Restor Dent Endod. 2018 Feb;43(1):e4
https://doi.org/10.5395/rde.2018.43.e4
pISSN 2234-7658·eISSN 2234-7666

Restorative Dentistry & Endodontics RDE

Open Lecture on Statistics

Check for updates

Statistical notes for clinical researchers: covariance and correlation

Hae-Young Kim *

Department of Health Policy and Management, College of Health Science, and Department of Public Health Science, Graduate School, Korea University, Seoul, Korea

OPEN ACCESS

***Correspondence to**
Hae-Young Kim, DDS, PhD
Department of Health Policy and Management, College of Health Science, and Department of Public Health Science, Graduate School, Korea University,
145 Anam-ro, Seongbuk-gu, Seoul 02841, Korea.
Tel: +82-2-3290-5667
Fax: +82-2-940-2879
E-mail: kimhaey@korea.ac.kr

ORCID iDs
Hae-Young Kim
https://orcid.org/0000-0003-2043-2575

Covariance and correlation are basic measures describing the relationship between two variables. They are a broad class of statistical tools which evaluate how two variables are related with dependence or association, especially for linear relationship. Difference between the two is that covariance is calculated under the original units of two variables, while correlation is obtained based on standardized scale resulting in a unit-less measure.

COVARIANCE

Covariance is defined as the expected value of variations of two variables from their expected values. More simply, covariance measures how much variables change together. The mean of each variable is used as reference and relative positions of observations compared to mean is important. Covariance is simply defined as the mean of multiplication of corresponding X and Y deviations from their mean, $(X - \bar{X})$ and $(Y - \bar{Y})$. Covariance is expressed as following formula:

$$\text{Covariance}(X, Y) = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{n - 1}$$

where n is the number of X and Y pairs.

Covariance mainly represents the direction of relationship of two variables. A positive sign of covariance value represents that two variables move to the same direction while a negative covariance value means that two variables move to opposite directions. **Figure 1** shows a coordinate plane made by the line X and Y (dotted line) as well as $\bar{X} = 5$ and $\bar{Y} = 6$ (solid line). In the quadrant I, x value moves positively from its mean and so does y value. Therefore, the points in the quadrant I represent positive relationship between two variables because Y values get larger as X values get larger relatively to their mean. Similarly, the points in the quadrant III also represent positive relationship because Y values get smaller than its mean as X values get smaller. The sign of the multiplicative value of each deviation, $(X - \bar{X})(Y - \bar{Y})$ is positive in the quadrants I and III because the pair of deviations have the same sign that both are positive or negative. Contrary, the points in quadrants II and IV show negative relationship as one variable gets larger than its mean as the other gets smaller (**Figure 1**). In quadrants II and IV, the sign of the multiplicative value of each deviation is negative because the pair of deviations have different signs each other.

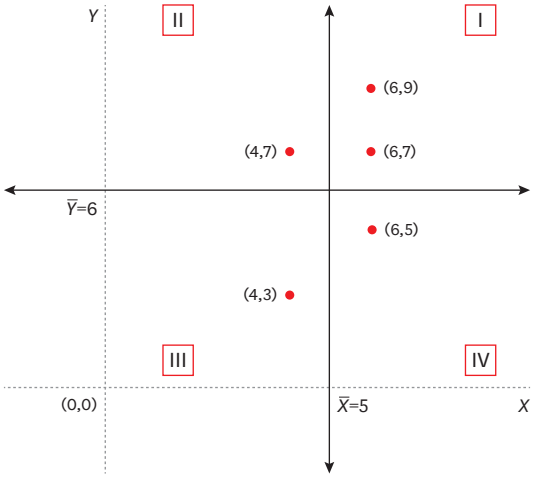


Figure 1. Illustration of relative positions of two variables in reference to their means.

A positive sign of covariance means that points in the quadrants I and III are predominant than those in the quadrants II and IV. A negative sign represents predominance of points in the quadrants II and IV. Therefore, positive and negative signs of covariance values can be interpreted as positive and negative relationships between 2 variables, respectively. If the covariance value is near zero, we may interpret there is no clear positive or negative relationship. The example data in **Table 1** shows positive covariance value, 109.1, which means a positive, increasing relationship between *X* and *Y*.

