The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: http://www.tandfonline.com/loi/utas20

A Graphic Representation of a Three-Way Contingency Table: Simpson's Paradox and Correlation

Minja Paik

To cite this article: Minja Paik (1985) A Graphic Representation of a Three-Way Contingency Table: Simpson's Paradox and Correlation, The American Statistician, 39:1, 53-54

To link to this article: http://dx.doi.org/10.1080/00031305.1985.10479387

A Graphic Representation of a Three-Way Contingency Table: Simpson's Paradox and Correlation

MINJA PAIK*

A simple graphic representation is given to illustrate the relationship between two variables in the presence of a possibly confounding variable. The graph makes it easy to understand Simpson's paradox by using the concept of correlation.

KEY WORDS: Correlation, within group vs. overall; Simpson's paradox.

Consider as an example the problem of sex bias in the hiring policy of a large company with two hypothetical departments, H and L. Suppose a $2 \times 2 \times 2$ table relating hiring status to the sex of the applicant by department is constructed, as in Table 1. The letter in each cell denotes the frequency for the cell. (See Bickel et al. 1975 for real data of this type.)

Let $\Pr(mL)$, $\Pr(fL)$, $\Pr(mH)$, and $\Pr(fH)$ be, respectively, the probabilities of hiring within the male-L, female-L, male-H, and female-H subgroups. For the above data, $\Pr(mL) = .275$, $\Pr(fL) = .313$, $\Pr(mH) = .738$, and $\Pr(fH) = .800$. Furthermore, let $\Pr(m)$ and $\Pr(f)$ be the overall hiring rates for the male and the female groups, respectively. Notice that $\Pr(mL) < \Pr(fL)$, and $\Pr(mH) < \Pr(fH)$, but $\Pr(m) = .583 > \Pr(f) = .410$. This apparent contradiction is an example of *Simpson's paradox*. Exactly how the paradox occurs and what happens in variations of this problem can be etiologically explained to students by means of a graph and the concept of correlation.

Draw a circle graph, as in Figure 1. Here, let each circle represent a sex $\times$ department subgroup. $X = 0$ for males and 1 for females, and the y coordinate of the center of each circle stands for the subgroup-specific hiring rate. The area of each circle is proportional to the sample size of the subgroup it represents.

Consider, first, the line segments connecting the two H circles and the two L circles. Their slopes β_H and β_L are both positive. But the slope β_p (p indicating pooled) of the middle line that connects $\Pr(m)$ and $\Pr(f)$ is negative. This is the slope representation of Simpson's paradox.

Table 1. A $2 \times 2 \times 2$ Table

	Department L		Department H	
Status	*Male*	*Female*	*Male*	*Female*
Hired	550 = a	1,250 = b	2,950 = A	800 = B
Denied	1,450 = c	2,750 = d	1,050 = C	200 = D
Total	2,000	4,000	4,000	1,000

*Minja Paik is Assistant Professor of Mathematical Sciences, George Mason University, Fairfax, VA 22030.

The paradox is more clearly visualized by the circle graph when we use the concept of correlation. In a 2×2 table, the phi coefficient, defined as $(\chi^2/N)^{1/2}$, is precisely the ordinary correlation coefficient r applied to two dichotomous variables (see Kendall and Stuart 1979). Let r_H, r_L, and r_p be the correlation coefficients in the 2×2 tables for departments H, L, and pooled, respectively. The slope of the line connecting the circle centers and the correlation in each department are related and share the same sign as shown, for example, by

$$\beta_H = [B/(B + D)] - [A/(A + C)]$$
$$= (BC - AD)/(A + C)(B + D)$$

and

$$r_H = (BC - AD)/\sqrt{(A + C)(B + D)(A + B)(C + D)}.$$

This fact and the circle graph help one to determine (if one has some training in looking at a bivariate distribution) the sign of r_p (and β_p) without the step of pooling the tables and actually computing. Consider all four circles in Figure 1 *together*. We see a negative overall correlation, since the two larger circles on the negative diagonal dominate the positive ones. This *simultaneous* consideration corresponds to the pooling of the tables as shown by the following claim. Define a bivariate variable (X, Y), taking the values at $[0, \Pr(mH)]$, $[0, \Pr(mL)]$, $[1, \Pr(fH)]$, and $[1, \Pr(fL)]$ with probabilities proportional to the sizes of the corresponding circles. It can then be easily shown that r_c, the correlation coefficient of X and Y, is equal to $\beta_p \cdot sd(X)/sd(Y)$, where sd is the standard deviation. Thus r_c, β_p, and r_p share the same sign. A visual inspection of the graph is often sufficient to determine the sign of these coefficients.

