Simpson's Paradox and Investment Management

Gilbert W. Bassett Jr.*

Professor of Economics and Finance at the University of Illinois at Chicago. This paper benefited from research I am doing at the Quantitative Equities division of Weiss, Peck, and Greer; Investments.

An earlier version of this paper appeared in Stats(1994). This version was reprinted in Peter Bernstein's December 1,1996 newsletter.

PORTFOLIO RETURNS

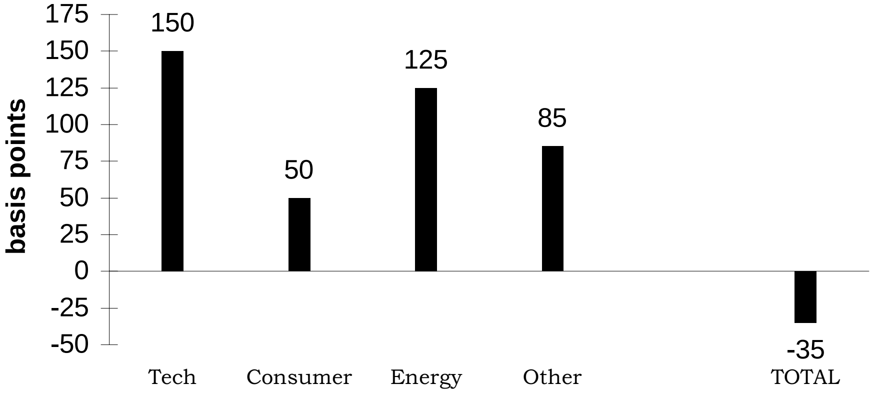
One last report and he'd be batting a thousand. The strategy of hiring specialized sector managers was looking better than he could have hoped. He had disaggregated the S&P500 into four homogeneous sectors and hired managers to handle each separate portfolio. He was waiting for the numbers from the last manager. The other three had already reported. Each had outperformed their sector benchmarks. One more and it would be bonus time.

The assistant handed him the fax from the last manager. It was a clean sweep. All the managers outperformed their benchmarks. Active returns--returns in excess of the benchmark--ranged from +50 to +150 basis points. He told his assistant to compute aggregate performance numbers and prepare a chart for presentation to the executive committee.

An hour later his assistant returned saying something was very wrong. The computer must be broken. The chart in Figure 1 showed that overall performance had been worse than the benchmark. The data seemed to have been entered correctly, but showed overall returns were 35 basis points worse than the S&P500. He knew this was impossible. Overall performance could be below the benchmark if at least one of the sector returns had been negative, but every return was positive. There was no way that the weighted average of four positive numbers could be negative. He called Research.

Figure 1

Active Returns



Research checked the data and confirmed that nothing was wrong. The chart was accurate: they had underperformed the S&P500 even though every manager had positive active returns. It wasn’t a problem with the computer, it was Simpson’s paradox. Not OJ, Homer, or Thomas, but E.H. Simpson who had written an article in the early 1950's. Research knew all about the paradox because of recent data anomalies they had encountered.

Research had wanted to assess fundamental (F) and technical (T) investment strategies. Managers of many funds were divided into F or T according to their main strategy, and their historical performance was categorized as good or poor. Results are summarized in Table 1; complete data is in the appendix. The clear winner was F. Good performance was found for 68.75% of the F, but for only 55.2% of the T strategies. The test statistic indicated the probability was virtually zero that such a difference could have arisen by chance.

The initial findings were subsequently reanalyzed by technical analysts in Research. They suspected that while T did not do well overall, it might do better in certain market situations such as bull markets. The data was partitioned by market type and relative performance was assessed restricting attention to bull markets.

Table 1

	Strategy	
	Fundamental	Technical
Good	68.75%	55.20%
Poor	31.25%	44.80%

The investigation showed what had been suspected, namely that T outperformed F in bull markets; see Table 2. So, at least in certain situations, T was better than F. Technical analysts argued that they should not be downsized because performance assessment needed to account for differing market conditions, and at least in bull markets their T strategy was best. (Technical analysts did not look at T and F in bear markets. This was deemed unnecessary: since F was best overall, it had to have done much better in bear markets).

Table 2

	Strategy-Bull Markets	
	Fundamental	Technical
Good	71.50%	76.00%
Poor	28.50%	24.00%

When Research eventually did investigate how much better F did in bear markets it discovered that the impossible had happened. The data presented in Table 3 shows that F did worse than T in both bull and bear markets. During bear markets neither strategy did very well, but T did better. Hence, the T strategies had outperformed F in bull and bear markets even though the F strategies had done better overall. The original data, once so compelling for F, now seemed to support technical strategies.

