Simpson's Paradox and the Hot Hand in Basketball

Robert L. WARDROP

A number of psychologists and statisticians are interested in how laypersons make judgments in the face of uncertainties, assess the likelihood of coincidences, and draw conclusions from observation. This is an important and exciting area that has produced a number of interesting articles. This article uses an extended example to demonstrate that researchers need to use care when examining what laypersons believe. In particular, it is argued that the data available to laypersons may be very different from the data available to professional researchers. In addition, laypersons unfamiliar with a counterintuitive result, such as Simpson's paradox, may give the wrong interpretation to the pattern in their data. This paper gives two recommendations to researchers and teachers. First, take care to consider what data are available to laypersons. Second, it is important to make the public aware of Simpson's paradox and other counterintuitive results.

KEY WORDS: Hot hand phenomenon; McNemar's test; Multiple analyses; Simpson's paradox.

1. INTRODUCTION

Schoolchildren routinely learn to identify optical illusions. It is arguably as important that the general public learn to identify statistical illusions. Many outstanding researchers have addressed this issue. As examples, Diaconis and Mosteller (1989) investigate computing the probabilities of coincidences; Kahneman, Slovic, and Tversky (1983) consider judgments made in the presence of uncertainty; and Tversky and Gilovich (1989) investigate the popular belief in the hot hand phenomenon in basketball. This article examines some of the data presented by Tversky and Gilovich.

Suppose that a basketball player plans to attempt 20 shots, with each shot resulting in a hit or a miss. A statistician might assume tentatively that the assumptions of Bernoulli trials are appropriate for this experiment. Suppose next that the experiment is performed and the player obtains the following data:

HMHMM MHHHM HHHMM HMHHH

Do these data provide convincing evidence against the tentative assumption of Bernoulli trials? Are the three occurrences of three successive hits convincing evidence of the player having a "hot hand"? These are difficult questions to answer because of the myriad of possible alternatives to Bernoulli trials that exist. It is mathematically and conceptually convenient to restrict attention to alternatives that allow the probability of success on any trial to depend on the outcome of the previous trial or, perhaps, the outcomes of some small number of previous trials. (This restriction may be unrealistic, but that issue will not be addressed in this article.) With the restrictive class of alternatives described here, Tversky and Gilovich devised a clever experiment to obtain convincing evidence that knowledgeable basketball fans are much too ready to detect occurrences of streak shooting—the hot hand—in sequences that are, in fact, the outcomes of Bernoulli trials.

Having established that basketball fans detect the hot hand in simulated random data, Tversky and Gilovich next examined three sets of real data. The data sets are: shots from the field during National Basketball Association (NBA) games; pairs of free throws shot during NBA games; and a controlled experiment using college varsity men and women basketball players. Using the restrictive alternatives described above, Tversky and Gilovich found no evidence of the hot hand phenomenon in any of their data sets. In addition, using a test statistic that is sensitive to certain time trends in the probability of success, they again found no evidence of the hot hand phenomenon.

This article examines the free throw data presented by Tversky and Gilovich. Tversky and Gilovich began by asking a sample of 100 "avid basketball fans" from Cornell and Stanford: "When shooting free throws, does a player have a better chance of making his second shot after making his first shot than after missing his first shot?" A "Yes" response was interpreted as indicating belief in the existence of the hot hand phenomenon, and a "No" as indicating disbelief. (Actually, a "No" response combines persons who believe in independence with those who believe in a negative association between shots; but the researchers apparently were not interested in separating these groups.) Sixty-eight of the fans responded "Yes" and the other 32 "No." Thus, a large majority of those questioned believed in the hot hand phenomenon for free throw shooting. Tversky and Gilovich investigated the above question empirically by examining data they obtained on a small group of well-known and widely viewed basketball players, namely, nine regulars on the 1980–1981 and 1981–1982 Boston Celtics basketball team.

After their analysis of the Celtics data, Tversky and Gilovich concluded that "These data provide no evidence that the outcome of the second shot depends on the outcome of the first." Section 2 of this article will examine the Celtics data with the goal of reconciling what Tversky and Gilovich found and what their basketball fans believed. In particular, it will be shown that, in a certain sense, the prevalent fan belief in the hot hand is not necessarily at odds with Tversky and Gilovich's conclusion.

The analysis presented in Section 3 of this paper indicates that several Celtics players were better at their second shots than at their first.

