Spearman correlation trendline

E-mail distributed on 27-11-2025

Dear all,

For this month's newsletter I would like to discuss a caution and a recommendation on plotting Spearman correlation trendlines. In publications it is common to see scatterplots where Spearman correlation coefficients are printed alongside a linear trendline (e.g., Fig. 1, left panel). However, this is technically misleading, as the trendline corresponds to the Pearson correlation, which may conflict with the conclusion of the Spearman correlation. How does one plot an accurate Spearman trendline, in this case? Below I recap first some brief facts about Spearman correlation and then discuss the correct method for producing a trendline, including an R function (`spearline`) to automate its calculation.

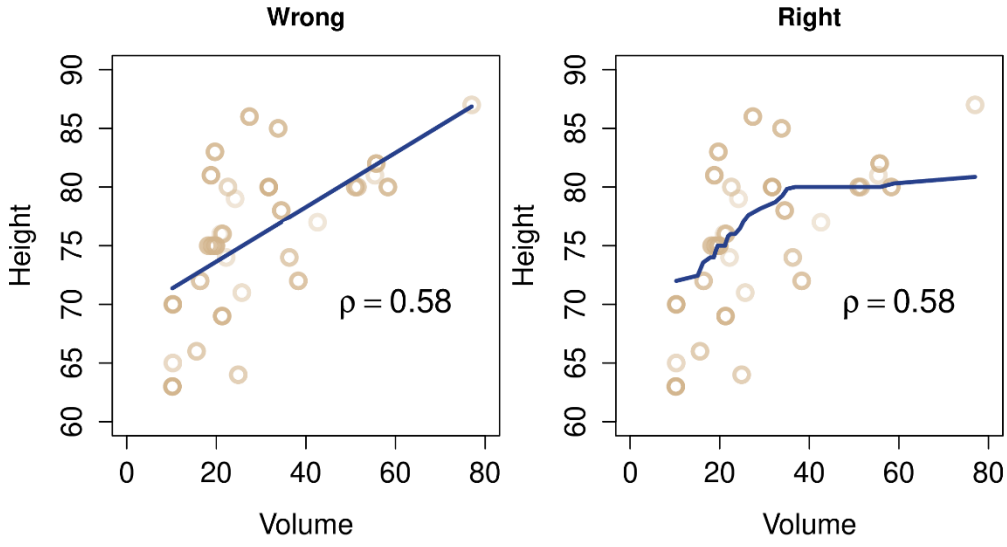

Figure 1. Scatterplot with Pearson trendline (left) and Spearman trendline (right).

1. Spearman correlation

The Spearman rank correlation coefficient, commonly denoted with Greek letter rho (ρ), is a popular "non-parametric" alternative to the standard Pearson correlation coefficient. Foremost it should be noted that the Spearman correlation is, in fact, a Pearson correlation, but simply applied to rank-transformed data. In turn, Pearson correlation is itself equivalent to a one-way linear regression where

both predictor and outcome have been standardized. Consequently, the following will produce identical output in R:

```
x <- rnorm(30)
y <- rnorm(30)
cor.test(x,y,method="spearman") #Direct Spearman method
cor.test(rank(x),rank(y),method="pearson") #Pearson method
summary(lm(scale(rank(y))~scale(rank(x)))) #regression method
```

This also serves to illustrate once more that almost all common models and tests in statistics can be rewritten as a special case of linear regression (Chartier & Faulkner, 2008).[1] Spearman correlation is often recommended when **(1)** variables are non-normally distributed, **(2a)** outliers or influential cases are present in the data, and **(2b)** the association between variables exhibits mild non-linearity. These three scenarios actually reflect two distinct types of parametric assumptions, with (1) concerning the shape of a *distribution*, and (2a–b) concerning the shape of an *association*. Either of these features can be subjected to non-parametric assumptions without implicating the other. For example, an association may be non-linear, while the distribution of its residuals is normal, and vice-versa.

A rank-transformation reduces a set of values purely to their order which, in the absence of ties, amounts to a set of ordered and equally spaced integers, e.g.:

```
> rank(c(0,0.5,1,2,3,1000))
[1] 1 2 3 4 5 6
```

This process forces the values to be uniformly distributed, which is why Spearman correlation actually falls more under the second type of non-parametric than the first. It is commonly assumed that Spearman correlation allows non-normal data but in truth its rank-transformation itself produces non-normal data. This means that an inferential test on a Spearman coefficient still depends on asymptotic normality of the test statistic (ρ) to be valid.

