Classic Problems of Probability

Prakash Gorroochurn

WILEY

excess of heads over tails between 45 and 55% of the time is only about $(2/\pi)(\arcsin\sqrt{.55}-\arcsin\sqrt{.45}) \approx .06$!

There is at least one other arcsine law,[†] again due to Lévy (1965). If $Z \in [0,\ 1]$ is the time of first occurrence of the maximum of the random walk, then Z has an arcsine distribution function too.

[†] See, for example, Mörters and Peres (2010, pp. 136–137).

Problem 30

Simpson's Paradox (1951)

Problem. *Consider the three events X, Y, and Z. Suppose*

$$\Pr\{X|YZ\} > \Pr\{X|\overline{Y}Z\} \quad \text{and} \quad \Pr\{X|Y\overline{Z}\} > \Pr\{X|\overline{Y}\overline{Z}\}.$$

Prove that it is still possible to have

$$\Pr\{X|Y\} < \Pr\{X|\overline{Y}\}.$$

Solution. Using the law of total probability, we have

$$\begin{aligned}\Pr\{X|Y\} &= \Pr\{X|YZ\}\Pr\{Z|Y\} + \Pr\{X|Y\overline{Z}\}\Pr\{\overline{Z}|Y\} \\ &= s\Pr\{X|YZ\} + (1-s)\Pr\{X|Y\overline{Z}\},\end{aligned} \tag{30.1}$$

$$\begin{aligned}\Pr\{X|\overline{Y}\} &= \Pr\{X|\overline{Y}Z\}\Pr\{Z|\overline{Y}\} + \Pr\{X|\overline{Y}\overline{Z}\}\Pr\{\overline{Z}|\overline{Y}\} \\ &= t\Pr\{X|\overline{Y}Z\} + (1-t)\Pr\{X|\overline{Y}\overline{Z}\},\end{aligned} \tag{30.2}$$

where $s = \Pr\{Z|Y\}$ and $t = \Pr\{Z|\overline{Y}\}$. Therefore,

$$\begin{aligned}\Pr\{X|Y\}-\Pr\{X|\overline{Y}\} &= [s\Pr\{X|YZ\}-t\Pr\{X|\overline{Y}Z\}] \\ &\quad + [(1-s)\Pr\{X|Y\overline{Z}\}-(1-t)\Pr\{X|\overline{Y}\overline{Z}\}]\end{aligned} \tag{30.3}$$

We now consider the sign of Eq. (30.3). Let $t = s+\delta(-1 \le \delta \le 1)$, $u \equiv \Pr\{X|YZ\}-\Pr\{X|\overline{Y}Z\} \ge 0$, and $v \equiv \Pr\{X|Y\overline{Z}\}-\Pr\{X|\overline{Y}\overline{Z}\} \ge 0$. Then

$$\begin{aligned}\Pr\{X|Y\}-\Pr\{X|\overline{Y}\} &= \left[s\Pr\{X|YZ\}-s\Pr\{X|\overline{Y}Z\}-\delta\Pr\{X|\overline{Y}Z\}\right] \\ &\quad + \left[(1-s)\Pr\{X|Y\overline{Z}\}-(1-s)\Pr\{X|\overline{Y}\overline{Z}\}\right] + \delta\Pr\{X|\overline{Y}\overline{Z}\} \\ &= su + (1-s)v - \delta w,\end{aligned}$$

where $w \equiv \Pr\{X|\overline{Y}Z\}-\Pr\{X|\overline{Y}\overline{Z}\}$. Therefore, $\Pr\{X|Y\}-\Pr\{X|\overline{Y}\}$ is negative if $su+(1-s)v<\delta w$.

Classic Problems of Probability, Prakash Gorroochurn.

30.1 Discussion

The algebra can mask the real implication of the three inequalities given in **Problem 30**. Consider the following example.[†] In a given University 1, 200 of 1000 males, and 150 of 1000 females, study economics. In another University 2, 30 of 100 males, and 1000 of 4000 females, study economics. Thus, in each university, more males than females study economics (1: 20% vs. 15%, 2: 30% vs. 25%). However, when the universities are combined, 230 of 1100 males (20.9%), and 1150 of 5000 females (23.0%), study economics. It now appears that, overall, more females than males study economics! If we define the event X that a student studies economics, the event Y that a student is male, and the event Z that a student goes to University 1, we obtain the counterintuitive set of inequalities in **Problem 30**. This is the essence of *Simpson's Paradox*[‡]: a reversal of the direction of association between two variables (gender and study economics) when a third variable[§] (university) is controlled for.

