## Correlation Example

### Time since Ph.D. vs. Publications

<table>
<thead>
<tr>
<th>Time since Ph.D.</th>
<th>(X)</th>
<th>(X - \bar{X})</th>
<th>((X - \bar{X})^2)</th>
<th>(Y)</th>
<th>(Y - \bar{Y})</th>
<th>((Y - \bar{Y})^2)</th>
<th>((X - \bar{X})(Y - \bar{Y}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>-4.67</td>
<td>21.81</td>
<td></td>
<td>18</td>
<td>-1.93</td>
<td>3.74</td>
<td>9.02</td>
</tr>
<tr>
<td>6</td>
<td>-1.67</td>
<td>2.79</td>
<td></td>
<td>3</td>
<td>-16.93</td>
<td>286.74</td>
<td>28.22</td>
</tr>
<tr>
<td>3</td>
<td>-4.67</td>
<td>21.81</td>
<td></td>
<td>2</td>
<td>-17.93</td>
<td>321.60</td>
<td>83.69</td>
</tr>
<tr>
<td>8</td>
<td>0.33</td>
<td>0.11</td>
<td></td>
<td>17</td>
<td>-2.93</td>
<td>8.60</td>
<td>-0.98</td>
</tr>
<tr>
<td>9</td>
<td>1.33</td>
<td>1.77</td>
<td></td>
<td>11</td>
<td>-8.93</td>
<td>79.80</td>
<td>-11.91</td>
</tr>
<tr>
<td>6</td>
<td>-1.67</td>
<td>2.79</td>
<td></td>
<td>6</td>
<td>-13.93</td>
<td>194.14</td>
<td>23.22</td>
</tr>
<tr>
<td>16</td>
<td>8.33</td>
<td>69.39</td>
<td></td>
<td>38</td>
<td>18.07</td>
<td>326.40</td>
<td>150.56</td>
</tr>
<tr>
<td>10</td>
<td>2.33</td>
<td>5.43</td>
<td></td>
<td>48</td>
<td>28.07</td>
<td>787.74</td>
<td>65.49</td>
</tr>
<tr>
<td>2</td>
<td>-5.67</td>
<td>32.15</td>
<td></td>
<td>9</td>
<td>-10.93</td>
<td>119.54</td>
<td>61.96</td>
</tr>
<tr>
<td>5</td>
<td>-2.67</td>
<td>7.13</td>
<td></td>
<td>22</td>
<td>2.07</td>
<td>4.27</td>
<td>-5.51</td>
</tr>
<tr>
<td>5</td>
<td>-2.67</td>
<td>7.13</td>
<td></td>
<td>30</td>
<td>10.07</td>
<td>101.34</td>
<td>-26.84</td>
</tr>
<tr>
<td>6</td>
<td>-1.67</td>
<td>2.79</td>
<td></td>
<td>21</td>
<td>1.07</td>
<td>1.14</td>
<td>-1.78</td>
</tr>
<tr>
<td>7</td>
<td>-0.67</td>
<td>0.45</td>
<td></td>
<td>10</td>
<td>-9.93</td>
<td>98.67</td>
<td>6.62</td>
</tr>
<tr>
<td>11</td>
<td>3.33</td>
<td>11.09</td>
<td></td>
<td>27</td>
<td>7.07</td>
<td>49.94</td>
<td>23.56</td>
</tr>
<tr>
<td>18</td>
<td>10.33</td>
<td>106.71</td>
<td></td>
<td>37</td>
<td>17.07</td>
<td>291.27</td>
<td>176.36</td>
</tr>
</tbody>
</table>

\[
\bar{X} = 7.67 \quad \sum (X - \bar{X})^2 = 293.33 \quad \bar{Y} = 19.93 \quad \sum (Y - \bar{Y})^2 = 2674.93 \quad \sum (X - \bar{X})(Y - \bar{Y}) = 581.67
\]

\[
r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}} = \frac{581.67}{\sqrt{(293.33)(264.93)}} = .66
\]

\[
t = \frac{r \sqrt{N-2}}{\sqrt{1-r^2}} = \frac{.66 \sqrt{13}}{\sqrt{1-.66^2}} = \frac{2.38}{.75} = 3.17
\]

\[
df = N-2 = 15-2 = 13, t_{crit, .05} = 2.160
\]

