A theoretical expectancy table calculator

BRETT MYORS

University of New South Wales, Kensington
New South Wales, Australia

A program for calculating theoretical expectancy tables is described. These tables are useful for communicating the meaning of correlation coefficients, since they display the relationship in a simple way. The program allows users to go beyond empirical expectancy tables, which are routinely calculated, by allowing the table to reflect corrections for effects such as unreliability and restriction of range.

The value of expectancy tables in conveying the meaning of correlation coefficients has been known for many years. They are easy to construct from raw data and readily facilitate communication with lay people. Preparation of an expectancy table from raw data involves three simple steps:

1. Group the x and y distributions into a few meaningful categories. This step forms the basic structure of the table. It is common practice to place the x variable along the rows of the table and the y variable along the columns.
2. Determine the joint frequency distribution; that is, tally up the number of data points falling into each cell in the table.
3. Divide each cell by its appropriate row total. The resulting proportion may be expressed as a percentage.

Modern computer programs, especially spreadsheets, make this basic procedure easy. Step 3 converts the raw tallies into conditional probabilities, which indicate the probability of obtaining a certain y value given the particular x value. Such tables are called *empirical expectancy tables* (Lawshe & Bolda, 1958), because they are based on tallies of the raw data.

Of equal, if not greater, interest are the *theoretical expectancy tables* (Lawshe, Bolda, Brune, & Auclair, 1958), which are based not on raw tallies, but on the overall correlation coefficient. The advantage of theoretical tables over empirical ones is that the value in each cell is influenced by all data points rather than by the small subset that happens to fall into a particular cell; thus, the individual cells do not exhibit local distortions due to sampling error. The probabilities change smoothly throughout the table and do not display any “bumps” or “reversals” (Lawshe et al., 1958). On the other hand, since theoretical tables are computed from correlation coefficients, they do not reveal nonlinear trends, which may be obvious in empirical tables. Modern computer packages render construction of empirical tables trivial,

but this is not the case for theoretical tables, which are based on the bivariate correlational surface.

To facilitate the construction of theoretical expectancy tables, Lawshe et al. (1958) published a series of five tables for the upper 30%, 40%, 50%, 60%, and 70% portions of the y distribution. In each table, the x distribution was split into five equal portions of 20%, and the correlations ranged between .15 and .85 in increments of .05. Only the upper portion of the y distribution was shown, because Lawshe et al. preferred expectancy tables containing one column pertaining to the “superior” level. The program described here allows the user to specify any desired partitioning of the x and y distributions and any correlation coefficient.

Method

A program called “Expect” was written for IBM-PC-compatible microcomputers using the Microsoft QuickC Compiler (Microsoft, 1990). First, the program prompts for the correlation coefficient. It then asks for the number of rows in the table, which may range between 2 and 15. It then prompts for the proportion of the x distribution in each row, beginning with the first row. It is preferable to start with the low end of the distribution so that x scores increase from top to bottom in the table. Proportions for the y distribution are input in a similar fashion, so that y scores increase from left to right in the table.

Once all the information is entered, three separate tables and one disk file are produced. The first table gives the raw bivariate probabilities under the correlational surface, which provides the probabilities for each cell. In an empirical table, this corresponds to dividing each cell tally by the total N of the table, rather than by the row total in Step 3. Lawshe (1948) called this a “simple” expectancy table. The second table printed is the theoretical version corresponding to Steps 1–3 and is called a “successive” expectancy table (Lawshe, 1948). Finally, a third table is printed. This table is called a “cumulative” expectancy table (Lawshe, 1948). At this point, Expect also prints a file called “expect.out” to disk. This file contains the coordinates of 500 points that, when plotted as a scatterplot in a spreadsheet or via one of the many charting programs available, graphically display the desired correlation. Thus, users have the option of presenting their correlations in either tabular or graphical form with enough points to make the relationships clear. The reader is reminded that the points in expect.out are not empirical data, but are merely theoretical (simulated) data. Once all printing is completed, users are asked if they want to print another table. An example session is shown in Table 1.

The example shown in the table corresponds to an entry in Lawshe et al.’s (1958) Table 4. Other input could have been used to produce results outside the published

Address correspondence to B. Myors, School of Psychology, University of New South Wales, P.O. Box 1, Kensington, NSW 2033, Australia.

