

Associations, Correlations, and Causation



Spearman rank correlation (Nonparametric correlation)

- Pearson's correlation coefficient is very sensitive to extreme values
- Spearman rank correlation calculates the Pearson correlation on the ranks of each variable
- The Pearson correlation coefficient is calculated, but the data values are replaced by the ranks

Spearman rank correlation (nonparametric)

- The Spearman rank correlation ranges between -1 and 1 as does the Pearson correlation
- We can test the null hypothesis that $\rho=0$
- The test statistic for $n>10$ is with $n-2$ degrees of freedom is compared to the t distribution

```
spearman cd4count age, stats(rho obs p)
```

```
Number of obs =      999  
Spearman's rho =     -0.1319
```

```
Test of Ho: cd4count and age_b are independent  
Prob > |t| =      0.0000
```

$$t_s = r_s \sqrt{\frac{n-2}{1-r_s^2}}$$

Matrix of Pearson correlations

- We can calculate a correlation matrix by listing multiple variable names
- Beware of which n's are used (use listwise option to get all n's equal so you are testing similar relationships)

```
pwcorr cd4count age_nsupport_b auditc, sig obs
```

	cd4count	age_b	nsupport_b	auditc
cd4count	1.0000			
	999			
age_b	-0.1133	1.0000		
	0.0003			
	999	3387		
nsupport_b	0.0067	0.3597	1.0000	
	0.8340	0.0000		
	994	3370	3372	
auditc	0.0670	0.0580	0.0468	1.0000
	0.0392	0.0009	0.0078	
	948	3242	3230	3244

Matrix of Pearson correlations

- Using listwise option to get all n's equal

```
pwcorr cd4count age_b nsupport_b auditc, sig obs listwise
```

	cd4count	age_b	nsupport_b	auditc
cd4count	1.0000			
	944			
age_b	-0.1053	1.0000		
	0.0012	944		
nsupport_b	0.0107	0.3209	1.0000	
	0.7416	0.0000	944	
auditc	0.0668	-0.0075	0.0475	1.0000
	0.0401	0.8181	0.1449	944
	944	944	944	944

Matrix of Spearman correlations

```
spearman cd4count age nsupport_b auditc, pw stats(rho obs p)
```

```
+-----+
| Key |
+-----+
| rho |
| Number of obs |
| Sig. level |
+-----+
```

```

      | cd4count   age_b nsupport_b   auditc
-----+-----
cd4count | 1.0000
          |      999
          |
          |
          | age_b | -0.1319   1.0000
          |      999   3387
          |      0.0000
          |
          | nsupport_b | 0.0118   0.4055   1.0000
          |      994   3370   3372
          |      0.7093   0.0000
          |
          | auditc | 0.0945   0.0777   0.0447   1.0000
          |      948   3242   3230   3244
          |      0.0036   0.0000   0.0111
          |

```

Matrix of Spearman correlations

```
spearman cd4count age nsupport_b auditc, stats(rho obs p)
```

```
+-----+
| Key          |
|-----|
| rho          |
| Number of obs |
| Sig. level    |
+-----+
```

