Learn About Pearson's Correlation Coefficient in SPSS With Data From the Global Health Observatory Data (2012)

Learn About Pearson's Correlation Coefficient in SPSS With Data From the Global Health Observatory Data (2012)

Student Guide

Introduction

This dataset example introduces researchers to Pearson's correlation coefficient, which is a measure of association between two continuous variables. Pearson's correlation coefficient measures the positive or negative linear relationship between two continuous variables. This example shows how to calculate and interpret the Pearson correlation coefficient using a subset of data from the Global Health Observatory data, 2012, looking specifically at International Health Regulation (IHR) scores on disaster preparedness and food safety across 137 countries. Analysis like this might help researchers and policy makers make better plans in anticipation of potential disasters.

What Is Correlation?

Correlation measures the association or dependence between two continuous variables. In theory, continuous variables can take on any numerical value within their range. In practice, variables that take on a large number of different values within their range are treated as continuous. Variables like age measured in years, income measured in dollars, or unemployment measured in percentages are all good examples. In this example, we explore one of the most common measures

of correlation between two continuous variables – the Pearson correlation coefficient.

Values for the Pearson correlation coefficient can range from +1 to −1, with 0 indicating that no correlation between the variables exists. In addition to assessing whether the two variables are related, the coefficient indicates both the direction and the strength of the correlation. The closer the coefficient is to +1 or −1, the *stronger* the relationship, while the sign indicates if the relationship is positive or negative. Researchers tend to put arbitrary evaluation points on different values of the coefficient. While this differs from source to source, the rule of thumb we suggest is:

- Coefficient between −0.3 and +0.3 = weak correlation.
- Coefficient less than −0.7 or greater than +0.7 = strong correlation.
- Coefficient between −0.3 and −0.7 or between +0.3 and +0.7 = moderate correlation.

We can also conduct a hypothesis test of the estimated coefficient. When computing formal statistical tests, it is customary to define the null hypothesis (H_0) to be tested. In this case, the standard null hypothesis is that the two variables in question are independent of each other (i.e. that the correlation between them equals zero). It is unlikely that the correlation between two variables would ever be observed to be exactly zero – some association in the data is likely simply as a consequence of random chance. A statistical test helps us determine if the observed correlation coefficient is large enough to declare the correlation statistically significant. Doing so would lead us to reject the null hypothesis (H_0) of independence and conclude that there likely is a relationship between the two variables. Researchers generally define “large enough” to be when the associated significance level, or *p*-value, of a statistical test is less than or equal to 0.05.

Pearson's Correlation Coefficient

Pearson's correlation coefficient measures the linear association between two continuous variables. This means that when values of one of the variables in question tend to be high, the values of the other variable in question also tend to be high (positive correlation) or low (negative correlation). The formula in Equation 1 shows how to compute Pearson's correlation coefficient, represented as the Greek letter ρ, between two variables named *X* and *Y*:

(1)

$$\rho = \frac{\sum_{i=1}^{N} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{N} (X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{N} (Y_i - \overline{Y})^2}}$$

where:

- X_i = the value for individual values of *X*
- Y_i = the value for individual values of *Y*
- $\overline{X}$ = the mean of the values for *X*
- $\overline{Y}$ = the mean of the values for *Y*
- *N* = the sample size.

The numerator of Equation 1 determines whether the correlation will turn out to be positive, negative, or zero. The denominator of Equation 1 scales the entire calculation so that the highest possible positive correlation is +1 and the lowest possible negative correlation is −1.

Assumptions Behind the Method

Nearly every statistical test relies on some underlying assumptions, and they are all affected by the characteristics of the data you have. Critical considerations for Pearson's correlation coefficient include:

- The relationship between them, if it exists, is best characterized as linear.
- The two variables in question are continuous.
- The two variables in question are approximately normally distributed.

Pearson's correlation coefficient is robust to violations of the normality assumption if the distributions of the two variables in question are at least symmetric. This can be explored by producing histograms of each variable.

In practice, the linearity assumption is often the most important. This can be checked by producing a two-way scatter plot of the two variables. If the relationship is monotonic, but not linear, or if the variables are not normally distributed, researchers should consider using the less restrictive Spearman's rank-order correlation coefficient.