Confidence limits seem to be rarely reported, perhaps because software packages do not often compute them and perhaps because they are a bit complicated to calculate. They are useful for having a sense of sampling variability of the sample correlation estimate, however, and should probably be used more often. To compute them, the \(r\) value must be converted to a \(z\) value first, then the intervals are calculated, and then the interval values are converted back to the \(r\) scale. (my values differ somewhat from those obtained with SPSS and R below because of rounding.) I show the computations below, but I have created an Excel spreadsheet that will do these computations automatically. Visit the class website to download it.

\[
z = .5 \left[ \ln \left( \frac{1+r}{1-r} \right) \right] = .5 \left[ \ln \left( \frac{1+.66}{1-.66} \right) \right] = .787 \quad SE_z = \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{15-3}} = .289
\]

\[
z \pm (1.96)(SE_z) = .874 \pm (1.96)(.289) = .222, 1.353
\]

\[
r = \frac{e^{z^2} - 1}{e^{z^2} + 1} = \frac{e^{.787^2} - 1}{e^{.787^2} + 1} = .559 = .218
\]

\[
r = \frac{e^{z^2} - 1}{e^{z^2} + 1} = \frac{e^{1.96^2} - 1}{e^{1.96^2} + 1} = 13.969 = .875
\]

\[\text{1 Numerical example from Cohen, Cohen, West, and Aiken (2003) Table 2.2.2.}\]
SPSS Syntax

graph
   /scatterplot(bivar)=numpubs with yrsphd.

correlations vars=yrsphd numpubs.

Correlations

```
R
> #clear active frame from previous analyses
> rm(mydata)
> library(haven)
> mydata = read_sav("c:/jason/spsswin/uvclass/ccwa2_2_2.sav")
> Correlation(YRSPHD, NUMPUBS, data=mydata)

Correlation Analysis for Variables YRSPHD and NUMPUBS
-----------------------------------------------------
>>> Pearson's product-moment correlation
Number of paired values with neither missing, n = 15
Number of cases (rows of data) deleted: 0

Sample Covariance: s = 41.548
Sample Correlation: r = 0.657

Hypothesis Test of 0 Correlation: t = 3.139, df = 13, p-value = 0.008
95% Confidence Interval for Correlation: to
```
> #base R function
> #cor.test(mydata$YRSPHD,mydata$NUMPUBS)
> #scatterplot
> ScatterPlot(YRSPHD, NUMPUBS)

>>> Suggestions
Plot(YRSPHD, NUMPUBS, fit="lm", fit_se=c(.90,.99))  # fit line, standard errors
Plot(YRSPHD, NUMPUBS, out_cut=.10)  # label top 10% potential outliers
Plot(YRSPHD, NUMPUBS, enhance=TRUE)  # many options

>>> Pearson's product-moment correlation

Number of paired values with neither missing, n = 15

Sample Correlation of YRSPHD and NUMPUBS: \( r = 0.657 \)

Hypothesis Test of 0 Correlation: \( t = 3.139, \; df = 13, \; p\text{-value} = 0.008 \)
95% Confidence Interval for Correlation: \( 0.2 \) to \( 0.9 \)

Example Write-up

A Pearson correlation coefficient was computed to examine the relationship between the amount of time since a faculty member received a PhD and the number of peer-reviewed publications. There was a significant positive correlation between the time since receiving the PhD and the number of publications, \( r = .66, \; p < .01, \; 95\% \text{[CI = .22, .88]} \). Approximately 44\% of the variance was shared between the two variables, \( r^2 = .44 \).