

10. Use the bilinearity of the covariance operator to show that

$$\text{cov}(M(\mathbf{X}), M(\mathbf{Y})) = \frac{\text{cov}(X, Y)}{n}$$

11. Expand and sum term by term to show that

$$\sum_{i=1}^n (X_i - M(\mathbf{X}))(Y_i - M(\mathbf{Y})) = \sum_{i=1}^n X_i Y_i - n M(\mathbf{X}) M(\mathbf{Y})$$

12. Use the result of Exercises 10 and 11, and basic properties of expected value, to show that

$$\mathbb{E}\left(\sum_{i=1}^n (X_i - M(\mathbf{X}))(Y_i - M(\mathbf{Y}))\right) = (n - 1) \text{cov}(X, Y)$$

Therefore, to have an unbiased estimator of  $\text{cov}(X, Y)$ , we should define the **sample covariance** to be the random variable

$$S(\mathbf{X}, \mathbf{Y}) = \frac{1}{n - 1} \sum_{i=1}^n (X_i - M(\mathbf{X}))(Y_i - M(\mathbf{Y}))$$

As with the sample variance, when the sample size  $n$  is large, it makes little difference whether we divide by  $n$  or  $n - 1$ .

**Properties**

The formula in the following exercise is sometimes better than the definition for computational purposes.

13. With  $\mathbf{X} \mathbf{Y}$  defined as in [Exercise 3](#), show that

$$S(\mathbf{X}, \mathbf{Y}) = \frac{1}{n - 1} \sum_{i=1}^n X_i Y_i - \frac{n}{n - 1} M(\mathbf{X}) M(\mathbf{Y}) = \frac{n}{n - 1} (M(\mathbf{X} \mathbf{Y}) - M(\mathbf{X}) M(\mathbf{Y}))$$

14. Use the result of the previous exercise and the strong law of large numbers to show that  $S(\mathbf{X}, \mathbf{Y}) \rightarrow \text{cov}(X, Y)$  as  $n \rightarrow \infty$  with probability 1.

The properties established in the following exercises are analogies of properties for the distribution covariance

15. Show that  $S(\mathbf{X}, \mathbf{X}) = S^2(\mathbf{X})$

16. Show that  $S(\mathbf{X}, \mathbf{Y}) = S(\mathbf{Y}, \mathbf{X})$

17. Show that if  $a$  is a constant then  $S(a \mathbf{X}, \mathbf{Y}) = a S(\mathbf{X}, \mathbf{Y})$

18. Show that  $S(\mathbf{X} + \mathbf{Y}, \mathbf{Z}) = S(\mathbf{X}, \mathbf{Z}) + S(\mathbf{Y}, \mathbf{Z})$

19. Show that

$$S(\mathbf{X}, \mathbf{Y}) = \frac{n}{n-1} (W(\mathbf{X}, \mathbf{Y}) - (M(\mathbf{X}) - \mathbb{E}(X))(M(\mathbf{Y}) - \mathbb{E}(Y)))$$

The following exercise gives a formula for the sample variance of a sum. The result extends naturally to larger sums.

$$\blacksquare 20. \text{ Show that } S^2(\mathbf{X} + \mathbf{Y}) = S^2(\mathbf{X}) + S^2(\mathbf{Y}) + 2S(\mathbf{X}, \mathbf{Y})$$

## Variance

In this subsection we will derive the following formula for the variance of the sample covariance. The derivation was contributed by Ranjith Unnikrishnan, and is similar to the derivation of the variance of the [sample variance](#).

$$\text{var}(S(\mathbf{X}, \mathbf{Y})) = \frac{1}{n} \left( d(\mathbf{X}, \mathbf{Y}) + \frac{1}{n-1} \text{var}(X) \text{var}(Y) - \frac{n-2}{n-1} \text{cov}^2(\mathbf{X}, \mathbf{Y}) \right)$$

$\blacksquare$  21. Verify the following result. *Hint:* Start with the expression on the right. Expand the product  $(X_i - X_j)(Y_i - Y_j)$ , and take the sums term by term.

$$S(\mathbf{X}, \mathbf{Y}) = \frac{1}{2n(n-1)} \sum_{i=1}^n \sum_{j=1}^n (X_i - X_j)(Y_i - Y_j)$$

It follows that  $\text{var}(S(\mathbf{X}, \mathbf{Y}))$  is the sum of all of the pairwise covariances of the terms in the expansion of Exercise 21.

