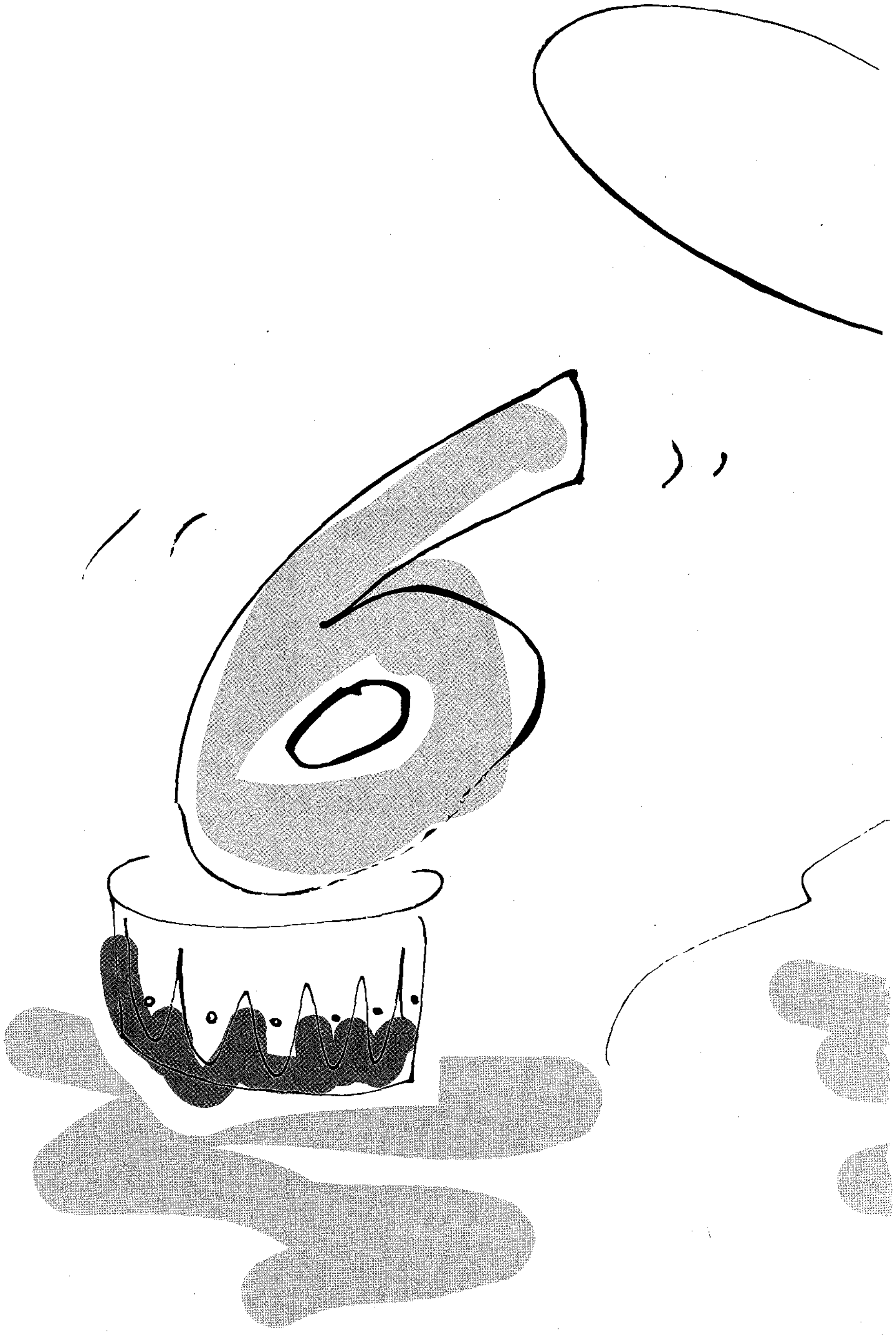
How Numbers Can Trick You

The statistics that fill the media are often subtly misleading. Here's a guide to the most common types of error.