Intuitively, why does *Simpson's Paradox* occur? First note that University 2 has a higher study rate in economics for both males and females, compared to University 1 (see Table 30.1). However, out of the total 1100 males, only about 9% (i.e., 100) go to University 2. On the other hand, out of the total 5000 females, 80% (i.e., 4000) go to University 2. Thus, the university that has the higher study rate in economics, namely University 2, takes many more females than males, relatively speaking. No wonder when looking at the combined data we get the impression that a higher proportion of females than males study economics!

Simpson's Paradox shows the importance of carefully identifying the third variable(s) before an analysis involving two variables on aggregated data is carried out. In our example, when university is not controlled for, we get the wrong impression that more females than males study economics. Furthermore, from Eqs. (30.1) and (30.2), we observe that if $s = t$ then $\Pr\{X|Y\} > \Pr\{X|\overline{Y}\}$, that is, if gender is independent of university (i.e., there is no gender differences across universities), Simpson's Paradox is avoided. The tables for each university are then said to be collapsible.

An interesting alternative demonstration of *Simpson's Paradox* can be obtained through a graphical approach.[**] Consider the data in Table 30.2.

[†] Taken from Haigh (2002, p. 40), courtesy of Springer.

[‡] First named by Blyth (1972). Simpson's paradox is also discussed in Dong (1998), Rumsey (2009, p. 236), Rabinowitz (2004, p. 57), Albert et al. (2005, p. 175), Lindley (2006, p. 199), Kvam and Vidakovic (2007, p. 172), Chernick and Friis (2003, p. 239), Agresti (2007, p. 51), Christensen (1997, p. 70), and Pearl (2000, p. 174).

[§] Also sometimes called the lurking variable.

[**] See Kocik (2001), and Alsina and Nelsen (2009, pp. 33–34).

Table 30.1 Illustration of Simpson's Paradox Using University Data

	University 1				University 2				Pooled			
	E	$\bar{E}$	Total	Study rate (%)	E	$\bar{E}$	Total	Study rate (%)	E	$\bar{E}$	Total	Study rate (%)
Female	150	850	1000	15	1000	3000	4000	25	1150	3850	5000	23.0
Male	200	800	1000	20	30	70	100	30	230	370	1100	20.9

We wish to show that, given

$$\frac{a}{b} < \frac{A}{B} \quad \text{and} \quad \frac{c}{d} < \frac{C}{D}, \tag{30.4a}$$

it is still possible to have

$$\frac{a+c}{b+d} > \frac{A+C}{B+D}. \tag{30.4b}$$

We represent the proportions of the different students who study economics by vectors on a Cartesian plane such that the proportions are equal to the slopes of the corresponding lines (see Fig. 30.1). For example, the proportion of females who study economics in University 1 is a/b; we represent this proportion by the vector joining $(0, 0)$ and (b, a). Since $a/b < A/B$, the segment joining $(0, 0)$ and (b, a) has a smaller slope than the segment joining $(0, 0)$ and (B, A). Similarly for $c/d < C/D$. By addition of vectors, we see that it is possible for the slope joining $(0, 0)$ and $(b + d, a + c)$ to be larger than the slope joining $(0, 0)$ and $(B + D, A + C)$, that is, it is possible to have $(a+c)/(b+d) > (A+C)/(B+D)$.