$\blacksquare$  22. Now, derive the formula for  $\text{var}(S(\mathbf{X}, \mathbf{Y}))$  by showing that

- $\text{cov}((X_i - X_j)(Y_i - Y_j), (X_k - X_l)(Y_k - Y_l)) = 0$  if  $i = j$  or  $k = l$  or  $i, j, k, l$  are distinct.
- $\text{cov}((X_i - X_j)(Y_i - Y_j), (X_i - X_j)(Y_i - Y_j)) = 2d(\mathbf{X}, \mathbf{Y}) + 2\text{var}(X)\text{var}(Y)$  if  $i \neq j$ , and there are  $2n(n-1)$  such terms in the sum of covariances.
- $\text{cov}((X_i - X_j)(Y_i - Y_j), (X_k - X_j)(Y_k - Y_j)) = d(\mathbf{X}, \mathbf{Y}) - \text{cov}^2(\mathbf{X}, \mathbf{Y})$  if  $i, j, k$  are distinct, and there are  $4n(n-1)(n-2)$  such terms in the sum of covariances.

$\blacksquare$  23. Show that  $\text{var}(S(\mathbf{X}, \mathbf{Y})) > \text{var}(W(\mathbf{X}, \mathbf{Y}))$ . Does this seem reasonable?

$\blacksquare$  24. Show that  $\text{var}(S(\mathbf{X}, \mathbf{Y})) \rightarrow 0$  as  $n \rightarrow \infty$ . Thus, the sample covariance is a **consistent** estimator of the distribution covariance.

## Sample Correlation

By analogy with the distribution correlation, the **sample correlation** is obtained by dividing the sample covariance by the product of the sample standard deviations:

$$R(X, Y) = \frac{S(X, Y)}{S(X)S(Y)}$$

25. Use the strong law of large numbers to show that  $R(X, Y) \rightarrow \text{cor}(X, Y)$  as  $n \rightarrow \infty$  with probability 1.

26. Click in the [interactive scatterplot](#) to define 20 points and try to come as close as possible to the following conditions: sample means 0, sample standard deviations 1, sample correlation as follows: 0, 0.5, -0.5, 0.7, -0.7, 0.9, -0.9.

27. Click in the [interactive scatterplot](#) to define 20 points and try to come as close as possible to the following conditions:  $X$  sample mean 1,  $Y$  sample mean 3,  $X$  sample standard deviation 2,  $Y$  sample standard deviation 1, sample correlation as follows: 0, 0.5, -0.5, 0.7, -0.7, 0.9, -0.9.

## The Best Linear Predictor

### The Distribution Version

Recall that in the section on (distribution) [correlation and regression](#), we showed that the best linear predictor of  $Y$  based on  $X$ , in the sense of minimizing mean square error, is the random variable

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

Moreover, the (minimum) value of the mean square error is

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

The [distribution regression line](#) is given by

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

### The Sample Version

Of course, in real applications, we are unlikely to know the distribution parameters  $\mathbb{E}(X)$ ,  $\mathbb{E}(Y)$ ,  $\text{var}(X)$ , and  $\text{cov}(X, Y)$ . Thus, in this section, we are interested in the problem of estimating the best linear predictor of  $Y$  based on  $X$  from our random sample  $((X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n))$ . One natural approach is to find the line  $y = Ax + B$  that fits the sample points best. This is a basic and important problem in many areas of mathematics, not just statistics. The term *best* means that we want to find the line (that is, find  $A$  and  $B$ ) that minimizes the average of the squared errors between the actual  $y$  values in our data and the predicted  $y$  values:

$$\text{MSE}(A, B) = \frac{1}{n-1} \sum_{i=1}^n (Y_i - (AX_i + B))^2$$

Finding  $A$  and  $B$  that minimize MSE is a standard problem in calculus.

28. Show that MSE is minimized for

$$A(\mathbf{X}, \mathbf{Y}) = \frac{S(\mathbf{X}, \mathbf{Y})}{S^2(\mathbf{X})}, \quad B(\mathbf{X}, \mathbf{Y}) = M(\mathbf{Y}) - \frac{S(\mathbf{X}, \mathbf{Y})}{S^2(\mathbf{X})} M(\mathbf{X})$$

and thus the **sample regression line** is

$$y = M(\mathbf{Y}) + \frac{S(\mathbf{X}, \mathbf{Y})}{S^2(\mathbf{X})} (x - M(\mathbf{X}))$$

29. Show that the minimum mean square error, using the coefficients in the previous exercise, is

$$\text{MSE}(A(\mathbf{X}, \mathbf{Y}), B(\mathbf{X}, \mathbf{Y})) = S^2(\mathbf{Y}) (1 - R^2(\mathbf{X}, \mathbf{Y}))$$

30. Use the result of the previous exercise to show that

- $-1 \leq R(\mathbf{X}, \mathbf{Y}) \leq 1$
- $R(\mathbf{X}, \mathbf{Y}) = -1$  if and only if the sample points lie on a line with negative slope.
- $R(\mathbf{X}, \mathbf{Y}) = 1$  if and only if the sample points lie on a line with positive slope.

