World Academy of Science, Engineering and Technology
International Journal of Humanities and Social Sciences
Vol:9, No:5, 2015

Choosing between the Regression Correlation, the Rank Correlation, and the Correlation Curve

Roger L Goodwin

Abstract: This paper presents a rank correlation curve. The traditional correlation coefficient is valid for both continuous variables and for integer variables using rank statistics. Since the correlation coefficient has already been established in rank statistics by Spearman, such a calculation can be extended to the correlation curve.

This paper presents two survey questions. The survey collected non-continuous variables. We will show weak to moderate correlation. Obviously, one question has a negative effect on the other. A review of the qualitative literature can answer which question and why. The rank correlation curve shows which collection of responses has a positive slope and which collection of responses has a negative slope. Such information is unavailable from the flat, "first-glance" correlation statistics.

Keywords: Bayesian estimation, regression model, rank statistics, correlation, correlation curve

I. Background

Reference [4] points out six factors that affect the size of a correlation. Those six factors include:

1) The amount of variability in either variable, X or F.
2) Differences in the shapes of the two distributions, X or F.
3) Lack of linearity in the relationship between X and F.
4) The presence of one or more "outliers" in the dataset.
5) Characteristics of the sample used for the calculation of the correlation.
6) Measurement error. Where possible, we illustrate the effects of these characteristics on the size of a correlation with a hypothetical data example.

The authors present a hypothetical dataset. The data is obviously made-up of integer responses. Given the graphs in the paper and the dataset, it is clear that the response variable is not normally distributed. Using the same hypothetical dataset, we can calculate the rank sum statistic as 0.83. Given that this is a hypothetical dataset, it is not possible to perform a qualitative analysis on the variables.

Reference [7] describes a way to model the covariance matrix and the coefficient matrix. Unfortunately, the coefficient matrix is limited to positive values only. Reference [1, pp. 401-402] give the estimators for a non-parametric approach to coefficient estimation. The author uses Pearson's original data from 1905 as examples. Those datasets appear to be real world data. This paper will present the correlation among the responses to two questions in an actual survey.

We will perform a qualitative analysis and quantitative analysis of the two survey questions. One question consistently, negatively correlates with the other questions in the survey. We examine its effects with a seemly, harmless question in the survey.

Roger L. Goodwin is with the US Government Publishing Office, Washington DC, 20401 (email: rgoodwin@gpo.gov).

II. The Two Survey Questions

The purpose of the survey was to measure customer satisfaction. The questionnaire contained 35 questions. Five-hundred, ninety-six people responded to two particular questions. This survey focused on government agencies that held contracts with the US Government Printing Office. The survey excluded entities under the Library program. The survey excluded bookstore patrons. Both the Library program and bookstore patrons will be queried under different surveys. The mode of delivery was via email. Potential respondents must have a DOT .mil or a DOT .gov email address to receive the questionnaire. Commercial (.com), organizations (.org) and education (.edu) institutions were out-of-scope. The previous survey performed in 2007 included the contractors that serviced the agencies. In 2011, only agencies were included in the survey.

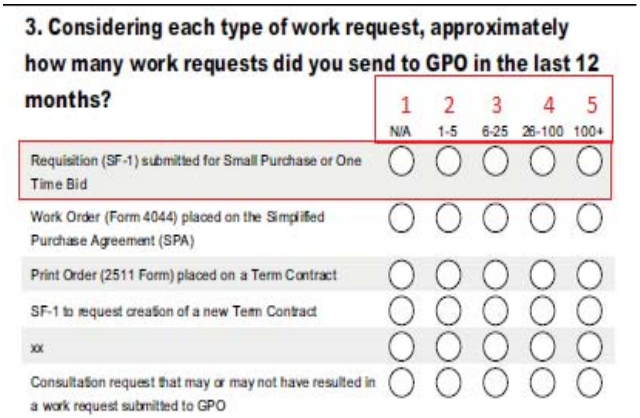

Fig. 1. This figure shows Question 3 in the 2011 Survey. Question 3 had six questions below it. Each of those questions required a response on the reader's part from N/A (i.e. not applicable) to $100+$.

The two questions have coded responses. First, we will show the two questions and the response coding. Then we will briefly discuss the wording of the questions. Table I shows the numerical codes for the responses to Question 3a. Table II shows the coded responses to Question 16.

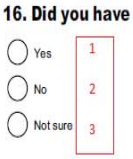

Fig. 2. This figure shows Question 16 in the 2011 Survey. Question 16 requires a response: Yes, No, and Not sure.

