A nonparametric measure of correlation

- we have seen that Pearson's correlation coefficient
  - measures only linear association between variables
  - can be greatly affected by outlying values
- Spearman's correlation coefficient is designed to overcome these problems
- to calculate Spearman's rho
  - rank the x and y values separately
  - calculate the usual (Pearson) coefficient on the ranks

Example: The data on the diameter and useable volume of wood is given below with the ranks, calculated separately for each variable.

diameter	rank	volume	rank
36	15.0	192	15
28	10.5	113	11
28	10.5	88	10
41	20.0	294	20
19	3.5	28	4
32	13.0	123	12
22	6.0	51	6
38	17.0	252	18
25	8.5	56	7
17	1.5	16	1
31	12.0	141	13
20	5.0	32	5
25	8.5	86	9
19	3.5	21	2
39	18.5	231	17
33	14.0	187	14
17	1.5	22	3
37	16.0	205	16
23	7.0	57	8
39	18.5	265	19

- the Pearson correlation (on the original values) is

```
MTB> corr c1 c2

Correlations: C1, C2

Pearson correlation of C1 and C2 = 0.976
```

- the Spearman correlation is the Pearson correlation of the ranks

```
MTB > corr c3 c4

Correlations: C3, C4

Pearson correlation of C3 and C4 = 0.989
```

- the Spearman value is larger, reflecting the curvature in the plot of the data

Example: The bottom right panel of the figure showing various correlations was dominated by one disparate value. The values and their ranks are shown below.

x	rank	y	rank
8	1	7	5
9	2	6	4
10	3	5	3
11	4	4	2
12	5	3	1
20	6	15	6

- the Pearson correlation is $r = .79$
- the Spearman correlation is $r_s = -.14$
- the Spearman measure has downweighted the unusual value
- when the two quantities are quite different, it is important to investigate whether there are unusual values or a curved relationship