Table 1. The calculation procedure of covariance and Pearson correlation coefficient

No	X	Y	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
1	73	90	0.55	7.65	0.30	58.52	4.21
2	52	74	−20.45	−8.35	418.20	69.72	170.76
3	68	91	−4.45	8.65	19.80	74.82	−38.49
4	47	62	−25.45	−20.35	647.70	414.12	517.91
5	60	63	−12.45	−19.35	155.00	374.42	240.91
6	71	78	−1.45	−4.35	2.10	18.92	6.31
7	67	60	−5.45	−22.35	29.70	499.52	121.81
8	80	89	7.55	6.65	57.00	44.22	50.21
9	86	82	13.55	−0.35	183.60	0.12	−4.74
10	91	105	18.55	22.65	344.10	513.02	420.16
11	67	76	−5.45	−6.35	29.70	40.32	34.61
12	73	82	0.55	−0.35	0.30	0.12	−0.19
13	71	93	−1.45	10.65	2.10	113.42	−15.44
14	57	73	−15.45	−9.35	238.70	87.42	144.46
15	86	82	13.55	−0.35	183.60	0.12	−4.74
16	76	88	3.55	5.65	12.60	31.92	20.06
17	91	97	18.55	14.65	344.10	214.62	271.76
18	69	80	−3.45	−2.35	11.90	5.52	8.11
19	87	87	14.55	4.65	211.70	21.62	67.66
20	77	95	4.55	12.65	20.70	160.02	57.56
	$\bar{X}$ = 72.45	$\bar{Y}$ = 82.35			Σ = 2,912.95	Σ = 2,742.55	Σ = 2,072.85

$$\text{Covariance}(X,Y) = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{n-1} = \frac{2{,}072.85}{20-1} = 109.10,$$

$$\text{SD}(X) = \sqrt{\frac{\sum(X-\bar{X})^2}{n-1}} = \sqrt{\frac{2912.95}{20-1}} = 12.38, \quad \text{SD}(Y) = \sqrt{\frac{\sum(Y-\bar{Y})^2}{n-1}} = \sqrt{\frac{2{,}742.55}{20-1}} = 12.01,$$

$$\text{Pearson correlation coefficient } (r) = \frac{Cov(x,y)}{\text{SD}(X)\times\text{SD}(Y)} = \frac{109.10}{12.38\times12.01} = 0.73.$$

https://doi.org/10.5395/rde.2018.43.e4

Then what can we say about the size of covariance values? Magnitude of the relationship? However, the problem is that the absolute value of covariance depends on the unit of variables. For example, if we change the unit of a variable from kilometer to meter unit, then the deviance from mean of 1 in kilometer units (km) is changed into 1,000 in meter units (m). The unit change makes huge difference in the value of covariance, even when the relationship of 2 variables is the same. Therefore, the size of covariance value cannot be interpretable as the magnitude of a relationship. Also, a covariance value has neither upper or lower bound nor any standard to determine the degree of relationship. There is a need of unit standardization procedure on covariance.

Table 1 shows the calculation procedure of covariance and Pearson correlation coefficient. Deviations of *X* and *Y* are multiplied, summed-up, and finally divided by n-1 to get covariance value. The Pearson correlation coefficient is obtained by dividing covariance value with standard deviations (SDs) of *X* and *Y* variables.

PEARSON CORRELATION COEFFICIENT

Correlation is the standardized form of covariance by dividing the covariance with SD of each variable under normal distribution assumption. Generally, we use '*r*' as sample correlation coefficient and 'ρ' as population correlation coefficient. The Pearson correlation coefficient has following formula.

$$\text{Pearson Correlation coefficient } (\rho \text{ or } r) = \frac{Cov(x, y)}{\text{SD}(X) \times \text{SD}(Y)}$$

The Pearson correlation coefficient is also the covariance of standardized form of *X* and *Y* variables. The correlation coefficient is unit-less, being independent of the scale of variables and the range is between –1 and +1. The interpretation of the Pearson correlation coefficient was provided by Cohen [1]. He proposed a small, medium, and large effect size of r as 0.1, 0.3, and 0.5, respectively, and explained that the medium effect size represents an effect likely to be visible to the naked eye of a careful observer. Also he subjectively set a small effect size to be noticeably smaller than medium but not so small as to be trivial and set a large effect size to be the same distance above the medium as small was below it [1]. His standard is generally accepted at the present.