2. INDEPENDENCE

It is instructive to begin by considering just two of the nine Boston Celtics players who are represented in the free throw data, namely, Larry Bird and Rick Robey. During

Robert L. Wardrop is Associate Professor, Department of Statistics, University of Wisconsin—Madison, Madison, WI 53706. The author thanks the referees and associate editor for helpful comments.

Table 1. Observed Frequencies for Pairs of Free Throws by Larry Bird and Rick Robey, and the Collapsed Table

Larry Bird				Rick Robey				Collapsed Table			
	Second:				Second:				Second:		
First:	Hit	Miss	Total	First:	Hit	Miss	Total	First:	Hit	Miss	Total
Hit	251	34	285	Hit	54	37	91	Hit	305	71	376
Miss	48	5	53	Miss	49	31	80	Miss	97	36	133
Total	299	39	338	Total	103	68	171	Total	402	107	509

the 1980–1981 and 1981–1982 seasons, Larry Bird shot a pair of free throws on 338 occasions. Five times he missed both shots, 251 times he made both shots, 34 times he made only the first shot, and 48 times he made only the second shot. These data are presented in Table 1, as are the same data for Rick Robey. Let $\hat{p}_{\text{hit}}$ and $\hat{p}_{\text{miss}}$ denote the proportion of first shot hits that are followed by a hit and the proportion of first shot misses that are followed by a hit, respectively. For Bird, $\hat{p}_{\text{hit}} = 251/285 = .881$ and $\hat{p}_{\text{miss}} = 48/53 = .906$. For Robey, these numbers are .593 and .612, respectively. Note that, contrary to the hot hand theory, each player shot slightly better after a miss than after a hit, although, as shown below, the differences are not statistically significant.

It is possible, of course, to ignore the identity of the player attempting the shots and examine the data in the collapsed table in Table 1. For example, on 509 occasions either Bird or Robey attempted two free throws, on 305 of those occasions both shots were hit, and so on. For the collapsed table, $\hat{p}_{\text{hit}} = .811$ and $\hat{p}_{\text{miss}} = .729$. These values support the hot hand theory—a hit was much more likely than a miss to be followed by a hit.

The data from Bird and Robey illustrate Simpson's paradox (Simpson 1951), namely, $\hat{p}_{\text{hit}} < \hat{p}_{\text{miss}}$ in each component table, but $\hat{p}_{\text{hit}} > \hat{p}_{\text{miss}}$ in the collapsed table. For further examples and discussion of Simpson's paradox, see Shapiro (1982), Wagner (1982), the essay by Alan Agresti in Kotz and Johnson (1983), and their references.

Figure 1 provides a visual explanation of Simpson's paradox. The top picture in the figure presents the proportion of second-shot successes after a hit for Bird, Robey and the collapsed table. The bottom picture in the figure presents the same three proportions for second shots attempted after a miss. It is easy to verify algebraically that the proportion of successes for a collapsed table equals the weighted average of the individual player's proportions, with weights equal to the proportion of data in the collapsed table that comes from the player. For the after-a-hit condition, for example, the weight for Bird is $285/376 = .758$, the weight for Robey is $91/376 = .242$, and the proportion of successes for the collapsed table, $305/376 = .811$, is

$$\frac{285}{376} \times \frac{251}{285} + \frac{91}{376} \times \frac{54}{91}.$$

In Figure 1, the heights of the four rectangles above the Bird and Robey proportions equal the weights associated with the relevant player–condition pair. For example, the height of the rectangle for Bird in the after-a-hit condition equals .758, in agreement with the computation of the previous paragraph. Thus, the proportion of successes for each collapsed table in the figure is located at the center of gravity of the two rectangles. As a result, even though both Bird and Robey shot better after a miss than after a hit, the collapsed values show the reverse pattern due to the huge variation in weights associated with each player. In short, Simpson's paradox has occurred because the after-a-miss condition, when compared to the after-a-hit condition, has a disproportionately large share of its data originating from the far inferior shooter Robey.

When I first examined the Bird and Robey data several years ago, my immediate reactions were that this is an interesting example of Simpson's paradox, the analysis of individual tables is "correct," and the analysis of the collapsed table is "incorrect." Now I believe these labels were applied too hastily. The reasons I changed my mind are discussed below after the entire data set is examined.

