Correlation using SPSS

Durgesh Chandra Pathak
Correlation using SPSS*

D.C. Pathak


*This Presentation has borrowed heavily from the aforesaid book.
What happens when there are two variables? 

**Covariance: Measuring Relationships (how?)**

**Covariance**

Together Changing

We get Cross-product deviations which are deviation of each variable from its mean.

\[
\text{Cov}_{x,y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(N - 1)}
\]

The numerator of the above equation is Cross-product deviation.

**Problem with Covariance:**

It depends on the scales of measurement. When two variables are measured on different units; e.g., Age and Memory.

**How to solve this problem?**

Standardization: Converting covariance into a standard set of units by dividing it with standard deviations of the two variables;
Correlation: Standardized covariance; What is the relationship between two (or more) variables.

A measure of Linear relationship between variables.

$$r = \frac{\text{COV}_{xy}}{S_x S_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(N - 1)S_x S_y}$$

Where $r$=Pearson’s Correlation Coefficient.
The value of $r$ varies between -1 to +1 through 0.

When one variable varies, there are three possibilities:
1. The second one increases when the first one increases,
2. The second one decreases when the first one increases,
3. The second one remains unchanged when the first one varies.
So, $r$ can also be positive (case 1), negative (case 2) or zero (case 3).
When $r=+1$, it implies that the two variables have a perfect positive relation;

When $r=-1$, it implies that the two variables are perfectly related in a negative manners (when one increases, the other decreases);

When $r= 0$, it implies that the two variables are not related to each other (they don’t change together).
Scatter Plots: Graphs of Relationship

1. Simple Scatter Plot: Used when there are two variables,
Simple Scatterplot

Exam Anxiety

Exam Performance (%)
3-D Scatterplot: Used to show relationship among three variables; Cumbersome to use,
3-D Scatterplot

Gender
- Red circle: Male
- Blue circle: Female
**Overlay Scatterplot**: When one variable is held constant and is plotted against several other variables;

E.g., if we are interested in knowing the relation between Anxiety and Exam performance and Revision time and Exam performance but not in Anxiety and Revision time.

Several pairs of variables are plotted on the same axes.
Exam Anxiety/Revision Time

Exam Performance (%)

Overlay Scatterplot
Matrix Scatterplot: Can be used in place of a 3-D Scatterplot.

B1: Exam Performance vs. Anxiety
C1: Exam Performance vs. Rev. Time
A2: Anxiety vs. Exam Performance
C2: Anxiety vs. Rev. Time
A3: Rev. Time vs. Exam Performance
B3: Rev. Time vs. Anxiety
Types of Correlation

- Bivariate Correlation
- Partial (& Part) Correlation

Correlation between Two variables

**How to do Bivariate Correlation in SPSS?**

1. **Analyze**
2. **Correlate**
3. **Bivariate**
4. **Transfer variables**
5. **Choose type of Correlation**
6. **Choose type of test: One-tailed or Two-Tailed**
7. **OK**

Observing relationship between two variables while **controlling** the effect of one or more additional variables.
Now, we are going to use a data set of 103 students (52 Male and 51 Female) regarding the time spent in revision of subject, anxiety before the exam, and marks obtained in exam of these students. We shall calculate Pearson’ Product-Moment Correlation Coefficient for these variables. This data set has been downloaded from Andy Field’s website.

1. Analyze
2. Correlate
3. Bivariate
   - Transfer variables of interest to ‘Variables box’
4. Decide on Test of Significance: One-tailed/two-tailed
5. Choose ‘Pearson’ in Correlation Coefficient
6. OK
SPSS Output:
Statistically significant Correlations are flagged by *. One * means the result is significant at .05 level and ** (two asterisk) implies that the result is significant at .01 level.

<table>
<thead>
<tr>
<th></th>
<th>Time Spent Revising</th>
<th>Exam Performance (%)</th>
<th>Exam Anxiety</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Spent Revising</td>
<td>Pearson Correlation</td>
<td>1</td>
<td>.397**</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>103</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>103</td>
<td>103</td>
</tr>
<tr>
<td>Exam Performance (%)</td>
<td>Pearson Correlation</td>
<td>.397**</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.000</td>
<td>103</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>103</td>
<td>103</td>
</tr>
<tr>
<td>Exam Anxiety</td>
<td>Pearson Correlation</td>
<td>-.709**</td>
<td>-.441**</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>103</td>
<td>103</td>
</tr>
</tbody>
</table>

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

**Reporting the Correlation:**
There is a statistically significant, positive relationship between Time spent in revision by an individual and his exam performance, \( r = .397 \), \( p \) (two-tailed) < .01
**Correlation and Causality:** Correlation does not imply causality, why?

- **Third variable problem:** there can be some other third variable affecting the relation between the two variables under question. E.g., Brain size and gender can be affected by size of the person.
- **Direction of causality:** correlation coefficient says nothing which variable is causing the other to change.

**Correlation and Effect Size:**
Small Effect: \( r = \pm .1 \)
Medium Effect: \( r = \pm .3 \),
Large Effect: \( r = \pm .5 \).