A natural question remains: how should we combine the data from the two universities in order to obtain "correct" economics study rates for females and males? Clearly, just adding the numbers in each university is not appropriate since it gives the two proportions $(a+c)/(b+d)$ and $(A+C)/(B+D)$ for females and males, respectively (see Table 30.2). Following Tamhane and Dunlop (2000, p. 132), we calculate the adjusted proportion of females who study economics across the two

Table 30.2 General Distribution of Frequencies Across Universities and Gender

	University 1			University 2			Pooled		
	E	$\bar{E}$	Total	E	$\bar{E}$	Total	E	$\bar{E}$	Total
Female	a	$b-a$	b	c	$d-c$	d	$a+c$	$(b+d)-(a+c)$	$b+d$
Male	A	$B-A$	B	C	$D-C$	D	$A+C$	$(B+D)-(A+C)$	$B+D$

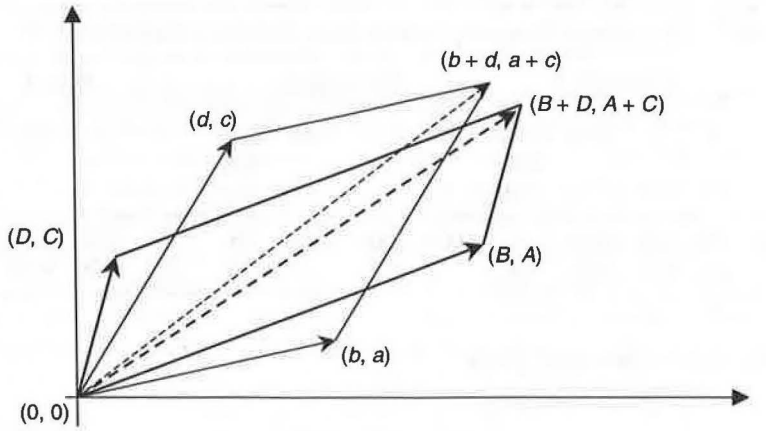

Figure 30.1 A graphical illustration of Simpson's Paradox.

universities as the sum of the weighted proportions of females in each university, where the weight is the relative size of the university:

$$\frac{a}{b}\left(\frac{b+B}{b+B+d+D}\right)+\frac{c}{d}\left(\frac{d+D}{b+B+d+D}\right)=\frac{a(b+B)/b+c(d+D)/d}{b+B+d+D}.$$

Likewise, the adjusted proportion of males who study economics across the two universities is

$$\frac{A}{B}\left(\frac{b+B}{b+B+d+D}\right)+\frac{C}{D}\left(\frac{d+D}{b+B+d+D}\right)=\frac{A(b+B)/B+C(d+D)/D}{b+B+d+D}.$$

For the data presented earlier, these formulae give 16.8% for females and 20.1% for males. We see that the directionality of the association is now preserved (i.e., a higher proportion of males than females in each university and also overall).

Simpson's Paradox is eponymously attributed to the statistician Edward H. Simpson (b. 1922) (Simpson, 1951), but a similar phenomenon was first mentioned by the eminent British statistician Karl Pearson (1857–1936) (Fig. 30.2) and his colleagues in 1899 (Pearson et al., 1899). These authors noted

> We are thus forced to the conclusion that a mixture of heterogeneous groups, each of which exhibits in itself no organic correlation, will exhibit a greater or less amount of correlation. This correlation may properly be called spurious, yet as it is almost impossible to guarantee the absolute homogeneity of any community, our results for correlation are always liable to an error, the amount of which cannot be foretold. To those who persist in looking upon all correlation as cause and effect, the fact that correlation can be produced between two quite uncorrelated characters A and B by taking an artificial mixture of two closely allied races, must come rather as a shock.

