# Spearman’s correlation

## Introduction

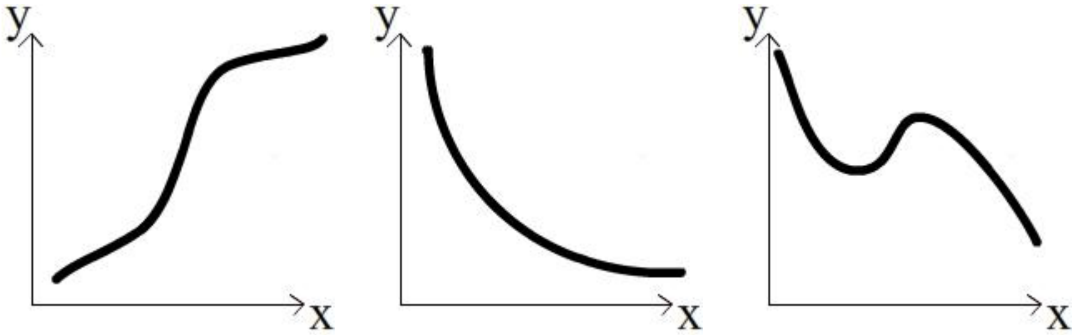
Before learning about Spearman’s correllation it is important to understand Pearson’s correlation which is a statistical measure of the strength of a *linear* relationship between paired data. Its calculation and subsequent significance testing of it requires the following data assumptions to hold:

- interval or ratio level;
- linearly related;
- bivariate normally distributed.

If your data does not meet the above assumptions then use Spearman’s rank correlation!

## Monotonic function

To understand Spearman’s correlation it is necessary to know what a monotonic function is. A monotonic function is one that either never increases or never decreases as its independent variable increases. The following graphs illustrate monotonic functions:



Monotonically increasing Monotonically decreasing Not monotonic

- Monotonically increasing - as the x variable increases the y variable never decreases;
- Monotonically decreasing - as the x variable increases the y variable never increases;
- Not monotonic - as the x variable increases the y variable sometimes decreases and sometimes increases.

## Spearman’s correlation coefficient

Spearman’s correlation coefficient is a statistical measure of the strength of a *monotonic* relationship between paired data. In a sample it is denoted by $r_s$ and is by design constrained as follows

$$-1 \leq r_s \leq 1$$

And its interpretation is similar to that of Pearsons, e.g. the closer $r_s$ is to $\pm 1$ the stronger the monotonic relationship. Correlation is an effect size and so we can verbally describe the strength of the correlation using the following guide for the absolute value of $r_s$:

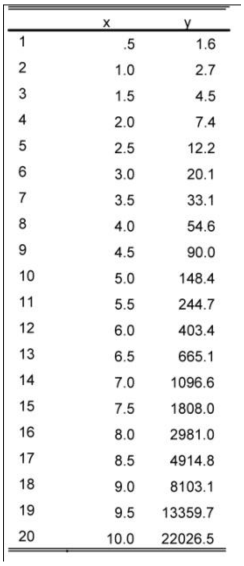
- .00-.19 “very weak”
- .20-.39 “weak”
- .40-.59 “moderate”
- .60-.79 “strong”
- .80-1.0 “very strong”

The calculation of Spearman’s correlation coefficient and subsequent significance testing of it requires the following data assumptions to hold:

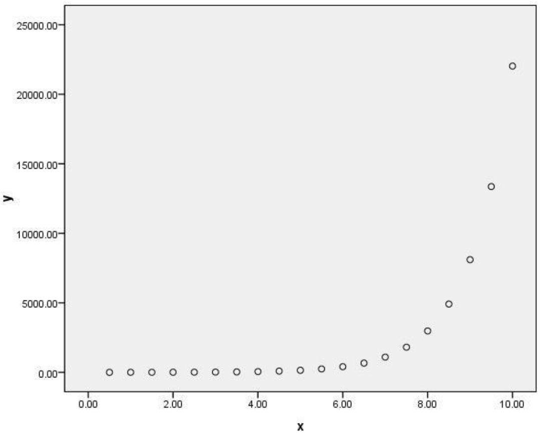
- interval or ratio level or ordinal;
- monotonically related.

Note, unlike Pearson’s correlation, there is no requirement of normality and hence it is a nonparametric statistic.