Table 2 introduces symbols to represent the various numbers in a 2×2 table. The values n_1, n_2, m_1, and m_2 denote the marginal totals, and the values of a, b, c, and d denote the cell counts. The null hypothesis states that the outcome of the second shot is statistically independent of the outcome of the first shot. If the null hypothesis is true, then conditional on the values of the marginal totals, the cell count a has a hypergeometric distribution with

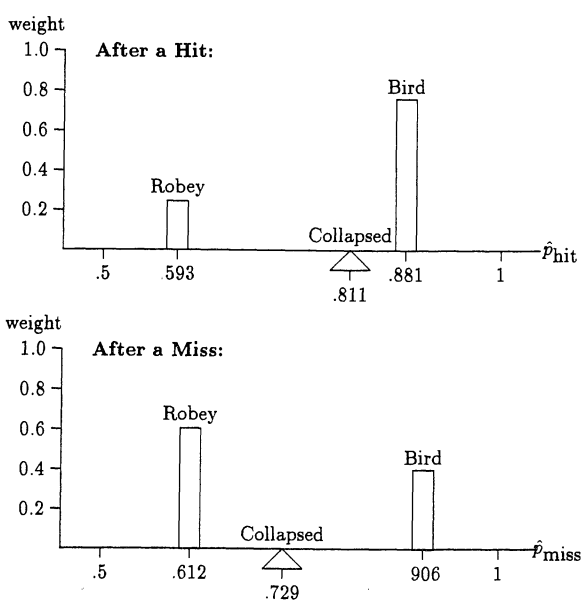

Figure 1. A Visual Explanation of Simpson's Paradox for the Free Throw Study.

Table 2. Standard Notation for a 2 × 2 Table

First:	Second: Hit	Miss	Total
Hit	a	b	n_1
Miss	c	d	n_2
Total	m_1	m_2	n

expectation and variance:

$$E(a) = \frac{n_1 m_1}{n} \tag{1}$$

and

$$\text{var}(a) = \frac{n_1 n_2 m_1 m_2}{(n-1)n^2}. \tag{2}$$

The null distribution of

$$Z = \frac{a - E(a)}{\sqrt{\text{var(a)}}} \tag{3}$$

can be approximated by the standard normal curve. For Larry Bird, $a = 251, E(a) = 252.12$, and $\text{var}(a) = 4.575$. Substituting these values into Equation (3) gives

$$z = \frac{251 - 252.12}{\sqrt{4.575}} = -.52.$$

Thus, as stated earlier, the results are not statistically significant. For Robey, $z = -.25$, and for the collapsed table, $z = 1.99$. Thus, an analysis of the collapsed table alone would lead one to conclude that there is statistically significant evidence in support of the hot hand theory.

Tversky and Gilovich report data for all nine men who played regularly for the Celtics during 1980–1982. The summaries needed for analysis are given in Table 3. The first column of the table lists the players' names. The second and third columns list the values of $\hat{p}_{\text{hit}}$ and $\hat{p}_{\text{miss}}$ defined above. The fourth, fifth, and sixth columns list the values of $a, E(a)$, and $\text{var}(a)$ which are obtained from their data and Equations (1) and (2). The seventh column lists the value of z from Equation (3) for each player. The men are listed in the table by decreasing values of $\hat{p}_{\text{hit}} - \hat{p}_{\text{miss}}$ which, not too surprisingly, also lists them by decreasing values of z. Thus, McHale, with a difference of $73 - 59 = 14$ percentage points, is listed first and Carr, with a difference of $68 - 81 = -13$ percentage points, is listed last. In terms of either the point estimates or the test statistic value, McHale provides the strongest evidence in support of the hot hand theory, and Carr provides the strongest evidence in support of an inverse relationship between the outcomes of the two shots. Note that four players—McHale, Maxwell, Parish, and Archibald—shot better after a hit, while the remaining five players shot better after a miss.

The data for McHale give a one-sided approximate P value of .0418. This is not particularly noteworthy for two reasons:

(1) It is difficult to justify the use of a one-sided alternative, especially given that five players shot better after a miss and four shot better after a hit.

