# Correlation Example[1]

| **Time since Ph.D.** | | | **Publications** | | | |
|---|---|---|---|---|---|---|
| $X$ | $X-\bar{X}$ | $(X-\bar{X})^2$ | $Y$ | $Y-\bar{Y}$ | $(Y-\bar{Y})^2$ | $(X-\bar{X})(Y-\bar{Y})$ |
| **3** | -4.67 | 21.81 | **18** | -1.93 | 3.74 | 9.02 |
| **6** | -1.67 | 2.79 | **3** | -16.93 | 286.74 | 28.22 |
| **3** | -4.67 | 21.81 | **2** | -17.93 | 321.60 | 83.69 |
| **8** | 0.33 | 0.11 | **17** | -2.93 | 8.60 | -0.98 |
| **9** | 1.33 | 1.77 | **11** | -8.93 | 79.80 | -11.91 |
| **6** | -1.67 | 2.79 | **6** | -13.93 | 194.14 | 23.22 |
| **16** | 8.33 | 69.39 | **38** | 18.07 | 326.40 | 150.56 |
| **10** | 2.33 | 5.43 | **48** | 28.07 | 787.74 | 65.49 |
| **2** | -5.67 | 32.15 | **9** | -10.93 | 119.54 | 61.96 |
| **5** | -2.67 | 7.13 | **22** | 2.07 | 4.27 | -5.51 |
| **5** | -2.67 | 7.13 | **30** | 10.07 | 101.34 | -26.84 |
| **6** | -1.67 | 2.79 | **21** | 1.07 | 1.14 | -1.78 |
| **7** | -0.67 | 0.45 | **10** | -9.93 | 98.67 | 6.62 |
| **11** | 3.33 | 11.09 | **27** | 7.07 | 49.94 | 23.56 |
| **18** | 10.33 | 106.71 | **37** | 17.07 | 291.27 | 176.36 |
| $\bar{X}=7.67$ | | $\sum(X-\bar{X})^2=293.33$ | $\bar{Y}=19.93$ | | $\sum(Y-\bar{Y})^2=2674.93$ | $\sum(X-\bar{X})(Y-\bar{Y})=581.67$ |

$$r=\frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2\sum(Y-\bar{Y})^2}}=\frac{581.67}{\sqrt{(293.33)(2674.93)}}=.66$$

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

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

Confidence limits seem to be rarely reported, perhaps because software packages have often not computed them in the past and perhaps because they are a bit complicated to calculate than other types of confidence intervals.[2] 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. You will not need to compute these by hand, but just for illustrating, I have included the computation ($\ln$ is the natural log function and $\mathrm{e}$ is Euler's mathematical constant).

$$z=.5\left[\ln\left(\frac{1+r}{1-r}\right)\right]=.5\left[\ln\left(\frac{1+.66}{1-.66}\right)\right]=.787$$

$$SE_z=\sqrt{\frac{1}{n-3}}=\sqrt{\frac{1}{15-3}}=.289$$

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

$$r=\frac{e^{2z}-1}{e^{2z}+1}=\frac{e^{2\times .222}-1}{e^{2\times .222}+1}=\frac{.559}{2.559}=.218$$

$$r=\frac{e^{2z}-1}{e^{2z}+1}=\frac{e^{2\times 1.353}-1}{e^{2\times 1.353}+1}=\frac{13.969}{15.969}=.875$$

---

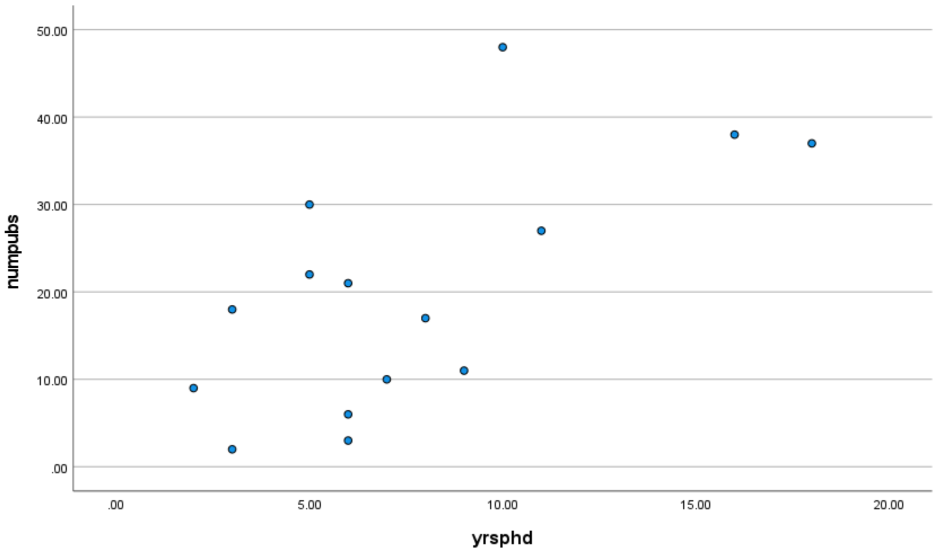
[1] Numerical example from Cohen, Cohen, West, and Aiken (2003) Table 2.2.2. Note that standard error in this text is $SE$ rather than $s$, which was used in Myers et al. text, and the current book often uses $r_{XY}$ in place of $r$ in many places.