Table 3

	Bull Market		**Bear Market**	
Performance	Fund.	Tech.	Fund.	Tech.
Good	71.50%	76.00%	44.00%	50.00%
Poor	28.50%	24.00%	56.00%	50.00%

Research eventually learned that the superior performance of the T strategy in bull and bear markets, despite doing worse with the combined data, is an instance of Simpson's paradox. It is clear that data showing F better overall, can be partitioned so that T does better in some particular subset of the data. It seems equally obvious however that there could be no partition such that the inferior T does better than F in every subset of the data. Simpson's paradox says however that this can occur; *it is possible for any relation between two variables to be reversed in every component subpopulation.*

A GRAPHICAL EXPLANATION

Fundamental and technical performance in bull and bear markets is depicted in Figure 1; this graphical explanation of the paradox follows Wittmer(1992). The vertical axis indicates how well each method performed. The figure shows T above F in both the bull and bear markets. The combined performance of a strategy is indicated by a point on the line segment connecting the two market conditions. The location of the combined point depends on the relative proportion of cases. If relative proportions were the same then T>F in both situations would mean the combined T must be located above the combined F. As indicated in the figure however, when relative proportions differ a Simpson reversal becomes possible and the combined F is higher than the combined T.

For example, the overall proportion of good performance for the F strategy must be between the 71.5% and 44% values in bull and bear markets. The combined value is the weighted average, 68.75% = .9(71.5%) + .1(44%) where the weights of .9 and .1 are determined by the fraction of F managers who participated in a bull and bear market, respectively. Likewise, for T we have 55.2% = .2(76%) + .8(50%)

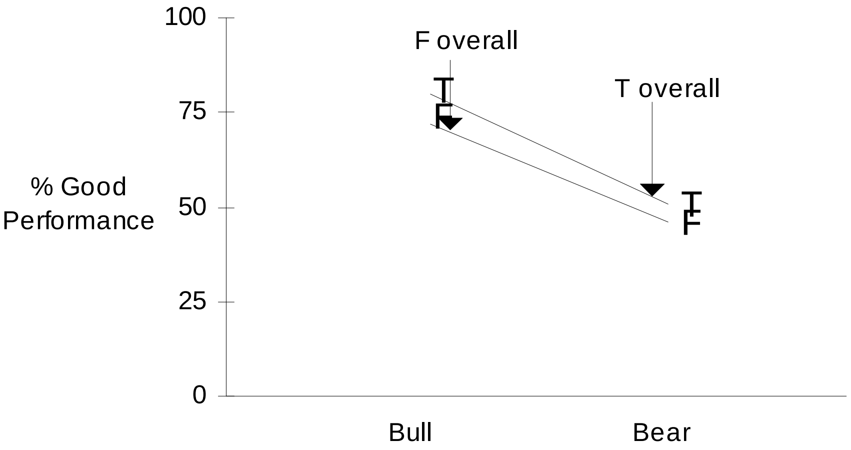


Figure 2

Thus, T is better than F in each type of market, but the weighted average for F is greater than the weighted average for T. We can now see why this occurred. Most F investors participated in a bull market, in which success rates were quite high, whereas most T investors participated in a bear market, in which success rates were low.

In a similar fashion, the negative returns for the firm occurred because the poorest sector returns had occurred for the sector that was slightly overweight relative to the S&P500 index; see Table 4.

Table 4

	Managed Portfolio		**S&P500**		
Sector	Weight	Return %	Weight	Return %	Active
Tech	0.20	31.50	0.23	30.00	1.50
Consumer	0.28	-10.00	0.25	-10.50	0.50
Energy	0.23	16.25	0.24	15.00	1.25
Other/Misc.	0.29	4.85	0.28	4.00	0.85

Managed (%)	8.64
S&P500(%)	9.00
Excess Return (bp)	-35

Simpson's paradox can arise in real life, and it can be used to construct interesting counterfactual situations.

Example 1 -- ACT scores. In 1987 the average ACT composite score was 18.7. In 1988 the average increased to 18.8. There are two categories of students who take the ACT: those who have completed a core curriculum of courses in preparation for college and those who have not. In 1987 the average score among those who had completed a core curriculum was 21.2; this average went down to 21.1 in 1988. Likewise, in 1987 the average score among those who had not completed a core curriculum was 17.3; this average went down to 17.1 in 1988. So, scores dropped in both groups, but the overall average went up; see Witmer(1992, p.113).