Let us consider some examples to illustrate it. The following table gives x and y values for the relationship $y = \exp(x)$. From the graph we can see that this is a perfectly increasing monotonic relationship.



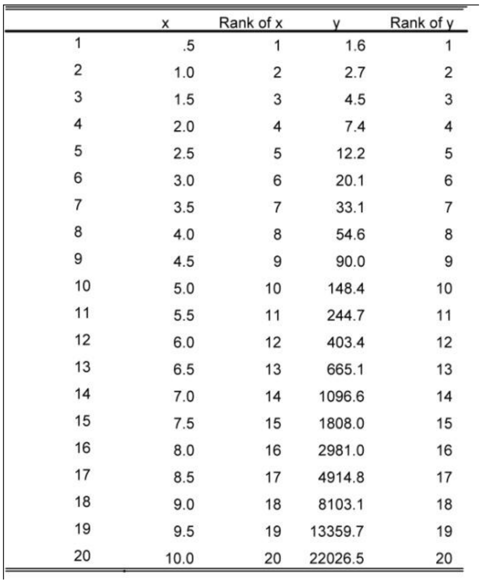
| | x | y |
|---|---|---|
| 1 | .5 | 1.6 |
| 2 | 1.0 | 2.7 |
| 3 | 1.5 | 4.5 |
| 4 | 2.0 | 7.4 |
| 5 | 2.5 | 12.2 |
| 6 | 3.0 | 20.1 |
| 7 | 3.5 | 33.1 |
| 8 | 4.0 | 54.6 |
| 9 | 4.5 | 90.0 |
| 10 | 5.0 | 148.4 |
| 11 | 5.5 | 244.7 |
| 12 | 6.0 | 403.4 |
| 13 | 6.5 | 665.1 |
| 14 | 7.0 | 1096.6 |
| 15 | 7.5 | 1808.0 |
| 16 | 8.0 | 2981.0 |
| 17 | 8.5 | 4914.8 |
| 18 | 9.0 | 8103.1 |
| 19 | 9.5 | 13359.7 |
| 20 | 10.0 | 22026.5 |



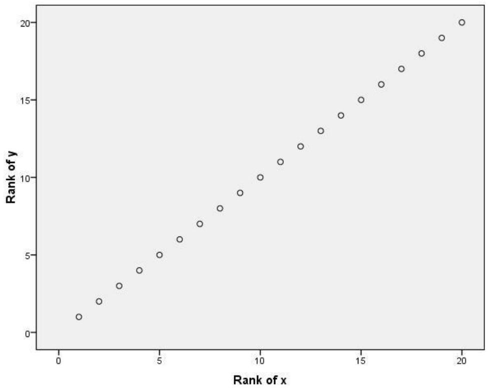
The calculation of Pearson’s correlation for this data gives a value of .699 which does not reflect that there is indeed a perfect relationship between the data. Spearman’s correlation for this data however is 1, reflecting the perfect monotonic relationship.

Spearman’s correlation works by calculating Pearson’s correlation on the ranked values of this data. Ranking (from low to high) is obtained by assigning a rank of 1 to the lowest value, 2 to the next lowest and so on.

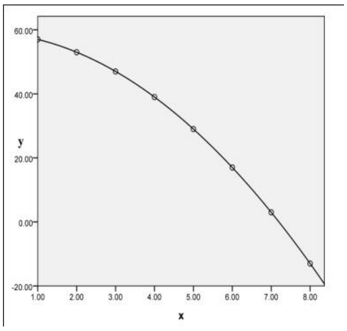
If we look at the plot of the ranked data, then we see that they are perfectly linearly related.