TABLE I. CODED RESPONSES TO QUESTION 3A

Choice	Coded Response
N/A	1
1-5	2
6-25	3
26-100	4
100+	5

A. Discussion

Reference [5] discusses the wording of questions and survey responses for personal interviews. The authors discuss the presentation of both factual questions and opinion questions for personal interview surveys. The authors reference [3], [8] for other modes of interviews such as self-completion questionnaires. We obviously ask for *factual* information from both Questions 3a and Question 16. Thus, we will concentrate on those issues only. The survey asks validity type questions at the agency level. Some problems with eliciting this information include:

1) Definition.
2) Comprehension.
3) Memory.
4) Social desirability.

From the factual question, what is the definition of the fact? Does the respondent understand the question and the appropriate answer? To give the correct answer, the respondent needs to have the necessary information accessible. If the question asks about the past, then the respondent must recall the information from memory (or records). The longer the period, the greater is the recall loss. Survey practice uses three procedures to minimize or avoid memory errors:

1) The use of records.
2) Aided recall techniques.
3) Diaries.

A source of invalidity in responses to factual questions is a social desirability bias where respondents distort their answers towards ones they consider more favorable to them. Asking for sensitive information falls into this category. Methods are available that desensitize a particular response by making it appear to be a common or acceptable one by. Another way to ask for sensitive information is the randomized response technique. The respondent chooses which of two (or more) questions he answers by a random device. In a personal interview, the respondent's identity is protected.

The objective of a survey is to have the respondent understand what is expected and have the respondent make the necessary effort to retrieve and organize the information into a suitable reporting form. To improve survey reporting (personal interviews),

TABLE II. CODED RESPONSES TO QUESTION 16

Choice	Coded Response
Yes	1
No	2
Not Sure	3

1) Include the use of respondent instructions.
2) The use of feedback.
3) The securing of respondent commitment.

A source of bias with factual questions includes the length of the list of items. The presentation order may affect the responses. From a lengthy list of items, respondents may select items from those items listed first. Respondents may select those items at the bottom of the list less often due to the list length.

B. Application to Question 3a

Let us look at Question 3a and go thru the discussion of [5]. This is a self-administered survey. Some of the same (or similar) issues still apply with factual questions. We do not attach the respondent's identity, such as the email address, to the survey responses.

Definition of terms:

Let us begin with the definition of the time "last 12 months." GPO delivered the survey instrument via email. The last 12 months can have the following definitions, depending on who received and interpreted the question.

1) The last 12 months can mean from the end date of receipt of the email.
2) The last 12 months can mean from the beginning the date the reader responds to the survey.
3) The Fiscal Year 2010, which would be October 1, 2009 to September 30, 2010.

In the absence of an instruction book, this phrase can mean different periods to different people. Stressing the word, "you" in italics would have helped convey some additional information. The person who received the questionnaire can respond for the agency or for himself.

Comprehension: Question 3a asks for the number of "small purchases and-one time bids," together. Small purchases and one-time bids are contract sizes. A Simplified Purchase Agreement (SPA) is a contract. Question 3b inquires about SPA contracts. A term contract is another type of contract. Questions 3c and 3d inquire about term contracts. There are specific differences among these contract types and sizes. Undoubtedly, reporting the number of contracts accurately will require referencing past records for most agencies.

Memory: Question 3a asks for the quantity of small purchases and one-times. This can become a survey coverage issue. In a small agency with only one person placing such contracts, the person can respond for himself. In a large agency with several regional offices, the answer is more complex. There could be complete coverage. GPO could have sent the survey instrument to those regional offices in which they would respond for themselves. On the other hand, there could be under-coverage. Those who received the survey instrument would have to account for the regional offices that did not

receive the survey instrument. Alternatively, there could be over-coverage. Several people in the same office received the survey instrument.

Since we are not asking for sensitive questions such as those oriented towards bankruptcy, drunken driving or abortion [5, p. 46], the issues of sensitivity and underreporting do not arise.

Next on the list is to provide the respondent with instructions. This is good advice towards definition of terms, comprehension, and respondent expectation. Detailed instructions were absent on this particular questionnaire. GPO provided feedback on the survey results via email, at a later date.

C. Application to Question 16

Definition of terms: Question 16 uses the phrase "past year." Question 3a uses the term "last 12 months." Consistent wording is an issue. Just as in Question 3a, we can find several different meanings for the term.

Comprehension: Overall, Question 16 seems simple and harmless.

Memory: Question 16 has the same over-coverage and under-coverage issues as Question 3a. Does the person who received the questionnaire respond for the agency or for himself? Stressing the word, "you" in italics would have helped convey some information. Defining the term, "you" in an instruction book would have certainly helped too.