Example 2a. A firm consists of ten subsidiaries. The firm is sued for discrimination because last year the proportion of women employed decreased in every one of its subsidiaries. The firm wins the case when it shows the court that the proportion of women employed by the firm last year actually increased.

Example 2b. A firm consists of ten subsidiaries. The firm is sued for discrimination because last year the proportion of women employed decreased. The firm wins the case when it shows the court that last year the proportion of women employed in each of its ten subsidiaries actually increased. (The two parts of this example are hypothetical, but for similar examples see Bickel et al.(1975), Wagner(1982), and Nozick(1993 p.104, 200n59).

Example 3. A drug is being tested at widely separated sites. At every location the drug leads to greatly improved recovery rates as compared with the standard treatment. The data is sent to Washington before asking for FDA approval. The combined data is never shown to the FDA because it is found that the drug has killed nearly everyone who has taken it. The mortality rate for the new drug is greater than 90%, as compared to 50% for the standard treatment. (This too is hypothetical, but see Blyth(1972) for similar examples.)

THE JANUARY EFFECT

New research began in search of the January effect. Some suspected that, after controlling for market conditions, F would do better than T during January.

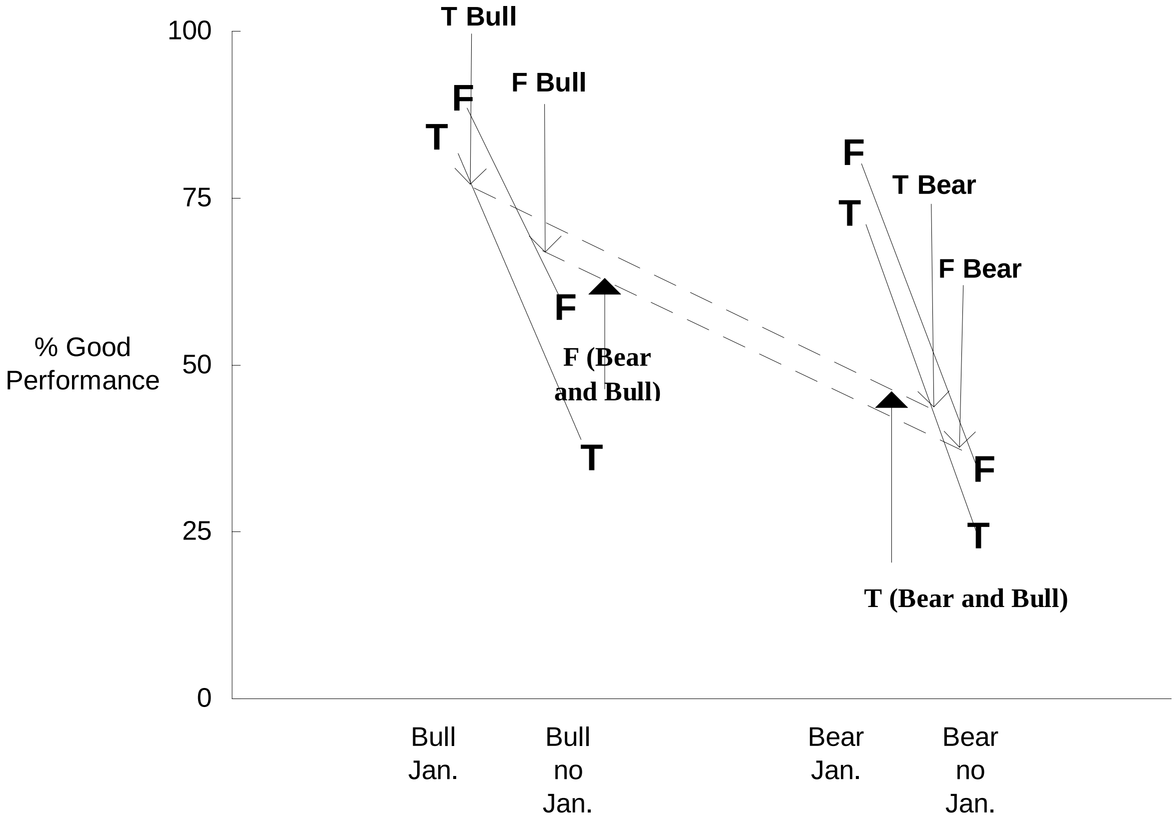
The data was partitioned again, now based on whether the strategy was in effect in January. It showed F doing better in January. But it also showed F doing better during the rest of the year as well. As Table 5 shows everything was re-reversed; F outperformed T in each of the comparisons. This happened even though with the same data F lost to T in bull and bear markets, and even though F beat T overall.