Illustrative Example: Disaster Preparedness and Food Safety

This example explores the correlation between the IHR preparedness score and the IHR food safety score across using data from the Global Health Observatory data from the World Health Organization, 2012. The research question guiding this example is:

> Are IHR preparedness scores correlated with IHR food safety scores at the country level?

This can be stated in the form of a null hypothesis:

> H_0 = There is no correlation between IHR preparedness scores and IHR food safety scores across countries.

The Data

This example uses two variables from the 2012 Global Health Observatory Data from the World Health Organization:

- IHR score on disaster preparedness (preparedness), which ranges from 0 to 100. This score measures how well a country provides for public health emergency response plans, with higher scores indicating better preparedness.
- IHR score on food safety (foodsafety), which ranges from 0 to 100. This score measures how well a country creates mechanisms for detecting and responding to food-borne disease and food contamination, with higher scores indicating better effort.

The preparedness variable has a mean of 62.3, a standard deviation of 31.3, a minimum value of 0 and a maximum value of 100. The food safety variable has a mean of 72.6, a standard deviation of 27.0, a minimum value of 0 and a maximum value of 100. Both variables are continuous measures, making them appropriate for the Pearson correlation.

Analyzing the Data

Correlations are often reported in tables where all the variables included are used to define both the rows and the columns of the table. The correlation between any variable and itself is always exactly 1. Table 1 reports the Pearson correlation coefficient evaluating the association between preparedness scores and food safety scores among countries.

Table 1: Pearson correlation between IHR preparedness score and IHR food safety score, Global Health Observatory Data from the World Health Organization, 2012.

		IHR Preparedness Score	IHR Food Safety Score
IHR Preparedness Score	Pearson Correlation	1	0.614
	Sig (2-tailed)		0.000
	Sample Size	137	137
IHR Food Safety Score	Pearson Correlation	0.614	1
	Sig (2-tailed)	0.000	
	Sample Size	137	137

Table 1 reports a positive correlation between these two variables of 0.614 that is statistically significant (p-value < 0.05). This indicates a moderate to strong positive correlation between the IHR preparedness score and the IHR food safety score across countries. Often it is good practice to compare the results of a Pearson correlation coefficient with a two-way scatter plot of the variables in question in order to have a graphical representation of the association between two variables and to evaluate the linearity of the relationship.

Presenting the Results

Results for Pearson's correlation coefficient can be presented as follows:

"We used a subset of data from the Global Health Observatory data to test the null hypothesis:

> H_0 = There is no correlation between IHR food safety scores and IHR preparedness scores across countries.

The data includes 137 countries. Table 1 reports the Pearson correlation coefficient between these two variables. The correlation is 0.614 and is statistically significantly different from zero (p-value < 0.05). This suggests a moderate to

strong positive linear relationship between the IHR preparedness score and the IHR food safety score across countries In other words, countries with better preparedness for disasters also have higher scores for food safety procedures. Further investigation of the relationship between these two variables using a two-way scatter plot is warranted."

Review

Pearson's correlation coefficient is one of the most common ways to measure association between two continuous variables. The other is Spearman's rank order correlation coefficient, which is most appropriate when the assumptions behind the Pearson's correlation coefficient do not hold. Pearson's correlation coefficient measures the strength of the linear relationship between two variables. Accepting or rejecting the null hypothesis associated with this measure does not say anything about whether there is some other form of association between the two variables in question. Two-way scatter plots are often useful ways of exploring more complicated relationships between two continuous variables.

You should know:

- What types of variable are suitable for Pearson's correlation coefficient.
- The basic assumptions underlying this statistical method.
- How to compute and interpret Pearson's correlation coefficient.
- How to report the results of a Pearson's correlation coefficient.

Your Turn

You can download this sample dataset along with a guide showing how to produce these two measures of correlation using statistical software. The sample dataset also includes another variable named surveillance that records the IHR surveillance score. See if you can reproduce the results presented here, and try

producing Pearson's correlation coefficient between the variables preparedness and surveillance.