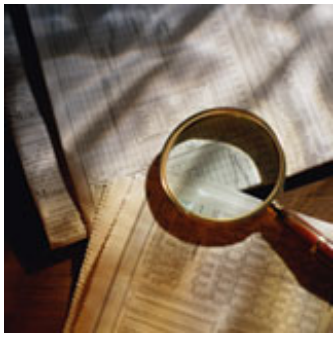
The Use -- and Misuse -- of Statistics: How and Why Numbers Are So Easily Manipulated

Published : April 02, 2008 in Knowledge@Wharton



When a report prepared by former Senator George J. Mitchell indicated that Roger Clemens and more than 80 other Major League Baseball players used illegal, performance-enhancing drugs, the former Cy Young award-winning pitcher went on the offensive to clear his name. Added to Clemens' testimony before Capitol Hill lawmakers was a voluminous report prepared by a marketing agency that relied on statistics to make the case for Clemens' innocence.

But an article in the February 10 edition of the *New York Times* written by four Wharton faculty makes another case: The validity of a statistical analysis is only as good as its individual components. It's a distinction that is likely to gain in importance as organizations and individuals try to make sense of an increasingly large and complex barrage of information.

"Today, consumers of information are drowning in data," says Justin Wolfers, Wharton professor of business and public policy. "Terabytes of data are being generated from the constant measurement of businesses, workers, government and other activity, and there are many ways to draw inferences from the raw data. Unfortunately, many of them lead in the wrong direction."

For example, he says a chain of retail stores may analyze its operations for a set period and find that those times when it reduced its sales prices coincided with times that overall sales fell. "That could lead the chain to conclude that low prices spurred a reduction in sales volume," says Wolfers. "But the true causal link may be deeper than that. Before the retailer raises prices in an attempt to increase sales, it should examine additional issues to see if overall demand during the period was influenced by other factors. For example, perhaps the firm historically runs its semi-annual sales during slow sales periods. If this is the case, low sales are causing price declines, rather than price declines lowering sales."

This illustrates a critical difficulty inherent in applying statistical analysis to business, social science and other settings, says Wolfers. "It's generally easier to isolate and exclude extraneous data when researchers deal with experimental or hard-sciences data, such as medicine," he notes. "In an experimental setting, a pharmaceutical company can randomly assign a drug to one set of subjects and a placebo to the other set. Assuming the researchers have randomized the people who received the drug, they can isolate the outcome to the effect of the drug or the placebo."

But in a business setting, that's not so easy. "In the example of the retail chain, it may be more difficult to isolate the effects of a variety of other influences," Wolfers says. Concerning the change in sale prices, "it would be necessary to consider the effects of sunny days and rainy days, or hot and cold ones, on the volume and behavior of shoppers."

In the Roger Clemens case, Wolfers worked with statistics professors Shane Jensen and Abraham Wyner and marketing professor Eric Bradlow to co-author the *Times* article titled, "Report Backing Clemens Chooses Its Facts Carefully."

In it, the researchers questioned the methodology used by Hendricks Sports Management to support Clemens' denial of using steroids. "The Clemens report tries to dispel this issue by comparing him with Nolan Ryan, who retired in 1993 at [age] 46," the authors write. "In this comparison, Clemens does not look atypical: Both enjoyed great success well into their 40s. Similar conclusions can be drawn when

comparing Clemens with two contemporaries, Randy Johnson and Curt Schilling."

But the Wharton researchers say those comparisons are incomplete. "By comparing Clemens only to those who were successful in the second act of their careers, rather than to all pitchers who had a similarly successful first act, the report artificially minimizes the chances that Clemens' numbers will seem unusual," they write. "Statisticians call this problem 'selection bias.'"

Just as a retailer needs to consider a plausible alternative forecast of what sales would have been in a price-comparison analysis, the Wharton researchers say that the performance of Clemens should be compared against "all highly durable starting pitchers." When that is done, Clemens' "second act is unusual," they write. Most pitchers improve steadily early in their careers, peak at about 30 and then slowly decline. In contrast, Clemens' career declined as he entered his late 20s and then improved through his mid-40s.

When it comes to "statisticians-for-hire," there's a tendency to choose comparison groups that support their clients, note the Wharton researchers. But what about when statistical analyses are used in a situation where the outcome is not tied to a particular point of view? Financial analysis, econometrics, auditing, production and operations are only some of the areas where parties seek unbiased data in order to make good decisions in the face of uncertainty.