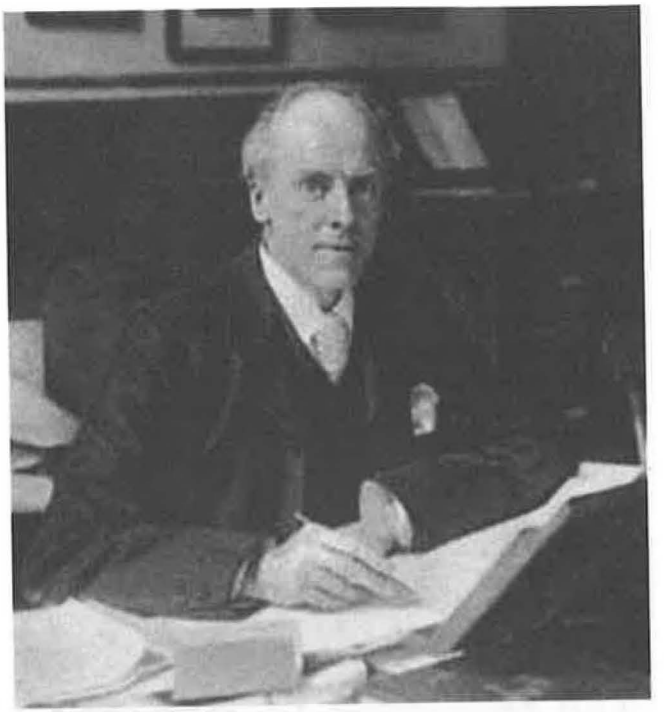

Figure 30.2 Karl Pearson (1857–1936).

Four years later, the renowned British statistician George Udny Yule (1871–1951) (Fig. 30.3), who had previously been Pearson's assistant, further delineated the problem (Yule, 1903).† Yule considers three attributes A, B, and C, such that A and B are independent of each other, given C or $\overline{C}$. Yule then proves that

> . . .there will be apparent association between A and B in the universe at large unless either A or B is independent of C.

The last two examples do not really conform to the probabilities given in **Problem 30** because there is no *reversal* in the direction of the association. When an actual reversal occurs, we have a strong form of *Simpson's Paradox*. On the other hand, we have a weak form of the paradox when the three expressions hold simultaneously:

$$\Pr\{X|YZ\}=\Pr\{X|\overline{Y}Z\},$$
$$\Pr\{X|Y\overline{Z}\}=\Pr\{X|\overline{Y}\overline{Z}\},$$
$$\Pr\{X|Y\}<\Pr\{X|\overline{Y}\}.$$

(The direction of the last inequality could, of course, be reversed.)

† Simpson's paradox is also sometimes called Yule's paradox or Yule–Simpson's paradox.

Figure 30.3 George Udny Yule (1871–1951).

The first actual instance of the strong form of *Simpson's Paradox* was demonstrated in Cohen and Nagel's acclaimed *An Introduction to Logic and Scientific Method* (Cohen and Nagel, 1934, p. 449). These authors present tables for mortality rates from tuberculosis in Richmond and New York City in 1910 as an exercise in their book (see Table 30.3).

They then write

> Notice that the death rate for Whites and that for Negroes were *lower* in Richmond than in New York, although the *total* death rate was *higher*. Are the two populations compared really *comparable*, that is, homogeneous?

Simpson's own 1951 paper was partly motivated by a set of 2×2 tables presented in Kendall's *Advanced Theory of Statistics* (Kendall, 1945, p. 317). In the first of these tables, Kendall displayed the frequencies for two attributes in a population, and showed that they were independent. The author then splits this 2×2 table into two 2×2 tables, one for males and one for females. Kendall then shows the two attributes are

Table 30.3 Rates from Tuberculosis in Richmond and New York City in 1910 as used by Cohen and Nagel (1934, p. 449)

	Population		Deaths		Death rate per 100,000	
	New York	Richmond	New York	Richmond	New York	Richmond
White	4,675,174	80,895	8,365	131	179	162
Colored	91,709	46,733	513	155	560	332
Total	4,766,883	127,628	8,881	286	187	226

Table 30.4 2×2 Tables Considered by Simpson (1951)

	Male		Female	
	Untreated	Treated	Untreated	Treated
Alive	4/52	8/52	2/52	12/52
Dead	3/52	5/52	3/52	15/52

positively associated in the male subpopulation and negatively associated in the female subpopulation. He concludes

> The apparent independence in the two together is due to the cancelling of these associations in the sub-populations.

Motivated by this example, Simpson considered the situation where two attributes are positively associated in each of two 2×2 tables (Simpson, 1951). He then showed that there is no net association in the aggregated 2×2 table (see Table. 30.4).

> This time we say that there is a positive association between treatment and survival both among males and among females; but if we combine the tables we again find that there is no association between treatment and survival in the combined population. What is the "sensible" interpretation here? The treatment can hardly be rejected as valueless to the race when it is beneficial when applied to males and to females.

We end by stressing that *Simpson's Paradox* is not a consequence of small sample size. In fact, it is a mathematical, rather than statistical, phenomenon, as can be seen from the algebraic inequalities in Eqs. (30.4a) and (30.4b).