Thus, the sample correlation measures the degree of linearity of the sample points. The results in the previous exercise can also be obtained by noting that the sample correlation is simply the correlation of the empirical distribution. Of course, properties (a), (b), and (c) are known for the distribution correlation.

The fact that the results in [Exercise 28](#) and [Exercise 29](#) are the sample analogies of the corresponding distribution results is beautiful and reassuring. Note that the sample regression line passes through  $(M(\mathbf{X}), M(\mathbf{Y}))$ , the center of the empirical distribution. Naturally, the coefficients of the sample regression line can be viewed as estimators of the respective coefficients in the distribution regression line.

31. Assuming that the appropriate higher order moments are finite, use the law of large numbers to show that, with probability 1, the coefficients of the sample regression line converge to the coefficients of the distribution regression line:

$$\frac{S(\mathbf{X}, \mathbf{Y})}{S^2(\mathbf{X})} \rightarrow \frac{\text{cov}(X, Y)}{\text{var}(X)} \text{ as } n \rightarrow \infty$$

$$M(\mathbf{Y}) - \frac{S(\mathbf{X}, \mathbf{Y})}{S^2(\mathbf{X})} M(\mathbf{X}) \rightarrow \mathbb{E}(Y) - \frac{\text{cov}(X, Y)}{\text{var}(X)} \mathbb{E}(X) \text{ as } n \rightarrow \infty$$

As with the distribution regression lines, the choice of predictor and response variables is important.

32. Show that the sample regression line for  $Y$  based on  $X$  and the sample regression line for  $X$  based on  $Y$  are not the same line, except in the trivial case where the sample points all lie on a line.

Recall that the constant  $B$  that minimizes

$$\text{MSE}(B) = \frac{1}{n-1} \sum_{i=1}^n (Y_i - B)^2$$

is the sample mean  $M(Y)$ , and the minimum value of the mean square error is the sample variance  $S^2(Y)$ .

Thus, the difference between this value of the mean square error and the one in [Exercise 29](#), namely  $S^2(Y) - R^2(X, Y)S^2(Y)$  is the reduction in the variability of the  $Y$  data when the linear term in  $X$  is added to the predictor. The fractional reduction is  $R^2(X, Y)$ , and hence this statistics is called the (sample) **coefficient of determination**.

## Exercises

### Simulation Exercises

33. Click in the [interactive scatterplot](#), in various places, and watch how the regression line changes.
34. Click in the [interactive scatterplot](#) to define 20 points. Try to generate a scatterplot in which the mean of the  $x$  values is 0, the standard deviation of the  $x$  values is 1, and in which the regression line has
- slope 1, intercept 1
  - slope 3, intercept 0
  - slope  $-2$ , intercept 1
35. Click in the [interactive scatterplot](#) to define 20 points with the following properties: the mean of the  $x$  values is 1, the mean of the  $y$  values is 1, and the regression line has slope 1 and intercept 2.

If you had a difficult time with the previous exercise, it's because the conditions imposed are impossible to satisfy!

36. Run the [bivariate uniform experiment](#) 2000 times, with an update frequency of 10, in each of the following cases. Note the apparent convergence of the sample means to the distribution means, the sample standard deviations to the distribution standard deviations, the sample correlation to the distribution correlation, and the sample regression line to distribution regression line.
- The uniform distribution on the square
  - The uniform distribution on the triangle.
  - The uniform distribution on the circle.
37. Run the [bivariate normal experiment](#) 2000 times, with an update frequency of 10, in each of the following cases. Note the apparent convergence of the sample means to the distribution means, the sample standard deviations to the distribution standard deviations, the sample correlation to the distribution correlation, and the sample regression line to the distribution regression line.
- $\text{sd}(X) = 1$ ,  $\text{sd}(Y) = 2$ ,  $\text{cor}(X, Y) = 0.5$
  - $\text{sd}(X) = 1.5$ ,  $\text{sd}(Y) = 0.5$ ,  $\text{cor}(X, Y) = -0.7$

### Data Analysis Exercises

38. Compute the correlation between petal length and petal width for the following cases in **Fisher's iris data**. Comment on the differences.

- All cases
- Setosa only
- Verginica only
- Versicolor only



39. Compute the correlation between each pair of color count variables in the **M&M data**



40. Consider all cases in **Fisher's iris data**.

- Compute the least squares regression line with petal length as the predictor variable and petal width as the response variable.
- Draw the scatterplot and the regression line together.
- Predict the petal width of an iris with petal length 40



41. Consider the Setosa cases only in **Fisher's iris data**.

- Compute the least squares regression line with sepal length as the predictor variable and sepal width as the unknown variable.
- Draw the scatterplot and regression line together.
- Predict the sepal width of an iris with sepal length 45.



[Virtual Laboratories](#) > [6. Random Samples](#) > [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#)

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