Because the correlation coefficient reflects only the strength of linear relationship, we need a cautious investigation of scatterplot before calculating it. As shown in **Figure 2**, a correlation coefficient of 0.8 can be obtained from totally different relationships between two variables. Only **Figure 2A** shows correct linear relationship, while curved relationship (**Figure 2B**), distorting effect of an outlier (**Figure 2C**), and strong effect of an outlier on the unrelated relationship (**Figure 2D**) show some relations different from a linear one. We need to keep in mind that all these different shapes of relationships could result in the same correlation coefficient. We should check those possibilities using scatterplots.

If the correlation coefficient is zero, there is also some caution needed in interpreting the meaning. We may expect no relationship such as **Figure 3A**, the shape of random scatter. However, U shape or reverse U shape relationship can show zero correlation coefficient (**Figure 3B and 3C**). An example of U shape relationship is the relationship between

https://doi.org/10.5395/rde.2018.43.e4

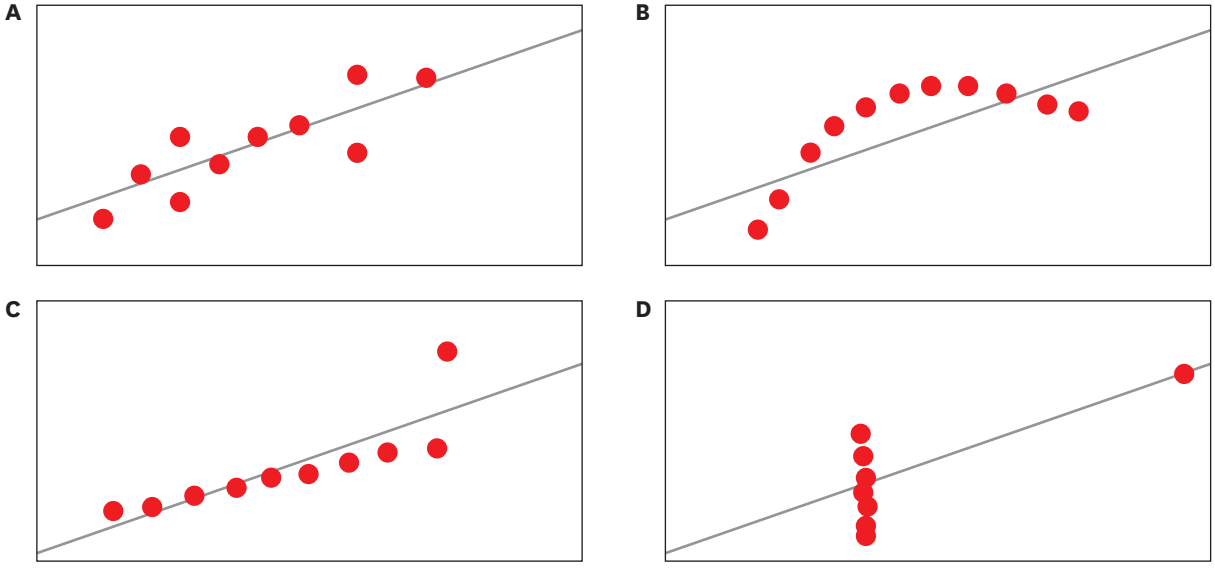


Figure 2. Four different relationship of two variables with the same correlation coefficient of around 0.8.