Table 1
Example Input and Output

Correlation coefficient:	.45	
How many rows in tables?	5	
Proportion in row 1:	.2	
Proportion in row 2:	.2	
Proportion in row 3:	.2	
Proportion in row 4:	.2	
How many columns in tables?	2	
Proportion in column 1:	.4	

Simple expectancy table for $r = .45$:

	40.0%	60.0%
20.0%	13.2%	6.8%
20.0%	9.9%	10.1%
20.0%	7.8%	12.2%
20.0%	5.8%	14.2%
20.0%	3.4%	16.6%

Successive expectancy table for $r = .45$:

	40.0%	60.0%
20.0%	65.9%	34.1%
20.0%	49.3%	50.7%
20.0%	38.9%	61.1%
20.0%	29.0%	71.0%
20.0%	16.9%	83.1%

Cumulative expectancy table for $r = .45$:

	40.0%	60.0%
100.0%	40.0%	60.0%
80.0%	33.5%	66.5%
60.0%	28.3%	71.7%
40.0%	23.0%	77.0%
20.0%	16.9%	83.1%

Note—Italics indicate values entered by the user.

tables. If a large table containing many rows and columns is desired, it is recommended that input for Expect be created in a text editor and sent to Expect via I/O redirection—for example, `expect <input.dat`. DOS manuals explain how I/O redirection works. This will allow the user to correct any errors in the text editor without having to reenter all the values.

Expect can also be used to create theoretical tables corrected for unreliability and range restriction, something which is impossible to do in an empirical table. Since a correlation coefficient is used by the program, the way to produce an expectancy table corrected for these effects is simply to enter the corrected correlation. In the example so far, the raw correlation was .45, and say that the criterion y variable had a reliability of .8. To produce an expectancy table corrected for this level of unreliability in the criterion, enter $.503 = .45/\sqrt{.8}$ into the program instead of .45. Corrections for range restriction may also be applied by using the familiar formulas (Thorndike, 1949). Please note that the program does not make these corrections. Users are expected to calculate them prior to running the program.

Successive and Cumulative Expectancy Tables

Lawshe and Bolda (1958) distinguished between successive and cumulative tables. These are also known as *individual* and *institutional* tables, respectively. Successive

tables probably have a wider range of application and conform to the description of Steps 1–3. That is, each row shows the conditional probability of lying in a particular column of the y distribution given your location in that row of the x distribution. For example, the value of 71.0% in row 4, column 3 of the successive table in the example is obtained from row 4, column 3 in the simple table by $71.0\% = 14.2\%/20.0\%$. This may have meaning in a selection context; for example, it indicates that someone who obtained a score of 4 out of 5 on a test with validity .45 has a 71% chance of being successful.

Each cell in the cumulative table shows the *cumulative* conditional probability of lying in a particular column of the table given your location in that row. For example, 77.0% in the cumulative table can be obtained from the simple table as $(16.6\% + 14.2\%)/(20.0\% + 20.0\%)$. Cumulative tables are particularly applicable in the context of selection validity, where they show the percentage of people falling into a particular region of performance given the proportion of people selected on the basis of a test of particular validity. In the example, if 60% of the applicants are considered successful, then the consequence of accepting the top 40% on the basis of a selection device with a validity of .45 would be that 77% of those selected would be considered successful—an improvement of 17% over random selection. If the cut score was lowered so that the top 60% were accepted, then only 71.7% would be successful, and so on.

Summary

Empirical expectancy tables are easy to develop, and they are useful for communicating the meaning of correlation coefficients to people who are unfamiliar with statistics. For this reason, they have been in widespread use for many years. Theoretical tables are not so easy to construct without knowledge of the bivariate correlational surface, but they have the advantage of presenting an idealized form of the empirical table that is free from local distortions due to sampling error. They also permit the tabulation of relationships that have been corrected for such things as unreliability and restriction of range—something that is impossible to do in an empirical table.

In the past, practitioners have had to rely on various look-up tables in order to obtain theoretical expectancies. This has the disadvantage of making available only a relatively few points under the correlational surface. If your data did not exactly match one of the tabulated entries, then you had to make do with an approximation. To solve this problem, a computer program that calculates the desired expectancies for any location on the surface was developed. This augments the use of computer-generated empirical tables and increases the accuracy of theoretical expectancies as currently constructed. It also increases the flexibility of how the surface may be partitioned.

Availability

Copies of the program are available from the author via e-mail (b.hesketh@unsw.edu.au).

REFERENCES

- LAWSHE, C. H. (1948). *Principles of personnel testing*. New York: McGraw-Hill.
- LAWSHE, C. H., & BOLDA, R. A. (1958). Expectancy charts: I. Their use and empirical development. *Personnel Psychology*, **11**, 353-365.
- LAWSHE, C. H., BOLDA, R. A., BRUNE, R. L., & AUCLAIR, G. (1958). Expectancy charts: II. Their theoretical development. *Personnel Psychology*, **11**, 545-559.
- MICROSOFT CORPORATION. (1990). *QuickC compiler*. Washington, DC: Author.
- THORNDIKE, R. L. (1949). *Personnel selection*. New York: Wiley.

(Manuscript received June 14, 1993;
revision accepted for publication March 22, 1994.)