Table 5

	January				Non January			
	Bull		Bear		Bull		Bear	
	Fund.	Tech.	Fund.	Tech.	Fund.	Tech.	Fund.	Tech.
Good	85%	80%	80%	75%	70%	40%	40%	25%
Poor	15%	20%	20%	25%	30%	60%	60%	75%

If data can be partitioned so that there is one Simpson reversal then it must be possible for reversals to continue as data is partitioned further. Such a situation is illustrated in Figure 3. It shows the performance of F and T strategies in bull/bear markets and, additionally, according to the new January condition. This partitioning leads to four subsets of the original data. As indicted in the figure, F is above T in every case. Aggregating up a level--to market conditions--leads to Figure 1 with the reversed conclusion that T did better. Aggregating up yet another level to the entire data set reverses this conclusion with F doing better than T. Which strategy really is best?

Figure 3



CONCLUSION

It seems obvious, at least at first glance, that positive returns in every sector guarantee that active total returns will be positive. Similarly, if T is better than F in every possible setting, then T must be better than F overall. As we have seen, this is not true. Beating every sector can result in total returns below the benchmark. Data compelling for one hypothesis can be reversed in every subpopulation thus supporting the contrary position.

The possibility of Simpson's paradox raises important issues regarding the interpretation of data. Knight used “risk” to refer to problems in which there is a well-defined model and data follows a given distribution. In this situation the statistical inference problem is well-specified and Simpson’s paradox can be hedged by making sure samples are random. Rational expectations and enough data will be sufficient to determine whether T or F is best.

This contrasts with a situation of “uncertainty” where the exact form of underlying distributions are not be well defined and the beliefs of rational investors need not converge. More than being uncertain about exact distributions, lack of convergence can arise from alternative models for the inference problem. Recall Keynes’ famous analogy for the professional investment problem:

> Or, to change the metaphor slightly, professional investment may be likened to the those newspaper competitions in which the competitors have to pick out the six prettiest faces from a hundred photographs, the prize being awarded to the competitor whose choice most nearly corresponds to the average preferences of the competitors as a whole; so that each competitor has to pick, not those faces which he himself finds the prettiest, but those which he thinks likeliest to catch the fancy of the other competitors, all of whom are looking at the problem for the same point of view. It is not a case of choosing those which, to the best of one's judgment, are really the prettiest, nor even those which average opinion genuinely thinks the prettiest. We have reached the third degree where we devote our intelligences to anticipating what average opinion expects the average opinion to be. And there are some, I believe, who practise the fourth, fifth and higher degrees. Keynes (1936, p.156)

The mere possibility of Simpson’s paradox raises issues regarding our ability to make conclusive inferences based on data alone, absent a model for the data. While models are essential, they are also provisional. Since real world data need not satisfy model assumptions exactly, our inferences are also provisional, always threatened by a reasonable partition of the data that would lead to a Simpson reversal.

REFERENCES

Bassett, Gilbert W.(1994). Learning About Simpson’s Paradox, *STATS*, No. 12, p. 13-17.

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

Blyth C. R. (1972). "On Simpson's Paradox and the Sure Thing Principle," *Journal of the American Statistical Association*, Vol.67 p.364-66.

Keynes, John Maynard (1936). *The General Theory of Employment, Interest, and Money*. Harbinger Edition, 1964.

Nozick, Robert (1993). *The Nature of Rationality*. Princeton.

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

Wagner, C. H. (1982). "Simpson's Paradox in Real Life," *The American Statistician.* February 1982, Vol.36, No. 1 p.46-48.

Witmer, Jeffrey A. (1992). *Data Analysis: An Introduction.* Prentice Hall.

APPENDIX

Data

	January				**Non January**			
	Bull		Bear		Bull		Bear	
Performance	Fund.	Tech.	Fund.	Tech.	Fund.	Tech.	Fund.	Tech.
Good	765	1440	80	3000	5670	80	360	1000
Poor	135	360	20	1000	2430	120	540	3000