Here if you drop the “pw” option you get all n’s equal

```

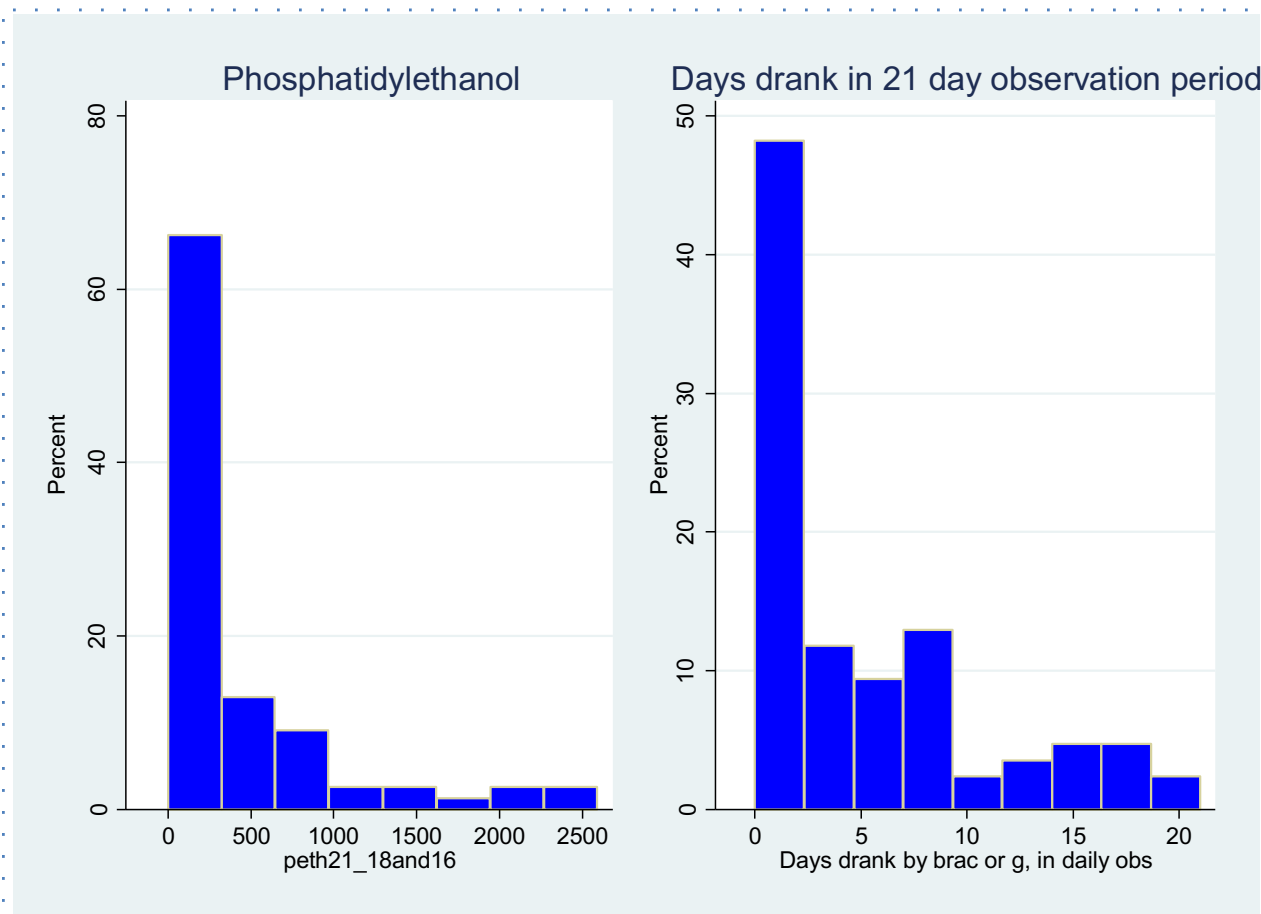
               | cd4count   age_b nsupport_b   auditc
-----+-----+-----+-----+-----+
cd4count |    1.0000
          |      944
          |
          |
age_b    |   -0.1261   1.0000
          |      944      944
          |    0.0001
          |
nsupport_b |    0.0127   0.3546   1.0000
          |      944      944      944
          |    0.6971   0.0000
          |
auditc   |    0.0940  -0.0116   0.0431   1.0000
          |      944      944      944      944
          |    0.0039   0.7221   0.1859
          |