Table 3. Selected Statistics for the Investigation of Independence of Shots for Nine Members of the Boston Celtics

Player	$\hat{p}_{hit}$	$\hat{p}_{miss}$	a	$E(a)$	var(a)	z
Kevin McHale	.73	.59	93	88.23	7.633	1.73
Cedric Maxwell	.81	.76	245	240.20	14.667	1.25
Robert Parish	.77	.72	164	160.75	13.061	.90
Nate Archibald	.83	.82	203	202.26	8.380	.26
Rick Robey	.59	.61	54	54.81	10.257	−.25
Gerald Henderson	.76	.78	77	77.58	4.858	−.26
Larry Bird	.88	.91	251	252.12	4.575	−.52
Chris Ford	.71	.77	36	37.03	3.100	−.58
M. L. Carr	.68	.81	39	41.20	3.620	−1.16

(2) Even if one believes a one-sided alternative is appropriate, on the assumption that all nine players have independence between shots, the approximate probability is $1 - (1 - .0418)^9 = .32$, or about one-third, that at least one of the nine P values would be as small or smaller than McHale's.

Table 4 presents the observed frequencies and row proportions for the free throw data collapsed over the nine Celtics under investigation. For the collapsed table, the relative frequency of a hit after a hit is $78.9 - 74.3 = 4.6$ percentage points higher than the relative frequency of a hit after a miss. Moreover, for the collapsed table, it can be shown that $a = 1{,}162, E(a) = 1{,}143.03$, and $\text{var}(a) = 72.015$, yielding $z = 2.24$, which is statistically significant.

To summarize, separate analyses of individual players indicate that four players shot better after a hit and five players shot better after a miss, but none of the individual player patterns is convincing. By contrast, the analysis of the collapsed table gives statistically significant evidence in support of the hot hand phenomenon.

In view of the Celtics data, what, if anything, are we to make of the fact that 68 out of 100 of Tversky and Gilovich's avid basketball fans believe in the hot hand phenomenon for free throw shooting? Perhaps these fans have been watching players who do exhibit the hot hand. Perhaps these fans see patterns in data where no patterns exist. I prefer the following explanation.

I am an avid basketball fan. Over the past 30 years, I have observed several thousand different players shooting free throws. It is difficult to imagine that I (or any other basketball fan) could remember the equivalent of thousands of 2 × 2 tables. Yet these individual tables are exactly what I would need in order to investigate properly the question of the hot hand phenomenon. It is much more reasonable to assume that I have a single 2 × 2 table

Table 4. Observed Frequencies and Row Proportions for Free Throw Data Collapsed Over Nine Celtics

First:	Second: Hit	Miss	Total	First:	Second: Hit	Miss	Total
Hit	1,162	311	1,473	Hit	.789	.211	1.000
Miss	428	148	576	Miss	.743	.257	1.000
Total	1,590	459	2,049				

in my mind, namely, the collapsed table for all players I have seen. Just like the Celtics data, my collapsed table indicates that a success is more likely than a failure to be followed by a success. Thus, there *is* a pattern in the data that are reasonably available to me and, I conjecture, in the data that are reasonably available to Gilovich and Tversky's 100 basketball fans. It seems reasonable to suggest to basketball fans that the mental equivalent of Simpson's paradox could lead to a cognitive statistical illusion that results in their "seeing patterns in the data that do not exist."

3. STATIONARITY

Tversky and Gilovich correctly concluded that there is no evidence of the hot hand phenomenon in the free throw data. In this section, it is demonstrated, however, that the simple model of Bernoulli trials is also inappropriate. In particular, it is shown that several of the Celtics players shot significantly better on their second free throw, perhaps as a result of the practice afforded by the first shot.

Look at Table 1 again. Larry Bird made 84.3% (285 of 338) of his first shots compared to 88.5% (299 of 338) of his second shots. Thus, there is evidence that he improved on his second shot. The null hypothesis that his probability of success was constant can be investigated with McNemar's test, which uses the fact that the null distribution of

$$Z_1 = \frac{b - c}{\sqrt{b + c}} \tag{4}$$

can be approximated by the standard normal curve. (Recall that b and c are defined in Table 2.) For Larry Bird, $b = 34$ and $c = 48$, giving

$$z_1 = \frac{34 - 48}{\sqrt{34 + 48}} = -1.55.$$

The same analysis can be performed for the other eight Celtics; the results are given in Table 5. The first column of the table lists the player's names. The second and third columns list, respectively, the relative frequencies of successes on the first and second shots. The remaining columns list the values of b and c from each player's 2×2 table and the value of z_1 computed from Equation (4). The players are listed according to the difference in relative frequencies between the first and second shots.