[2] Until the most recent version of SPSS, confidence intervals for correlations were not available through the CORRELATIONS procedure. They are available now, so I don't anticipate that you will ever need to compute confidence intervals manually. But just in case the occasion arises, I have created an Excel spreadsheet that will do these computations automatically here: http://web.pdx.edu/~newsomj/uvclass/correlation confidence intervals.xlsx.

## SPSS Syntax

```
graph
  /scatterplot(bivar)=numpubs with yrsphd.

 correlations vars=yrsphd numpubs
        /ci.
```

## Correlations




**Correlations**

| | | yrsphd | numpubs |
|---|---|---|---|
| yrsphd | Pearson Correlation | 1 | .657 |
| | Sig. (2-tailed) | | .008 |
| | N | 15 | 15 |
| numpubs | Pearson Correlation | .657 | 1 |
| | Sig. (2-tailed) | .008 | |
| | N | 15 | 15 |

**Confidence Intervals**

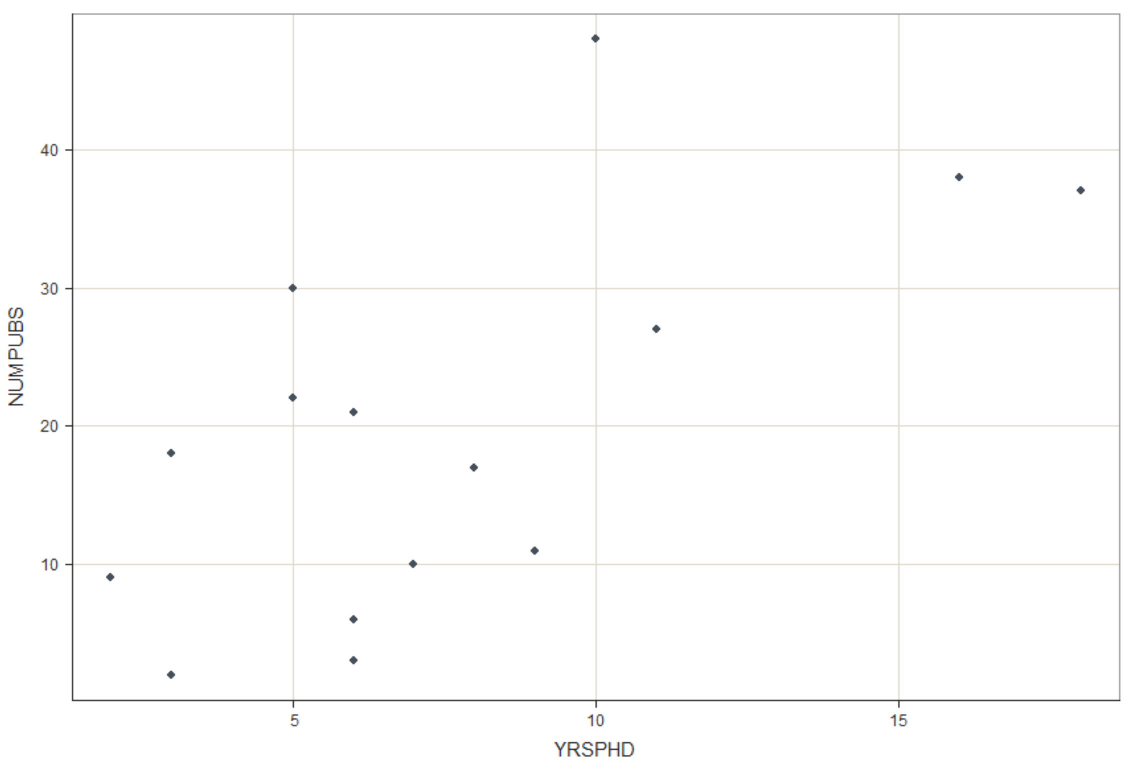
| | Pearson Correlation | Sig. (2-tailed) | 95% Confidence Intervals (2-tailed)[a] | |
|---|---|---|---|---|
| | | | Lower | Upper |
| yrsphd - numpubs | .657 | .008 | .218 | .875 |

a. Estimation is based on Fisher's r-to-z transformation.

## R

```
> #clear active frame from previous analyses
> rm(d)
> #clear console
> cat("\014")
> library(haven)
> d = read_sav("c:/jason/spsswin/uvclass/ccwa2_2_2.sav")
>
>
> #add digits to get more digits for CIs
> Correlation(YRSPHD, NUMPUBS, digits_d = 3, data=d)

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

> #an alternative method
> #base R function
> #cor.test(d$YRSPHD,d$NUMPUBS)
>
> #scatterplot (this function also reports the correlation)
> 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-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 = .008, 95% [CI = .22, .88]. Approximately 44% of the variance was shared between the two variables, $r^2$ = .44.