Forcing equal spacing between values is what makes Spearman correlation more robust against outliers and (some) non-linearity. In the above example, the gap between 3 and 1000 is reduced to 1 by transforming to 5 and 6, which vastly diminishes its impact on the resulting coefficient estimate. However, it is not a guaranteed failsafe against outliers, as rank-transformed data can still be outlying in the bivariate sense (e.g., a subject has the highest ranked score on both X and Y), and in the presence of many ties the distribution of the rank may be multimodal (e.g., zero-inflation).

Regarding non-linearity, Spearman correlation can detect **monotone** associations in addition to linear ones, that is, relationships which may have some curvature but are still one-directional, strictly increasing or decreasing, such as sigmoid or threshold relationships. It is not suitable for relationships where the direction of association changes. Moreover, in the rank-transformed value space, Spearman correlation is still linear (see next section).

[1] Social science textbooks and especially software such as SPSS have historically contributed to the confusion that tests like *t*-test, ANOVA, ANCOVA, correlation, etc., are distinct models when, in fact, they are all a special case of linear regression.

2. Spearman trendline

When reporting a scatterplot, the use of trendlines is recommended since eye-balling associations—even linear ones—can be surprisingly difficult. What is important is to be clear about what type of trend is being shown and for the visuals to be consistent with any information that is printed alongside as text. As noted, many publications print Spearman coefficient numbers on a scatterplot alongside the Pearson trendline, which may be misleading. The simplest solution in this case is to report a scatterplot in the rank-transformed space of the two variables. Here, the linear trendline will accurately reflect the Spearman correlation. Unfortunately, this will not have an intuitive interpretation for most readers, and it is generally advised to keep graphs on the original scale of the variables as much as possible. How can we link the rank values back to the original values then?

There is no general back-transformation for ranks, since by definition a rank-transformation erases the magnitude and direction of the original values. However, when the original values are available, one can approximately match the predicted rank of the Spearman correlation coefficient to the quantiles of the original values, e.g.:

```
x <- rnorm(30) ; rx <- rank(x)
y <- rnorm(30) ; ry <- rank(y)
rfit <- fitted(lm(ry~rx))
fit <- quantile(y,probs=rfit/length(y))
plot(x,y) ;   lines(x[order(x)],fit[order(x)])
```

Which will produce a Spearman trendline. At the bottom of this document, you can find an R script for the function `spearline` which will automate this for a pair of X and Y variables. The function will not literally plot the line but simply return line coordinates (ordered) in a list, so that it can be plugged into functions like `plot`, `lines`, or `points`. When missing values are present in X or Y, those cases will be removed and information is printed on how many cases were removed.

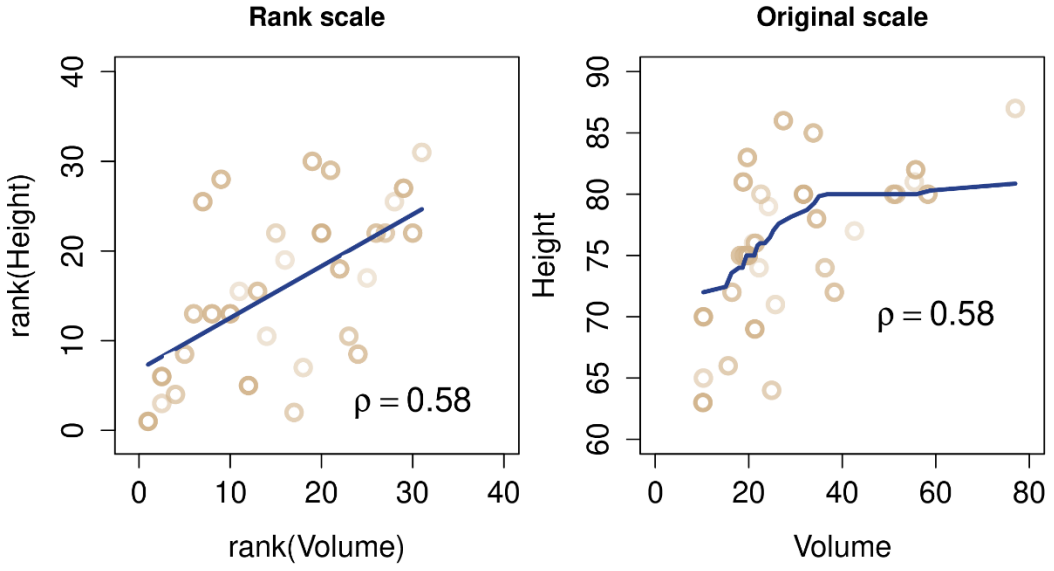

Figure 2. Scatterplot with Spearman trendline on the rank-transformed scale (left) and the original scale (right).