Table 5. Selected Statistics for Comparing the Success Rates on the First and Second Free Throws for Nine Members of Boston Celtics

Player	$\hat{p}(S_1)$	$\hat{p}(S_2)$	b	c	z_1
Cedric Maxwell	.70	.80	57	97	−3.22
Robert Parish	.67	.75	49	76	−2.41
Nate Archibald	.76	.83	42	62	−1.96
Rick Robey	.53	.60	37	49	−1.29
Larry Bird	.84	.88	34	48	−1.55
Gerald Henderson	.73	.77	24	29	−.69
M. L. Carr	.69	.72	18	21	−.48
Chris Ford	.70	.73	15	17	−.35
Kevin McHale	.72	.69	35	29	.75
Total	—	—	311	428	$z_2 = -4.30$

Thus, Maxwell, who shot ten percentage points better on the second shot than on the first, is listed first, and McHale, who shot three percentage points better on the first shot, is listed last. Note the following features of the data.

(1) Eight of nine players had a higher success rate on their second shots.

(2) Three players had one-sided approximate P values below .05: Maxwell (.0006), Parish (.0080), and Archibald (.0250). The interpretation of these P values should take into account that nine tests were performed. If, in fact, each player had a constant success rate on his two shots, the approximate probability of obtaining at least one P value equal to or smaller than .0006 is: $1 - (1 - .0006)^9 = .0054$. Similarly, the approximate probability of obtaining at least two P values equal to or smaller than .0080 is .0022. Finally, the approximate probability of obtaining at least three P values equal to or smaller than .0250 is .0012. Thus, the three statistically significant results do not seem to be attributable to the execution of many tests.

(3) McNemar's test can be viewed as testing that a Bernoulli trial success probability equals .5 based on a sample of size $b + c$. Thus, several of the analyses of individual players presented in Table 5 are based on very little data and, hence, have very low power. To combat this difficulty, it is instructive to combine the data across the nine players. In particular, if the null hypothesis of constant success probability is true for all nine players, then the observed value of

$$Z_2 = \frac{\sum(b - c)}{\sqrt{\sum(b + c)}},$$

where the sum is taken over the nine tables, can be viewed as an observation from a distribution that is approximately the standard normal curve. The observed value of Z_2 is −4.30, given in the bottom row of Table 5. This value indicates that there is overwhelming evidence against the assumption that all nine null hypotheses are true.

4. SUMMARY

This article puts forth an argument to reconcile what avid basketball fans believe and what Tversky and Gilovich found. It is argued that the fans and the researchers were analyzing different sets of data. While the researcher's data had no pattern, the fan's data had a pattern. This pattern, however, was due to the effects of aggregation and not the hot hand phenomenon. This finding indicates that researchers should take care to consider what data are available to laypersons. In addition, this finding underscores the importance of increasing the awareness of statistical fallacies among the general public.

This article also demonstrates that several Celtics players showed a significant improvement in their shooting ability on the second free throw. Thus, while the hot hand phenomenon is not supported by these free throw data, neither is the simple model of Bernoulli trials.

[*Received March 1992. Revised November 1993.*]

REFERENCES

Diaconis, P., and Mosteller, F. (1989), "Methods for Studying Coincidences," *Journal of the American Statistical Association*, 84, 853–861.

Kahneman, D., Slovic, P., and Tversky, A. (1983), *Judgement Under Uncertainty: Heuristics and Biases*, Cambridge, U.K.: Cambridge University Press.

Kotz, S., and Johnson, N. L. (eds.) (1983), *Encyclopedia of Statistical Science* (Vol. 3), New York: John Wiley, pp. 24–28.

Shapiro, S. H. (1982), "Collapsing a Contingency Table—A Geometric Approach," *The American Statistician*, 36, 43–46.

Simpson, E. H. (1951), "The Interpretation of Interaction in Contingency Tables," *Journal of the Royal Statistical Society*, Ser. B, 13, 238–241.

Tversky, A., and Gilovich, T. (1989), "The Cold Facts About the 'Hot Hand' in Basketball," *CHANCE: New Directions for Statistics and Computing*, 2, 16–21.

Wagner, C. H. (1982), "Simpson's Paradox in Real Life," *The American Statistician*, 36, 46–47.