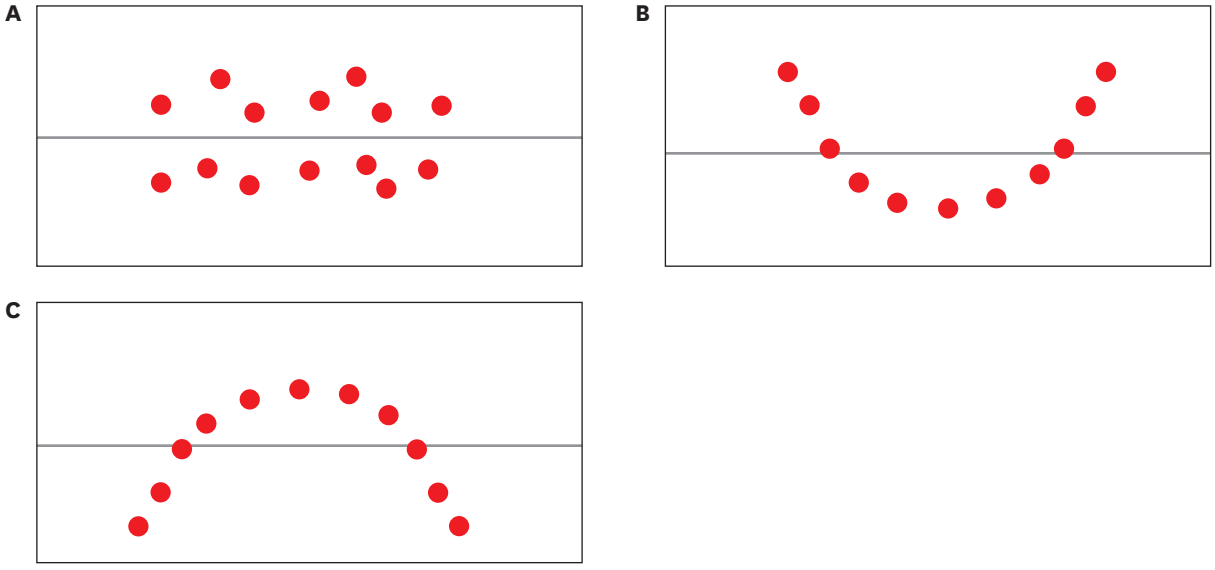


Figure 3. Typical types of zero correlation: (A) random scatter — true no association, (B) U shape, and (C) reverse U shape.

consumption of electricity and temperature. At very low temperature lots of electricity is consumed for warming and the need is gradually decreased with the increase of temperature. However, if temperature rises further we need more electricity for air conditioning. An example of reverse U shape is the relationship between stress and work performance. Performance may be improved if there is some stress, but too much stress can cause burn-out of the person which decreases performance. Therefore, it is always a good idea to examine the relationship between variables with a scatterplot.

https://doi.org/10.5395/rde.2018.43.e4

SPEARMAN'S RANK CORRELATION COEFFICIENT

Spearman's rank correlation coefficient is the non-parametric version of the Pearson correlation coefficient calculated using rank values of two variables. It is expressed as following formula.

$$\text{Spearman correlation coefficient } (\rho \text{ or } r_s) = 1 - \frac{6\sum d^2}{n(n^2-1)},$$

where d = Rank(Y) − Rank(X) and n = sample size.

While the Pearson correlation assesses linear relationships, Spearman correlation assesses monotonic relationship that two variables are related but not necessarily linear. Let's consider the relationship between 99 p values ranges from 0.01 to 0.99 by 0.01-unit increase and corresponding log odds values, $\ln\left(\frac{p}{1-p}\right)$. As seen **Figure 4**, Spearman correlation coefficient is 1 when y values always increase as x values increase, while Pearson correlation coefficient counts only linearity with coefficient of 0.97. Spearman rank correlation coefficient can be applied on continuous variables with influential outliers which are located far away from most observations. In such cases, it is not appropriate to apply Pearson. Also, it can be applied to assess relationships between ordered categorical values. The range of Spearman correlation coefficient is from –1 to +1, which represent perfect negative and positive relationships, respectively.

Table 2 shows the calculation procedure of the Spearman rank correlation. Difference of rank of two variables is used in calculating rank correlation coefficient.

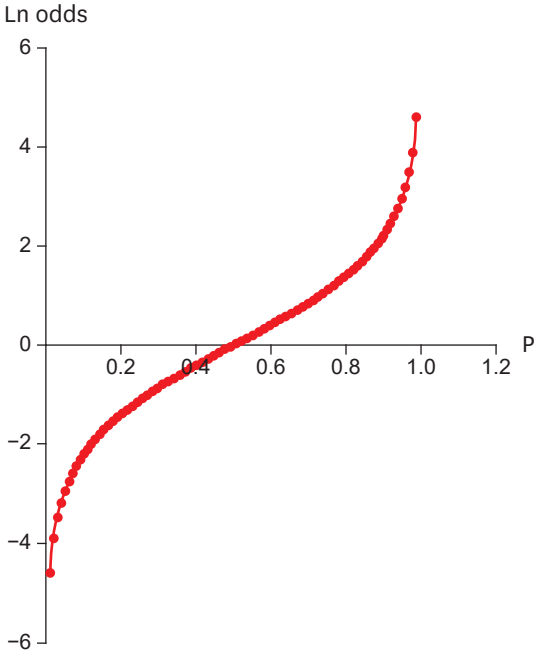


Figure 4. A curved relationship of 'p' and 'ln odds'. (Pearson correlation = 0.97 vs. Spearman's rank correlation = 1).