Fig. 2 shows an example for the relation between tree volume and tree height from the trees data set (default included in R), depicting the linear Spearman association on the rank-transformed scale of the data (left panel), and the nonlinear Spearman association on the original scale of the data (right panel). The nonlinear trendline very much resembles a non-parametric scatterplot smoother such as a LOESS curve and, in fact, I recommend LOESS curves as an alternative to a proper Spearman trendline.

A peculiarity of Spearman correlation is that it transforms both variables. For analyses where there is a clear distinction between predictor and outcome (e.g., multiple regression), typically only the outcome would be transformed.[2] For most practical purposes, this approach would probably also suffice to obtain a non-parametric estimate of a generic XY association (e.g., `cor(x,rank(y))`). However, transforming both X and Y maintains the symmetry in Spearman correlation that Pearson correlation also has. When only one is transformed, this no longer holds, that is:

$$Cor\big(X, rank(Y)\big) \neq Cor(rank(X), Y)$$

In addition, transforming both variables makes the resulting coefficient robust against outliers in both X and Y.[3]

Finally, it is reasonable to demand an estimate of precision on the Spearman trendline in the form of a confidence interval or an error band. While a similar back-transformation procedure could be attempted as outlined above for the lower and upper estimates of the predicted values, this may produce odd results and error bands that are visually incoherent. For this purpose, bootstrapping may be more useful although it should only be applied in large samples and I caution that bootstrapping rank statistics is its own complex topic. If one must plot a non-parametric trendline with an error band, it may be safer to plot a bootstrapped LOESS smoother instead.

My final advice regarding Spearman correlation and other non-parametric statistics is to remain consistent in your reporting. If you tested differences of medians, then medians should be plotted. If you tested rank associations, then rank-based trendlines should be reported. Avoid mixing parametric and non-parametric results in graphs or, if you do so, clarify what is being reported in your figure caption and/or highlight the comparative differences between the two.

References

Chartier, S., & Faulkner, A. (2008). General linear models: an integrated approach to statistics. *Tutorial in Quantitative Methods for Psychology*, *4*, 65–78.

Conover, W. J. (1999). *Practical Nonparametric Statistics* (Third Edition). John Wiley & Sons. New York.

Higgins, J. J. (2004). *An Introduction to Modern Nonparametric Statistics*. Pacific Grove, CA: Brooks/Cole.

Hollander, M., Wolfe, D. A., & Chicken, E. (2013). *Nonparametric Statistical Methods* (Third Edition). John Wiley & Sons. New York.

[2] In addition, linear regression only makes distributional assumptions about the outcome, *not* the predictors.

[3] Though not necessarily against bivariate outliers.

R script

```
################################################################################
## SPEARMAN TRENDLINE FUNCTION
################################################################################

spearline <- function(x,y) {

  ### MISSING DATA CHECK
  data <- data.frame(x,y)
  if(sum(complete.cases(data))!=nrow(data)) {
    cat(sum(!complete.cases(data)),"cases with missing values removed\n\n")
    flush.console()
    data <- data[complete.cases(data),]
  }

  ### MODEL
  rdata <- data.frame(apply(data,2,rank))
  model <- lm(y~x,data=rdata)
  newx <- seq(1,nrow(data),length.out=nrow(data))
  yout <- quantile(data$y,predict(model,newdata=data.frame(x=newx))/nrow(data))
  xout <- quantile(data$x,newx/nrow(data))

  ### OUTPUT
  invisible(list(x=xout[order(xout)],y=yout[order(xout)]))
}

################################################################################
```

--
Ben Meuleman, Ph.D.
Statistician
Swiss Center for Affective Sciences
University of Geneva | Campus Biotech
Chemin des Mines 9 | CH-1202 Genève
ben.meuleman@unige.ch | +41 (0)22 379 09 79