III. Regression Correlation

Reference [2, p. 179] gives the simple correlation coefficient in (1) for the regression model $\widehat{y}_i = b_0 + b_1 x_i$, with (x_i, y_i) $i = 1, 2, ...n$ paired observations. The random variable Y is the dependent variable in the model. The variable X is the independent variable in the model.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{n}(x_i - \bar{x})\sum_{i=1}^{n}(y_i - \bar{y})\right]^{1/2}} \tag{1}$$

where the estimators for $\bar{x}$ and $\bar{y}$ are

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

and

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

We can re-write (1) to (2).

$$r = b_1 \frac{s_x}{s_y} \tag{2}$$

where the slope b_1 equals

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

and the intersection with the Y-axis equals

$$b_0 = \bar{y} - b_1 \bar{x}.$$

The standard deviation estimates s_x and s_y are

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

and

$$s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

The simple correlation coefficient assumes:

1) Linearity.
2) Normality of the dependent variable.
3) Paired observations.

Which sums become more difficult to calculate in (1) - (2)? The estimate b_1 looks complicated even though it is only summations. Other summations in the other estimators are easier to calculate. We prefer (1) because there are less terms. Using the survey responses where Question 3a is the independent variable X and Question 16 is dependent variable Y, we obtain the simple correlation coefficient estimate:

$$r = \frac{-109.0285235}{\sqrt{727.2936242 \times 237.9513423}} = -0.262084394$$

Some authors call the simple correlation coefficient the Pearson coefficient. The regression model assumes paired observations. However, the dependent variable (Question 16) is not normally distributed. A summary of some of the data appears in III - IV in Section IV.

IV. Rank Correlation

Tables III - IV summarize the raw data for a rank analysis.

TABLE III. The following table shows the rank statistics for the independent variable (Question 3a)

X	No. Responses	Avg Ranks
1	124	62.5
2	233	241.0
3	134	424.5
4	74	528.5
5	31	581.0

TABLE IV. The following table shows the rank statistics for the dependent variable (Question 16)

Y	No. Responses	Avg Ranks
1	318	159.5
2	233	435.0
3	45	574.0

From [6, Ch. 7], we can analyze the survey responses in the context of testing for randomness against an upward trend. Since the response variable is not normally distributed, the Pearson correlation coefficient is not appropriate — nor is linear regression. Since a large number of ties occur in the data, we apply mid-rank statistics. We rank the two sets of N responses separately. For tied observations, we take the average of the ranks. To test for independence, we can use the Spearman rank correlation coefficient. It tests for the strength of association between two characteristics.

Equation (3) gives the Spearman correlation coefficient for the two sets of ranked observations.

$$r = \frac{\sum_{i=1}^{N}(R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{N}(R_i - \bar{R})^2 \sum_{i=1}^{N}(S_i - \bar{S})^2}} \tag{3}$$

where

$$\bar{S} = \bar{R} = \frac{1}{2}(N+1) \tag{4}$$

Using the survey data, we obtain the following correlation estimate:

$$r = \frac{-4274639.75}{16192638 \times 13900903.5} = -0.284917551$$

This is another one-glance statistic. It may give a clue to the correlation between two variables if the assumptions hold true. The correlation curve gives more detailed, quantitative information. We already discussed the qualitative issues in Section II.

V. A Rank Correlation Curve

Reference [1] gives estimators for the non-parametric correlation curve. The entire correlation curve concept relies on continuous variable(s). Since this paper has neither variable, X or Y, as continuous, those estimators need to be adapted for rank statistics. Equation (5) shows the estimator for the conditional ranked mean of the response variable.

$$\widehat{\mu}_i = \bar{S}_i = \frac{\sum_{j=1}^{n_i} S_{ij}}{n_i}. \tag{5}$$

Specifically to the survey outlined in Section II, $i = 1, 2, ...5$; and the total number of paired responses is $N = 596$.

Equation (6) gives a measure of the variance of the mid-ranks.

$$\widehat{\sigma}_i^2 = \sum_{j=1}^{n_i}(R_{ij} - \bar{R}_i)^2 \sum_{j=1}^{n_i}(S_{ij} - \bar{S}_i)^2 \tag{6}$$

Note that for the independent variable S, the terms $\sum_{j=1}^{n_i}(S_{ij} - \bar{S}_i)^2$ are the same for each i. From a computational aspect, it is simpler to perform the conditional summations for both R and S.

$$\widehat{\sigma}_z\widehat{\beta}_i = \sum_{j=1}^{n_i}(R_{ij} - \bar{R}_i)(S_{ij} - \bar{S}_i) \tag{7}$$

Equations (6) - (7) give the estimators for the rank correlation curve in (8).

$$\widehat{\xi}_i = \frac{\widehat{\sigma}_z\widehat{\beta}_i}{\widehat{\sigma}_i} \tag{8}$$

where $\widehat{\xi}_i$ is the rank correlation curve for given values of the independent variable X_{ij}, $j = 1, 2, ..., n_i$, $i = 1, 2, ..., 5$.