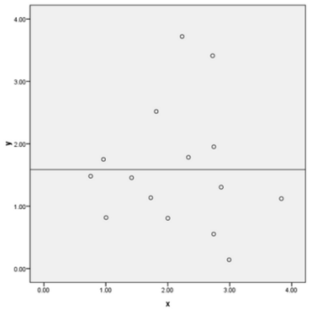
| | x | Rank of x | y | Rank of y |
|---|---|---|---|---|
| 1 | .5 | 1 | 1.6 | 1 |
| 2 | 1.0 | 2 | 2.7 | 2 |
| 3 | 1.5 | 3 | 4.5 | 3 |
| 4 | 2.0 | 4 | 7.4 | 4 |
| 5 | 2.5 | 5 | 12.2 | 5 |
| 6 | 3.0 | 6 | 20.1 | 6 |
| 7 | 3.5 | 7 | 33.1 | 7 |
| 8 | 4.0 | 8 | 54.6 | 8 |
| 9 | 4.5 | 9 | 90.0 | 9 |
| 10 | 5.0 | 10 | 148.4 | 10 |
| 11 | 5.5 | 11 | 244.7 | 11 |
| 12 | 6.0 | 12 | 403.4 | 12 |
| 13 | 6.5 | 13 | 665.1 | 13 |
| 14 | 7.0 | 14 | 1096.6 | 14 |
| 15 | 7.5 | 15 | 1808.0 | 15 |
| 16 | 8.0 | 16 | 2981.0 | 16 |
| 17 | 8.5 | 17 | 4914.8 | 17 |
| 18 | 9.0 | 18 | 8103.1 | 18 |
| 19 | 9.5 | 19 | 13359.7 | 19 |
| 20 | 10.0 | 20 | 22026.5 | 20 |



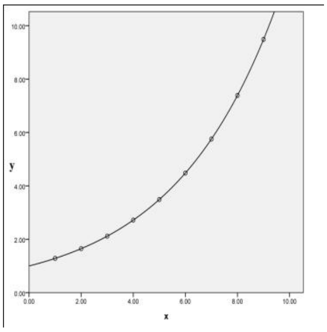
In the figures below various samples and their corresponding sample correlation coefficient values are presented. The first three represent the “extreme” monotonic correlation values of -1, 0 and 1:



$r_s = -1$
perfect –ve
monotonic correlation

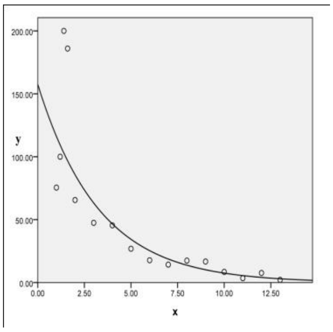


$r_s = 0$
no correlation

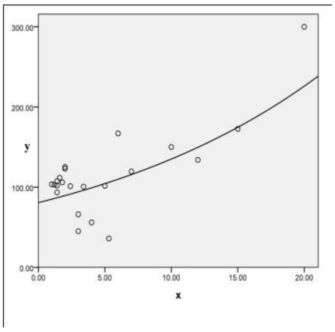


$r_s = 1$
perfect +ve
monotonic correlation

Invariably what we observe in a sample are values as follows:

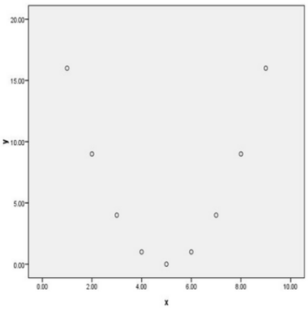


$r_s = -.941$
very strong -ve
monotonic correlation



$r_s = .372$
weak +ve
monotonic correlation

Note: Spearman’s correlation coefficient is a measure of a monotonic relationship and thus a value of $r_s = 0$ does not imply there is no relationship between the variables. For example in the following scatterplot $r_s = 0$ which implies no (monotonic) correlation however there is a perfect quadratic relationship:

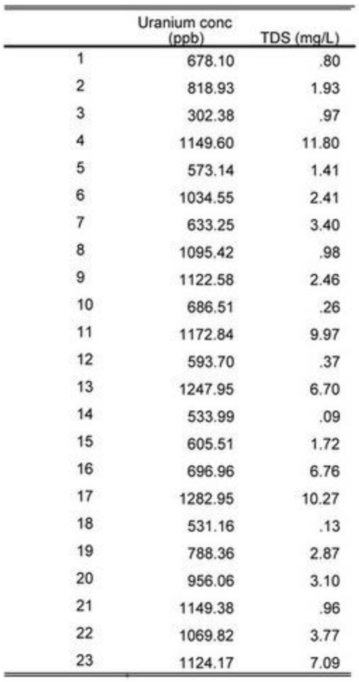


$r_s = 0$
perfect quadratic relationship

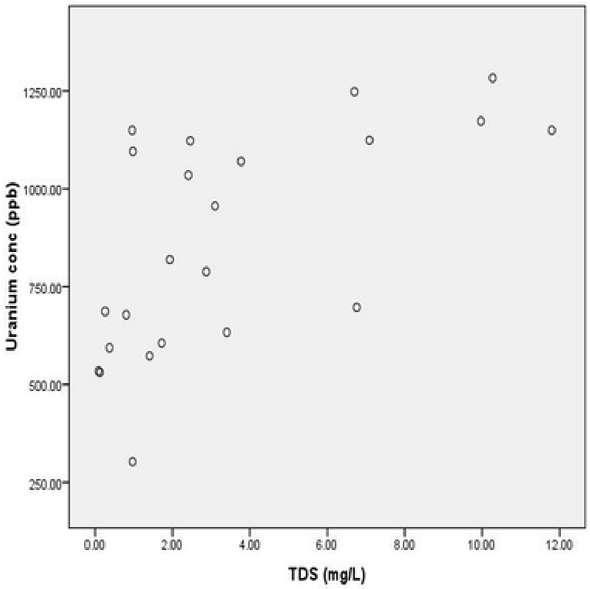
## Example

The following data comprises 23 groundwater samples that were collected recording the Uranium concentration (ppb) and the total dissolved solids (mg/L). It is of interest to know if the two variables are correlated?

We should initial consider if Pearson's correlation is appropriate or whether we should resort to Spearman's if there are assumption violations.

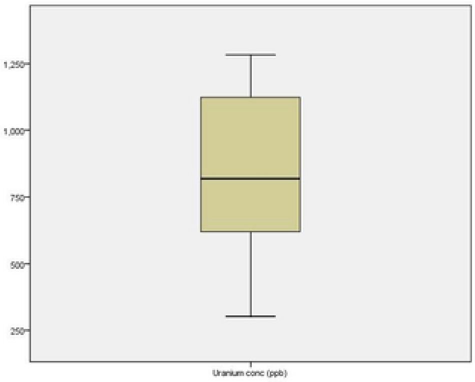


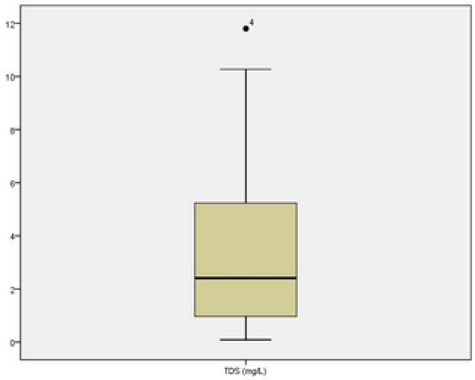
| | Uranium conc (ppb) | TDS (mg/L) |
|---|---|---|
| 1 | 678.10 | .80 |
| 2 | 818.93 | 1.93 |
| 3 | 302.38 | .97 |
| 4 | 1149.60 | 11.80 |
| 5 | 573.14 | 1.41 |
| 6 | 1034.55 | 2.41 |
| 7 | 633.25 | 3.40 |
| 8 | 1095.42 | .98 |
| 9 | 1122.58 | 2.46 |
| 10 | 686.51 | .26 |
| 11 | 1172.84 | 9.97 |
| 12 | 593.70 | .37 |
| 13 | 1247.95 | 6.70 |
| 14 | 533.99 | .09 |
| 15 | 605.51 | 1.72 |
| 16 | 696.96 | 6.76 |
| 17 | 1282.95 | 10.27 |
| 18 | 531.16 | .13 |
| 19 | 788.36 | 2.87 |
| 20 | 956.06 | 3.10 |
| 21 | 1149.38 | .96 |
| 22 | 1069.82 | 3.77 |
| 23 | 1124.17 | 7.09 |



The scatterplot suggests a definite positive correlation between Uranium and TDS. However, there is possibly slight evidence of non-linearity for TDS values close to zero. However, this is debateable and so we shall move on and consider the other normality assumption.

We need to perform some normality checks for the two variables. One simple way of doing this is to examine boxplots of the data. These are given below.

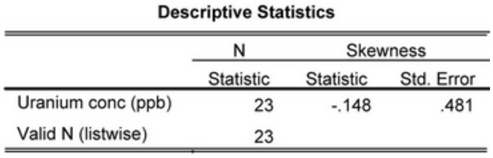




The boxplot for Uranium is fairly consistent with one from a normal distribution; the median is fairly close to the centre of the box and the whiskers are of approximate equal length.