https://doi.org/10.5395/rde.2018.43.e4

Table 2. The calculation procedure of the Spearman's rank correlation

No	*X*	*Y*	Rank(*X*)	Rank(*Y*)	d^*	d^2
1	73	90	11.5	15	3.5	12.25
2	52	74	2	5	3	9
3	68	91	7	16	9	81
4	47	62	1	2	1	1
5	60	63	4	3	−1	1
6	71	78	9.5	7	−2.5	6.25
7	67	60	5.5	1	−4.5	20.25
8	80	89	15	14	−1	1
9	86	82	16.5	10	−6.5	42.25
10	91	105	19.5	20	0.5	0.25
11	67	76	5.5	6	0.5	0.25
12	73	82	11.5	10	−1.5	2.25
13	71	93	9.5	17	7.5	56.25
14	57	73	3	4	1	1
15	86	82	16.5	10	−6.5	42.25
16	76	88	13	13	0	0
17	91	97	19.5	19	−0.5	0.25
18	69	80	8	8	0	0
19	87	87	18	12	−6	36
20	77	95	14	18	4	16
						Σ = 328.5

Spearman's correlation coefficient $(r_s) = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6 \times 328.5}{20(20^2-1)} = 1 - 0.25 = 0.75$.

*d = Rank(*Y*) − Rank(*X*).

REFERENCES

1. Cohen J. A power primer. Psychol Bull 1992;112:155-159.
PUBMED | CROSSREF

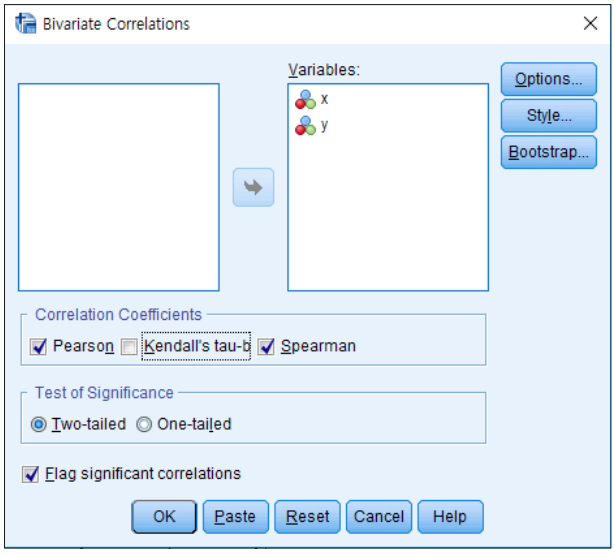
Appendix 1. Procedure of covariance and correlation using IBM SPSS

The procedure of logistic regression using IBM SPSS Statistics for Windows Version 23.0 (IBM Corp., Armonk, NY, USA) is as follows.

(A) Data

	x	y
1	73	90
2	52	74
3	68	91
4	47	62
5	60	63
6	71	78
7	67	60
8	80	89
9	86	82
10	91	105
11	67	76
12	73	82
13	71	93
14	57	73
15	86	82
16	76	88
17	91	97
18	69	80
19	87	87
20	77	95

(B) Analyze-Correlate-Bivariate logistic



(C) Pearson correlation coefficient

Correlations

		x	y
x	Pearson Correlation	1	.733**
	Sig. (2-tailed)		.000
	N	20	20
y	Pearson Correlation	.733**	1
	Sig. (2-tailed)	.000	
	N	20	20

**. Correlation is significant at the 0.01 level (2-tailed).

(D) Spearman's rank correlation

Correlations

			x	y
Spearman's rho	x	Correlation Coefficient	1.000	.752**
		Sig. (2-tailed)	.	.000
		N	20	20
	y	Correlation Coefficient	.752**	1.000
		Sig. (2-tailed)	.000	.
		N	20	20

**. Correlation is significant at the 0.01 level (2-tailed).

https://doi.org/10.5395/rde.2018.43.e4