```

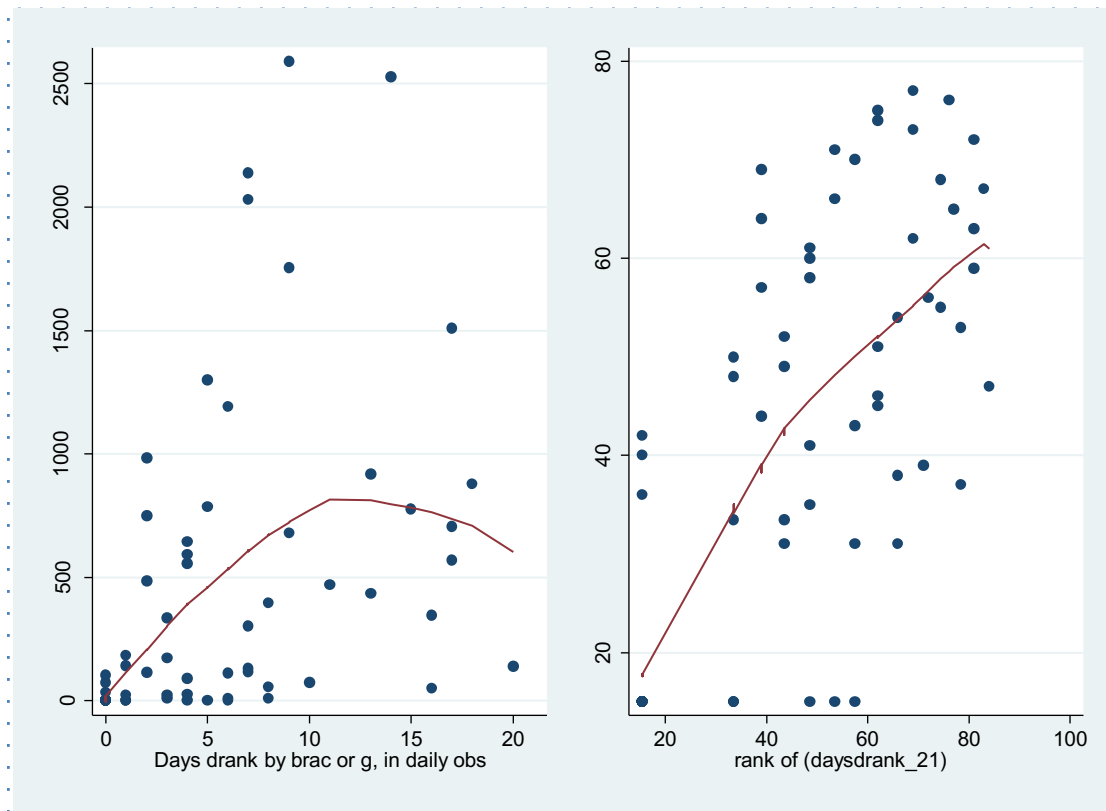
Pearson vs. Spearman

	Pearson	Spearman
Sensitivity to outliers	more sensitive	less sensitive
Can be used with ordinal data	no	yes
Assumes data are from a normal distribution	yes	no
Detects non-linear relationship	no	yes

Biomarker and alcohol consumption



PEth Biomarker of alcohol consumption vs. days drinking (raw data vs. ranks)



Pearson and Spearman correlations

```
. pwcorr peth21_18and16 daysdrank_21, obs sig
```

```
-----+-----  
| peth2~16 days~_21  
-----+-----  
peth21_18~16 | 1.0000  
|  
| 77  
|  
daysdrank_21 | 0.4717 1.0000  
| 0.0000  
| 77 85  
|  
-----+-----
```

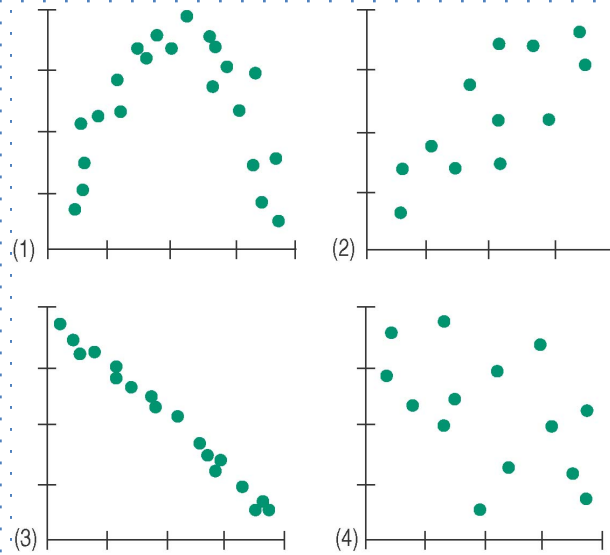
```
. spearman peth21_18and16 daysdrank_21
```

```
Number of obs = 77  
Spearman's rho = 0.7413  
Test of Ho: peth21_18and16 and daysdrank_21 are independent  
Prob > |t| = 0.0000
```

Association Identification

Which scatterplot below shows

- a) little association?
- b) negative association?
- c) linear association?
- d) moderately strong association?
- e) strong association?



Association Identification

Matching.

Here are several scatterplots.

The calculated correlations are:

a) -0.977

b) -0.021

c) 0.736

d) 0.951

Which is which?

