Covariance and Pearson r

Dr. J. Kyle Roberts

Southern Methodist University
Simmons School of Education and Human Development
Department of Teaching and Learning
Bivariate Data

- Up until now, we have only been looking at univariate measures. Now we will begin looking at bivariate measures of relationship.
- Typically, we have two measures (although there are more) that we frequently consult.
  - Covariance
  - Pearson r
- IMPORTANT: Correlation does not imply causation!
Formula for the Covariance

- Covariance is sometimes referred to as the sum of the cross-products.
- The covariance tells us if there is any relationship between the two variables.
- A positive value means that as one variable goes “up” the other variable tends to also go “up.”

\[ COV_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1} \]
Computing $COV_{xy}$ for a Given Variable Set

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>x</th>
<th>y</th>
<th>xy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>-1.5</td>
<td>-2</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>-0.5</td>
<td>-1</td>
<td>0.5</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>1.5</td>
<td>3</td>
<td>4.5</td>
</tr>
</tbody>
</table>

\[
COV_{xy} = \frac{\sum xy}{n-1} = \frac{8}{4-1} = 2.667
\]

> cov(1:4, c(1, 2, 3, 6))

[1] 2.666667
Correlation

• A correlation is a symmetric, scale-invariant measure of the (linear) association between two random variables.

• The correlation is completely symmetric between the two variables. We do not assume that one is the predictor and the other is the response. In most cases we assume that both variables are being driven by an unobserved, “hidden” or “lurking” variable.

• In other words correlation between variables is an observed or empirical trait. It does not imply causation.
Pearson correlation

- The Pearson correlation
  
  \[ r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \]
  
  \[ = \frac{COV_{xy}}{SD_x SD_y} \]

  is the most common measure of correlation.

- Both \( r \) and \( \rho \) are dimensionless and restricted to \([-1, 1]\).

- A correlation (theoretical or empirical) of 0 implies no linear dependence of the variables. If you assume a bivariate normal distribution it also implies independence of \( X \) and \( Y \).

- A correlation of \( \pm 1 \) implies a perfect linear dependence between the variables.
Heuristic Data Generation

```r
> set.seed(12346)
> cov.mat <- matrix(c(225, 200, 30, 200, 225, 15,
+     30, 15, 225), 3, 3, dimnames = list(c("reading",
+     "spelling", "math"), c("reading", "spelling",
+     "math")))
> studknow <- data.frame(mvrnorm(40, c(80, 78, 64),
+     cov.mat))
> head(studknow)

    reading spelling   math
1  103.46649  100.33448 43.75703
2   75.77614  75.87772  62.43045
3   94.60047  95.59099  38.27453
4   56.87628  49.39053  48.18428
5   62.71829  60.47098  76.07416
6  115.98872 107.98283  57.91762
```
Scatterplot matrix of heuristic measures

```r
> print(splom(~studknow, aspect = 1, type = c("g", "p")))
```
A parallel coordinate plot

An alternative graphical presentation of these three variables is called a parallel coordinate plot, because the axes for the variables are parallel, not perpendicular. The corresponding positions on consecutive pairs of axes are joined. Parallel lines indicate agreement of the scales; crossing lines indicate disagreement.

```r
> print(parallel(~studknow))
```
Computing Pearson $r$

- We can now compute the statistic for Pearson $r$ across all of our measures with:

```r
> cor(studknow)
```

```
    reading  spelling  math
reading  1.0000000 0.9288218 -0.1995084
spelling 0.9288218 1.0000000 -0.1789046
math    -0.1995084 -0.1789046 1.0000000
```

- The Pearson $r$ also answers the question for us of “How well does a single line represent the bivariate relationship between these two vectors of data?”

- By plotting this, we can see how this is true.
Plotting of bivariate relationships

$r = 0.929$

$r = -0.179$

$r = -0.200$
A Single Line Representing Relationships

\[
\begin{align*}
\text{ex1} & \leftarrow c(1, 2, 3, 4) \\
\text{ex2} & \leftarrow c(1, 2, 3, 5) \\
\text{ex3} & \leftarrow c(1, 4, 5, 2)
\end{align*}
\]
Square Before You Compare!

- The Pearson $r$ is on a ordinal scale.
- This means that an $r$ of 0.6 is greater than an $r$ of 0.5, but we do not know how much greater.
- In order to convert this to an interval scale, we must square the $r$.
- Therefore, two variables that have a 0.6 $r$ have a 11% stronger relationship than two variables with a 0.5 $r$.

$$0.6^2 - 0.5^2 = 0.36 - 0.25 = 0.11$$
Alternative measures of correlation (Spearman)

- The Pearson correlation assumes that the marginal distributions of the variables are more-or-less normal.
- Spearman’s $\rho$ uses only the ranks of the observations for each variable. As such, results are often similar to $r$, but slightly different since $r$ also takes into account relative distance.

$$\rho = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

- where $n$ is the number of paired ranks and $d$ is the difference between the paired ranks.

```r
> cor(studknow, method = "spearman")

          reading   spelling     math
reading  1.0000000  0.9105066 -0.2393996
spelling 0.9105066  1.0000000 -0.2534709
math    -0.2393996 -0.2534709  1.0000000
```
Another Example of $\rho$

```r
> height <- c(50, 45, 52, 44, 60, 38)
> weight <- c(100, 105, 122, 115, 149, 74)
> cbind(h = rank(height), w = rank(weight), d.2 = (rank(height) + rank(weight))^2)

         h  w  d.2
[1,]   4  2  4
[2,]   3  3  0
[3,]   5  5  0
[4,]   2  4  4
[5,]   6  6  0
[6,]   1  1  0

> cor(height, weight, method = "pearson")
[1] 0.8974731

> cor(height, weight, method = "spearman")
[1] 0.7714286