Fig. 3 shows the correlation curve using rank statistics for the two survey questions discussed Section II. The two straight lines represent the Spearman correlation coefficient and the Pearson correlation coefficient. Table V summarizes the results of the correlation curve. Notice that the correlations are negative for each conditioned value. The correlations tend to be moderately valued. The curve has a positive slope between the coded values of 1 and 2; and a negative slope thereafter.

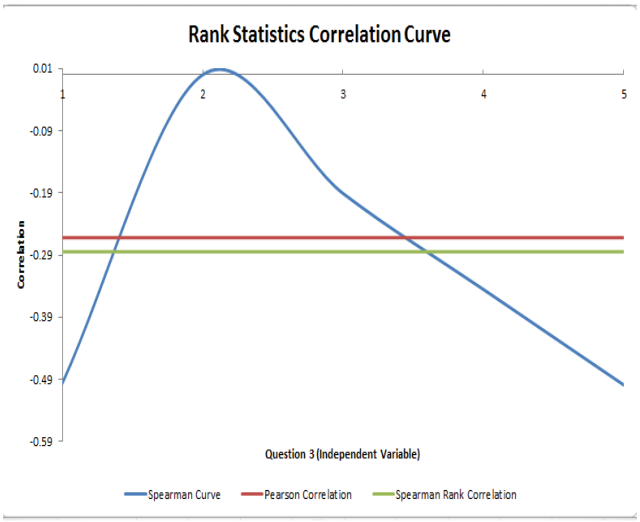

Fig. 3. This figure shows the rank correlation curve. The curve has a positive slope between the values 1 and 2; and a negative slope thereafter.

TABLE V. This table shows the correlation coefficients using rank statistics

Coded Response	Response	n_i	r
1	N/A	124	-0.496066354
2	$1-5$	233	-0.000223162
3	$6-25$	134	-0.191054933
4	$26-100$	74	-0.345224995
5	$100+$	31	-0.498809561

VI. Summary

This paper presented the correlation between two questions in a survey. One question consistently negatively correlates with the other questions in the survey. We reviewed the literature of the qualitative analysis survey questions and applied it to two survey questions. We reviewed the literature of the quantitative analysis of correlation. We presented a quantitative analysis of the data using first-glance statistics. We discussed, in detail, the computational aspects and assumptions to the correlation statistics.

REFERENCES

[1] S. Blyth, "Karl Pearson and the Correlation Curve," *International Statistical Review / Revue Internationale de Statistique,* Vol. 62, No. 3 (Dec., 1994), pp. 393-403.

[2] B. L. Bowerman and R. T. O'Connell, *Linear Statistical Models: An Applied Approach, Second Edition,* Duxbury Press, Belmont, CA, 1990.

[3] J. B. Forsythe and O. Wilhite, "Testing Alternative Versions of Agriculture Census Questionnaires," *Proceedings of the Business and Economic Statistics Section, American Statistical Society,* 1972, pp 206-215.

[4] L. D. Goodwin and N. L. Leech, "Understanding Correlation: Factors That Affect the Size of *r*," *The Journal of Experimental Education,* Vol. 74, No. 3 (Spring, 2006), pp. 251-266.

[5] G. Kalton and H. Schuman, "The Effect of the Question on Survey Responses: A Review," *Journal of the Royal Statistical Society,* Series A (General), Vol. 145, No. 1 (1982), pp.42-73.

[6] E. L. Lehmann and H. J. M. D'Abrera, *Nonparametrics: Statistical Methods Based on Ranks,* Holden-Day, Inc., Oakland, CA, 1975.

[7] J. C. Liechty, M. W. Liechty and P. Mller, "Bayesian Correlation Estimation," *Biometrika,* Vol. 91, No. 1 (Mar., 2004), pp. 1-14.

[8] N. D. Rothwell and A. M. Rustmeyer, "Studies of Census Mail Questionnaires," *Journal of Marketing Research,* 16, 1979, 401-406.

Roger L. Goodwin Roger Goodwin has 15 years' experience with several government agencies. Two of the agencies are statistical in nature; the third agency is both production and commercial in nature. Roger Goodwin completed statistical assignments on many computer platforms, which include PCs, Vax VMS OS, Unix OS, and IBM mainframes. He usually performs his statistical analyses in SAS and uses Excel for simpler calculations. He developed reports for cost, progress, and billing reports using SAS and SAP Business Objects.

Roger Goodwin holds a BS in Computer Science and an MS in Applied Statistics from Old Dominion University. He completed a certificate in Software Engineering Processes from Learning Tree. He completed the Project Management Professional certification from PMI. He authored several papers in IEEE conferences and two online journals that summarize his experiences in government. He authored papers in the North East SAS Users Group that describes some of the SAS code that he wrote.