Coca-Cola and Mutual Funds

Do things always go better with Coke? That appears to be at the heart of a lawsuit, seeking class action status, filed against the Coca-Cola Company's marketing for Enviga, its caffeinated green-tea drink. The ads for Enviga state that "it actually burns more calories than it provides, resulting in 'negative calories,'" according to the suit, filed in U.S. District Court in Camden, N.J.

It alleges that Coca-Cola's claims are based on "...the abstract of a single, small and short-term study funded by Coke...." The suit goes on to say that while the subjects in the clinical study were relatively lean individuals with an average Body Mass Index (BMI) of 22, "the great majority of Americans are overweight or obese," with a BMI of 25 or more, and would not be likely to lose weight by consuming Enviga. A spokesman for Coca-Cola says the company's study and its results are valid.

Another example of disputed statistics concerns a March *Wall Street Journal* advertisement for the Dreyfus Funds. The ad notes that its Intermediate Term Income Fund achieved a four-star Morningstar rating, says David Peterson, an independent statistical consultant based in the Research Triangle area of North Carolina and a member of the American Statistical Association.

"The ad was careful to point out that past results are no promise of future results, but fails to mention that Dreyfus has at least 19 mutual funds," says Peterson. "Naturally, the best among them at any moment in time is likely to be pretty good although conversely, the worst of them -- which are not mentioned in the advertisement -- are likely pretty bad, even if there is nothing fundamentally unusual about any of the 19 funds."

Using this same principle, he says, a pharmaceutical company "could conduct 10 separate and independent tests of the effectiveness of a new drug, and base its advertising only on the most favorable result."

Mistrust and Miscommunication

The possibility of unintentional errors in any study is also cause for concern, says Wharton's Jensen.

"Even if care is taken to establish a good sample, there are possibilities of misleading results," he notes. "One common problem is data mining. If someone analyses a large dataset for long enough, they are bound to find a statistically significant effect or difference between some set of variables." Unfortunately, he says, researchers often go on to simply report their single significant finding without acknowledging the "many insignificant tests that they did before finding that one result."

According to Jensen, "a proper accounting of the entire testing process is needed to keep these types of results in perspective." But at least two forces routinely work against effective analyses. "The first is a

mistrust of statistical analyses, and the second is a lack of dialogue between academic statisticians and practitioners." In fact, says Jensen, "I've read about many studies in medicine, economics and social science that could benefit from more discussion with statisticians about the analysis of collected data and the collection of the data itself."

Bradlow also voices concern over the interpretation of statistical outcomes. "I always say to my students that data-driven solutions can't always tell you the right answer. Instead, they can tell you which [answers] to eliminate because they are not supported by the data." The true value of a statistical analysis is that it helps users to properly characterize uncertainty as opposed to a "best guess," to realize what outcomes are statistically significant, and to answer specific hypotheses.

"The key issue here is representation," he says, referring back to the Roger Clemens study. "Researchers and users should always concern themselves with how the data are obtained and whether they represent a random sample. If they don't, then one has to be careful in one's conclusions."

Even researchers who do not have an agenda need to exercise caution, according to Bradlow. "In the late 1990s when we collected demographic data at a two-century-old cemetery, it was noted that people who were buried there at a later date [closer to the time of the study] had died at an earlier average age, compared to people who had been buried many years ago," says Bradlow, who wrote up the results in an article for *Chance* magazine titled, "A Selection of Selection Anomalies."

"It's tempting to conclude that mortality has gone up for younger people, but that would be an incorrect conclusion." Instead, he notes, the earlier deaths are a function of the fact that as one approaches the date of Bradlow's survey, the sample of people who were buried at the cemetery under study would be bound to include a disproportionate number of people who died young simply because they were born closer to the survey date.

For Wolfers, a key to minimizing the misuse of statistics involves intuitive plausibility, or understanding the researcher's approach and the interplay of forces. "It's important to know what the drivers are behind the variables," he says. "Once that is established, an observer can better understand and establish causality."

Jensen offers another example of that. "I'm involved in a study that models the fielding ability of major league baseball outfielders. One hypothesis going into the study is that outfielders would have a harder time catching balls hit behind them, forcing them to run backwards, than balls hit in front of them that would require them to run forwards."

But the results indicated that the opposite was true: At any given distance, fielders tended to catch more balls running backwards. "This seemed very counter-intuitive at first," Jensen says. "But it starts to make sense once you consider the hang time [length of time the ball remains in the air]. Balls hit farther are in the air longer, and so the fielder has more time to get to them and make the catch, even if the ball is hit behind them. This was an interesting case where the data clearly illuminated a weakness in our prior reasoning."