**Coefficient of Determination (R^2):** The correlation coefficient (r) if squared, is called the Coefficient of Determination (R^2) and can be used as a measure of the amount of variability in one variable that can be explained by the other.
E.g., in our previous example, the r between anxiety and exam performance was -0.441. So, \( R^2 = (-0.441)^2 = 0.194481 \).
We can multiply \( R^2 \) by 100 to express it in Percentage.
So, \( 0.194481 \times 100 = 19.4481\% \) = Variations in exam performance of students can be explained by variations in their anxiety.

**Word of Caution:** \( R^2 \) does not imply causal relationship.
Non-Parametric Correlations: Four assumptions should be met for the data to be Parametric:
1. Normally Distributed Data
2. Homogeneity of Variance
3. Interval Data
4. Independence

When the data has violated any or all of these assumptions, we can use non-parametric correlations.
How can we know if Our Distribution is Normal or not?
One can use Kolmogorov-Smirnov Test (K-S test) or Shapiro-Wilk Test. These test compare the scores in the sample to a normally distributed set of scores of same mean and standard deviation.

**Decision Rule:**
If the test is Non-significant (p > .05)＝ the distribution does not differ significantly from a normal distribution;
If the test is Significant (p < .05)＝ the distribution is Non-normal.
Tests of Normality

<table>
<thead>
<tr>
<th></th>
<th>Kolmogorov-Smirnov&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Shapiro-Wilk</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Statistic</td>
<td>df</td>
</tr>
<tr>
<td>Time Spent Revising</td>
<td>.179</td>
<td>103</td>
</tr>
<tr>
<td>Exam Performance (%)</td>
<td>.135</td>
<td>103</td>
</tr>
<tr>
<td>Exam Anxiety</td>
<td>.153</td>
<td>103</td>
</tr>
</tbody>
</table>

<sup>a</sup> Lilliefors Significance Correction

Thus, our data came out to be Non-normal.

**Testing Homogeneity of Variance:**

Levene’s test is used; It tests the hypothesis that the variances in the groups are equal;

**Decision Rule:**

- If Levene’s test is statistically non-significant (p > .05)
  
  ⇒ Homogeneity of variances is maintained.

- If Levene’s test is statistically significant (p < .05),

  ⇒ Variances are Heterogeneous.
## SPSS output for Levene’s Test:

### Test of Homogeneity of Variance

<table>
<thead>
<tr>
<th></th>
<th>Levene Statistic</th>
<th>df1</th>
<th>df2</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Time Spent Revising</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Based on Mean</td>
<td>.173</td>
<td>1</td>
<td>101</td>
<td>.678</td>
</tr>
<tr>
<td>Based on Median</td>
<td>.267</td>
<td>1</td>
<td>101</td>
<td>.606</td>
</tr>
<tr>
<td>Based on Median and</td>
<td>.267</td>
<td>1</td>
<td>99.318</td>
<td>.606</td>
</tr>
<tr>
<td>with adjusted df</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Based on trimmed mean</td>
<td>.247</td>
<td>1</td>
<td>101</td>
<td>.620</td>
</tr>
<tr>
<td><strong>Exam Performance (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Based on Mean</td>
<td>.160</td>
<td>1</td>
<td>101</td>
<td>.690</td>
</tr>
<tr>
<td>Based on Median</td>
<td>.068</td>
<td>1</td>
<td>101</td>
<td>.795</td>
</tr>
<tr>
<td>Based on Median and</td>
<td>.068</td>
<td>1</td>
<td>100.892</td>
<td>.795</td>
</tr>
<tr>
<td>with adjusted df</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Based on trimmed mean</td>
<td>.138</td>
<td>1</td>
<td>101</td>
<td>.711</td>
</tr>
<tr>
<td><strong>Exam Anxiety</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Based on Mean</td>
<td>.003</td>
<td>1</td>
<td>101</td>
<td>.956</td>
</tr>
<tr>
<td>Based on Median</td>
<td>.000</td>
<td>1</td>
<td>101</td>
<td>.989</td>
</tr>
<tr>
<td>Based on Median and</td>
<td>.000</td>
<td>1</td>
<td>99.177</td>
<td>.989</td>
</tr>
<tr>
<td>with adjusted df</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Based on trimmed mean</td>
<td>.000</td>
<td>1</td>
<td>101</td>
<td>.997</td>
</tr>
</tbody>
</table>
Thus, the assumption of Homogeneity of Variances is maintained for the data under consideration.