In summary, the circle graph enables us to discern all of the relationships existing within the separate tables, as well

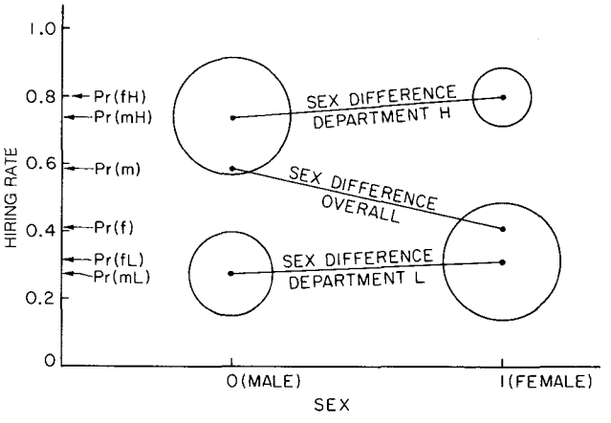

Figure 1. A 2×2 Circle Graph

as the summary table, at one glance. The top two circles and the bottom two circles show within-group correlations to be positive, and a negative correlation based on the four-circle distribution implies a negative overall correlation.

By varying the positions and sizes of the circles in Figure 1, one can easily see that all of the 3^3 combinations for the three correlations are actual possibilities. The sex bias problem represents only one of these combinations. In an example discussed by Mantel (1982), the overall correlation is zero, but creation of separate tables for the subgroups creates irrelevant, positive, within-group correlations. A third possibility that has zero within-group correlations but positive overall correlation assumes the probability of a male child at successive births to be independent and fixed but different among different families. In this case, the sex of successive children will be positively correlated overall, and the within-family correlation will remain zero. Phenomena of this type contribute to the confusion associated with the term *correlation*.

[*Received July 1983. Revised May 1984.*]

REFERENCES

Bickel, P. J., Hammel, E. A., and O'Connell, J. W. (1975), "Sex Bias in Graduate Admissions: Data From Berkeley," *Science*, 187, 398–404.

Kendall, S. M., and Stuart, Alan (1979), *The Advanced Theory of Statistics* (Vol. 2; 4th ed.), New York: Macmillan.

Mantel, Nathan (1982), "Simpson's Paradox in Reverse," *The American Statistician*, 36, 395.

Jensen's Inequality for General Location Parameter

C.H. SPIEGELMAN*

A wide class of location parameters is shown to satisfy Jensen's inequality. When the expectation EX exists and l is a convex function, Jensen's inequality states that $El(x) \geq l(EX)$. It is shown that for μ, a properly defined location parameter, $\mu(l(x)) \geq l(\mu(x))$.

KEY WORDS: Concave; Convex; Majorization.

For any convex function l, the first part of Jensen's inequality (Lehmann 1983, p. 50) states that $El(x) \geq l(EX)$ whenever the expectation EX exists. Though simple in concept, this inequality has had a phenomenal amount of use and is virtually the cornerstone of many inequalities. Because of this, many different types of extensions have been considered. In the present version, the expectation operator is replaced by another one for which the mean is a special case. This permits the inequality to be applied to a wide class of location parameters.

Let μ be the location parameter associated with the distribution function F; it is convenient to write $\mu(X)$ for $\mu(F)$, where X has distribution function F. It is shown that for μ, a location parameter appropriately defined, $\mu[l(x)] \geq l[\mu(x)]$.

The definition used is that of Bickel and Lehmann (1975, p. 1046), namely:

1. μ takes on larger values for random variables that are typically larger. Formally, they require $\mu(X) \leq \mu(Y)$ whenever Y is stochastically larger than X [i.e., $F_x(t) \geq F_y(t)$ for $-\infty < t < \infty$].
2. $\mu(aX + b) = a\mu(X) + b$ if $a > 0$.
3. $\mu(-X) = -\mu(X)$.

Doksum (1975, p. 11) gave an equivalent definition in approach 2.

Suppose that l is defined on a convex set containing the support of F.

Theorem. If $\mu(X)$ exists, then $\mu[l(X)] \geq l[\mu(X)]$.

Proof. Let $L(x) = a + bx$ be a line tangent to $l(x)$ at an arbitrary point $[x_0, l(x_0)]$. Then by 2 and 3, it follows that $\mu[L(X)] = L[\mu(X)]$. Since l is convex, $l(x) \geq a + bx$; and by 1, $\mu[l(X)] \geq L[\mu(X)]$. It only remains to choose $x_0 = \mu(X)$, which is possible by Bickel and Lehmann's (1975) Theorem 1.

Comment. The second part of Jensen's inequality, which states that "If l is strictly convex, the inequality is strict unless X is constant with probability 1," does not generalize to our family of location parameters. This can be easily seen in the case of the median.

There are many examples of location parameters that satisfy these assumptions. A few of the more popular ones are trimmed means, medians (for this special case there is a previous inequality in Tomkins 1975), and midrange.

[*Received February 1984. Revised February 1984.*]

REFERENCES

Bickel, P. J., and Lehmann, E. L. (1975), "Descriptive Statements for Nonparametric Models II: Location," *Annals of Statistics*, 3, 1045–1069.

Doksum, K. A. (1975), "Measures of Location and Asymmetry," *Scandinavian Journal of Statistics*, 2, 11–22.

Lehmann, E. L. (1983), *Theory of Point Estimation*, New York: John Wiley.

Tomkins, R. J. (1975), "On Conditional Medians," *Annals of Probability*, 3, 375–379.

*C.H. Spiegelman is Mathematical Statistician at the Center for Applied Mathematics, Statistical Engineering Division, National Bureau of Standards, Gaithersburg, MD 20899. The author thanks I. Olkin for comments and suggestions.