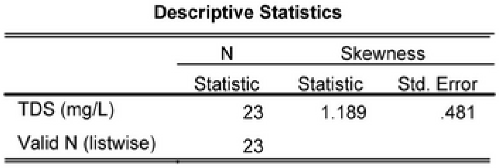
The boxplot for TDS is slightly disturbing in that the median is close to the lower quartile and the lower whisker is shorter than the upper one, which would be suggesting positive skewness. Also there is an outlier and Pearson's correlation is sensitive to these as well as skewness.

Since we have some doubts over normality, we shall examine the skewness coefficients to see if there is further evidence to suggest whether either of the variables is skewed.



**Descriptive Statistics**

| | N | Skewness | |
|---|---|---|---|
| | Statistic | Statistic | Std. Error |
| Uranium conc (ppb) | 23 | -.148 | .481 |
| Valid N (listwise) | 23 | | |

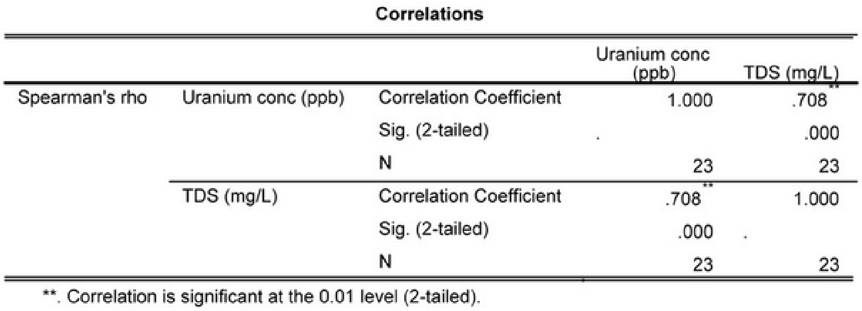


**Descriptive Statistics**

| | N | Skewness | |
|---|---|---|---|
| | Statistic | Statistic | Std. Error |
| TDS (mg/L) | 23 | 1.189 | .481 |
| Valid N (listwise) | 23 | | |

A quick check to see if the skewness coefficients are not sufficiently large to warrant concern is to see if the absolute values of the skewness coefficients are less than two times their standard errors. Using this guide, the Uranium data's skewness is consistent with the data being normal. However the TDS skewness coefficient appears to be large enough to warrant concern that ther is positive skewness present (1.189 > 2 x .481).

Hence we do have concerns over the normality of our data and should continue with a Spearman's correlation analysis. SPSS produces the following Spearman's correlation output:



**Correlations**

| | | | Uranium conc (ppb) | TDS (mg/L) |
|---|---|---|---|---|
| Spearman's rho | Uranium conc (ppb) | Correlation Coefficient | 1.000 | .708** |
| | | Sig. (2-tailed) | . | .000 |
| | | N | 23 | 23 |
| | TDS (mg/L) | Correlation Coefficient | .708** | 1.000 |
| | | Sig. (2-tailed) | .000 | . |
| | | N | 23 | 23 |

**. Correlation is significant at the 0.01 level (2-tailed).

The significant Spearman correlation coefficient value of 0.708 confirms what was apparent from the graph; there appears to be a strong positive correlation between the two variables. Thus large values of uranium are associated with large TDS values

However, we need to perform a significance test to decide whether based upon this sample there is any or no evidence to suggest that linear correlation is present in the population. To do this we test the null hypothesis, $H_{0,}$ that there is no monotonic correlation in the population against the alternative hypothesis, $H_{1,}$ that there is monotonic correlation; our data will indicate which of these opposing hypotheses is most likely to be true. Let $\rho_s$ be the Spearman's population correlation coefficient then we can thus express this test as:

$$\begin{aligned} H_0 &: \rho_s = 0 \\ H_1 &: \rho_s \neq 0 \end{aligned}$$

i.e. the null hypothesis of no monotonic correlation present in population against the alternative that there is monotonic correlation present.

Since SPSS reports the p-value for this test as being .000 we can say that we have very strong evidence to believe $H_1$, i.e. we have some evidence to believe that groundwater uranium and TDS values are monotonically correlated in the population.

This could be formally reported as follows:

"A Spearman's correlation was run to determine the relationship between 23 groundwater uranium and TDS values. There was a strong, positive monotonic correlation between  Uranium and TDS ($r_s$ = .71, n = 23, p < .001)."