**Non-Parametric Correlations:**

1. **Spearman’s Correlation Coefficient:**
   It first ranks the data and then applies Pearson’s correlation to these ranks.

\[
r_s = 1 - \frac{6 \times \sum d^2}{N^3 - N}
\]

Where \( r_s \) is Spearman’s correlation coefficient, \( d^2 \) is the difference between the ranks and \( N \) is the number of cases.

\[
\begin{array}{cccc}
\text{Spearman's rho} & \text{Time Spent Revising} & \text{Exam Performance} & \text{Exam Anxiety} \\
\text{Correlation Coefficient} & \text{Time Spent Revising} & \text{Exam Performance} & \text{Exam Anxiety} \\
\text{Sig. (2-tailed)} & . & .000 & .000 \\
\text{N} & 103 & 103 & 103 \\
\text{Exam Performance (%)} & \text{Correlation Coefficient} & \text{Exam Anxiety} \\
\text{Sig. (2-tailed)} & .350** & .000 & .000 \\
\text{N} & 103 & 103 & 103 \\
\text{Exam Anxiety} & \text{Correlation Coefficient} & \text{Exam Performance} \\
\text{Sig. (2-tailed)} & -.622** & .000 \\
\text{N} & 103 & 103 & 103 \\
\end{array}
\]

**.** Correlation is significant at the 0.01 level (2-tailed).
2. Kendall’s Tau (\(\tau\)):
A non-parametric correlation coefficient and can be used in place of Spearman’s coefficient when one has a small data set with a large number of tied ranks.

<table>
<thead>
<tr>
<th>Correlations</th>
<th>Time Spent Revising</th>
<th>Exam Performance (%)</th>
<th>Exam Anxiety</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kendall's tau_b</td>
<td>Correlation Coefficient</td>
<td>1.000</td>
<td>.263**</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>103</td>
<td>103</td>
</tr>
<tr>
<td>Exam Performance (%)</td>
<td>Correlation Coefficient</td>
<td>.263**</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.000</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>103</td>
<td>103</td>
</tr>
<tr>
<td>Exam Anxiety</td>
<td>Correlation Coefficient</td>
<td>-.489**</td>
<td>-.285**</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>103</td>
<td>103</td>
</tr>
</tbody>
</table>

**: Correlation is significant at the 0.01 level (2-tailed).
## Comparison between Pearson, Spearman and Kendall’s $\tau$

<table>
<thead>
<tr>
<th></th>
<th>Pearson</th>
<th>Spearman</th>
<th>Kendall</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Rev. Time vs. Performance</strong></td>
<td>+ .397**</td>
<td>+ .358**</td>
<td>+ .263**</td>
</tr>
<tr>
<td><strong>Anxiety vs. Performance</strong></td>
<td>- .441**</td>
<td>- .405**</td>
<td>- .285**</td>
</tr>
<tr>
<td><strong>Rev. Time vs. Anxiety</strong></td>
<td>- .709**</td>
<td>- .622**</td>
<td>- .489**</td>
</tr>
</tbody>
</table>

Thus, selection of appropriate correlation coefficient is important as it would affect the Coefficient of Determination (i.e., the amount of variance we can explain).
**Correlation between a Continuous and a Discrete variable:**

- **Biserial Correlation:** When one variable is a continuous dichotomy; e.g., passing or failing an exam;

- **Point-Biserial Correlation:** When one variable is a discrete dichotomy; e.g., pregnancy;

<table>
<thead>
<tr>
<th>Correlations</th>
<th>Exam Performance (%)</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exam Performance (%)</td>
<td>Pearson Correlation</td>
<td>1</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>103</td>
<td>103</td>
</tr>
<tr>
<td>Gender</td>
<td>Pearson Correlation</td>
<td>-.005</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td></td>
<td>.963</td>
</tr>
<tr>
<td>N</td>
<td>103</td>
<td>103</td>
</tr>
</tbody>
</table>

The sign of the correlation coefficient becomes irrelevant in case of Point-Biserial Correlation. It depends on the way variables are categorized. If we reverse the coding, the sign of Point-Biserial Correlation coefficient would also be reversed.
Partial Correlation:

In our data set, Exam Performance is negatively related to Exam Anxiety but positively related with Revision Time.

Lets assume that Anxiety explains x% of variance in the Exam performance and Revision Time explains y% of variance in Exam performance while Revision Time also explains z% of the variation in Anxiety. So, all three variables are interrelated.
Now, if we want to find out the unique portions of variance then we need to do a partial correlation.

A Correlation between two variables in which the effect of other variables are held constant is known as Partial Correlation (Field, Andy, 2005)

Now, if we want to find out the unique portions of variance then we need to do a partial correlation.

A Correlation between two variables in which the effect of other variables are held constant is known as Partial Correlation (Field, Andy, 2005)
**First Order Correlation:** When only one variable is controlled.

**Second Order Correlation:** When two variables are controlled, and so on.

<table>
<thead>
<tr>
<th>Control Variables</th>
<th>Exam Performance (%)</th>
<th>Exam Anxiety</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Spent Revising</td>
<td>Correlation: 1.000</td>
<td>Significance (2-tailed): .</td>
</tr>
<tr>
<td>Exam Performance (%)</td>
<td>Correlation: -.247</td>
<td>Significance (2-tailed): .012</td>
</tr>
</tbody>
</table>

The Partial Correlation Coefficient between Exam Performance and Exam Anxiety is statistically significant but the variance explained has been reduced.

Now, \( R^2 \) for the unique \( r = (0.247)^2 = 0.061009 \)

Or in percentage: 6.1009%.

Thus, the unique variance in Exam Performance explained by Exam Anxiety is only 6.1%.