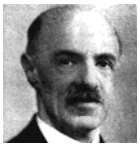
Spearman rank correlation



Developed 1904

Charles E. Spearman 1863-1945

Spearman rank correlation

- **Spearman rank correlation**
- it is a nonparametric measure of the relationship between 2 sets of ordinal (ranked) values
- calculation
  - where d is the difference between ranks

$$r_s = 1 - \frac{6\sum d^2}{n^3 - n}$$

Perfect Correlations of Two Ranked Variables

Case 1: Positive correlation

Variable X ranks	Variable Y ranks	Difference in ranks (d)	d^2
1	1	0	0
2	2	0	0
3	3	0	0
4	4	0	0
5	5	0	0
			$\Sigma d^2 = 0$

$$r_s = 1 - \frac{6(\Sigma d^2)}{N^3 - N} = 1 - \frac{6(0)}{5^3 - 5} = 1 - \frac{0}{120} = 1.0$$

Case 2: Negative correlation

Variable X ranks	Variable Y ranks	Difference in ranks (d)	d^2
1	5	−4	16
2	4	−2	4
3	3	0	0
4	2	2	4
5	1	4	16
			$\Sigma d^2 = 40$

$$r_s = 1 - \frac{6(\Sigma d^2)}{N^3 - N} = 1 - \frac{6(40)}{5^3 - 5} = 1 - \frac{240}{120} = -1.0$$

example

river	Catchment	rank	Discharge	rank	d	d^2
yellow	672	7	3.3	7	0	0
ganges	956	5	11.7	4	1	1
amazon	5775	1	175.0	1	0	0
missi.	3269	2	18.4	3	1	1
mekong	795	6	11.0	5	1	1
indus	969	4	5.6	6	2	4
yangtze	1942	3	22.0	2	1	1
					Σd^2	8

$$r_s = 1 - \frac{(6*8)}{(7^3 - 7)} = 1 - \frac{48}{336} = 0.86$$

- significance
  - H_0: there is no rank correlation
  - H_1: may be directional or nondirectional
  - a nondirectional test is a 2 tailed test
  - a directional test is a 1 tailed test
  - the df is the number of pairs of ranked values
  - you can use a table if the sample size is small <100

- but more generally the test statistic is
- where n is the number of observations
  - this might be a case where a one tailed test is desirable since most researchers have an idea of the direction of the sign of the coefficient

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

- we reject H_0, 2 tailed critical value=2.365
- H_0: no association between 2 variables
- H_1: association between 2 variables - 2 tailed
- H_1: +/- association between 2 variables - 1 tailed

$$t = 0.86\sqrt{\frac{7-2}{1-0.86^2}} = 0.86 * 4.38 = 3.76$$

Example 2

- It has been suggested that nonmetropolitan growth is increasing
- If so then there should be a relationship between population density and population growth
- Let's test and see

Spearman Correlation Example: State Population Change and Density, 1990–95

State*	Original data population: Percentage change 1990–95	Original data population: Density** 1990	Ranked data population: Percentage change 1990–95	Ranked data population: Density 1990	Difference (d)	d^2
Alabama	5.47	79.60	28	26	2	4
Alaska	5.80	1.00	32	1	31	961
Arizona	17.54	32.29	49	14	35	1225
Arkansas	5.53	45.14	29	16	13	169
California	6.04	190.40	33	39	−6	36
Colorado	13.60	31.80	46	13	33	1089
Vermont	3.55	60.71	15	21	−6	36
Virginia	6.69	155.83	35	36	−1	1
Washington	11.69	73.18	45	23	22	484
W. Virginia	1.62	74.34	8	25	−17	289
Wisconsin	4.52	89.88	22	27	−5	25
Wyoming	5.29	4.68	26	2	24	576

*Data listed for only 12 states
**Persons per square mile

$$r_s = 1 - \frac{6(\Sigma d^2)}{N^3 - N}$$

$$r_s = 1 - \frac{6(28968)}{50^3 - 50} = 1 - \frac{173808}{124950} = 1 - 1.391 = -.391$$

Source: Bureau of the Census, Dept. of Commerce.

Spearman Correlation Coefficients for State Population Change and Density

Time period	Spearman r_s
1960–1965	+.166
1965–1970	+.176
1970–1975	−.539
1975–1980	−.561
1980–1985	−.406
1985–1990	+.177
1990–1995	−.391

- correction for tied ranks
- from a practical viewpoint it is often not worth correcting for ties
- use of correction is advised if
  - 1) when 3 or more observations are tied equally
  - 2) when the number of pairs of ties is more than 1/4 of the number of observations

- the formula is
  - A=(n^3-n)/12
  - B=3((t_x^3-t_x)/12)
  - C=3((t_y^3-t_y)/12)
  - where t_x is the number of values of variable x tying at a given rank
  - t_y is the same for y
  - the effect of the correction for ties is to increase the value of r_s making it easier to reject the null hypothesis

$$r_s = \frac{(A - B) + (A - C) - \sum d^2}{2\sqrt{(A - B)(A - C)}}$$

Advantages

- Its nonparametric so you don't have to have a normal distribution
- Its less effected by outliers

- In principle, r_s is simply a special case of the Pearson product-moment coefficient in which the data are converted to ranks before calculating the coefficient.
- As we've seen, a simpler procedure is normally used to calculate it