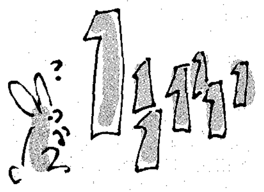
BY ARNOLD BARNETT



THINK tanks, government agencies, special-interest groups, and academics conduct myriad studies about health, safety, and the physical and social sciences. The popular press usually conveys the gist of these studies to the general public in terms of statistics that are meant to summarize the findings. Unfortunately, many such reports are compromised by errors in statistical reasoning. And while people have developed a healthy skepticism about advertisements that also appear in the media, the numbers in these paid-for messages can be even more distorted than we cynically imagine.

A substantial fraction of statistical misunderstandings fall into a half-dozen categories—the Six Deadly Sins of Statistical Misrepresentation. I offer examples of these errors below; while they are drawn mostly from criminology and aviation (domains with which I am particularly familiar), they have plenty of counterparts elsewhere. My hope is to help audiences of the popular media—that is, just about everybody—to detect difficulties often apparent only to those with independent information about the subject, and to discourage fellow citizens from taking a strong position or course of action based solely on a press report.

ILLUSTRATIONS
BY TAMAR HABER-SCHAIM



#1 GENERALIZING FROM NON-RANDOM SAMPLES

Statistics about unusual subpopulations are often interpreted as applying to an entire population. Such extrapolation can yield misleading and even ludicrous results.

ANGRY HEARTS

In March 1994, an Associated Press story in the *Boston Globe* reported of a study that had concluded that outbursts of anger "can double the chance for heart attack." In interviews with 1,500 people who had suffered heart attacks in the previous few days, researchers at Harvard Medical School found that a disproportionate number reported episodes of extreme anger in the two hours preceding the attack. A statistical analysis by the researchers led them to estimate that "anger was associated with 2.3 times the usual (heart attack) risk."

There's a problem here: the only contributors to the data analysis were people who had suffered heart attacks—and had survived them. Thus, the newspaper's implied advice to the broader population—"keep cool"—may have been misguided. Although the study indicated that vigorously "blowing off steam" seems to raise the *immediate* risk of heart attack, such releases of tension could serve to reduce the overall long-term risk. But people who had freely expressed anger throughout their lives—and who had, perhaps as a result, managed to live to old age without a heart attack—could never make it into the researchers' sample. It is also possible, though perhaps far-fetched, that those whose heart attacks were instigated by anger were better able to survive them than are other such victims. Were that the case, angry people could be overrepresented in the sample by virtue of their ability to survive a heart attack and thus become available for an interview.

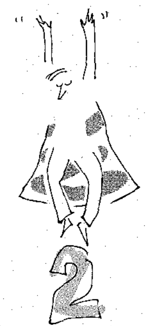
To illuminate the difficulty, let's look at a couple of examples—one real, the other hypothetical. If you looked at the age at death of deceased rock-and-roll stars (Buddy Holly, Jimi Hendrix, Janis Joplin, Jim Morrison, Kurt Cobain, et al.), you might superficially conclude that rock stars die about 40 years younger than the general population. This interpretation is invalid, though, because the sample is biased, systematically excluding those icons of rock who are still alive; for all we know, Mick Jagger might live to be 90. The same problem afflicts analyses of angry Americans—the analyses are restricted to those among them who *get* heart attacks.

More hypothetically, suppose that disease X, if untreated, is fatal 20 percent of the time. Now imagine that there is a widely used surgical procedure for this disease that kills 1 percent of the patients who undergo it but that cures the other 99 percent. Of those people whose deaths are attributable to disease X, an awful lot will have spent their final hours on the operating table. Viewed in isolation, this might suggest that the surgery is highly dangerous. But in neglecting to look at the 99 percent who were cured, this "last hour before death" analysis totally ignores the benefits of the procedure. In the same way, a study that limits its purview to known victims of heart attacks obscures any possible benefit to the heart of releasing tension through anger.

FREQUENT FLIERS

During the mid-1980s, Midway Airlines promoted its New York-to-Chicago service with ads in the *New York Times* and the *Wall Street Journal* claiming that "84 percent of frequent business travelers to Chicago prefer Midway Metrolink to American, United, and TWA." This figure seemed astounding because, at the time, Midway was carrying only 8 percent of the traffic between New York and Chicago. How could there be such a huge discrepancy between what people "prefer" and what they actually do?

The mystery was solved in the fine print at the bottom of the ad, which revealed that the relevant survey was "conducted among Midway Metrolink passengers between New York and Chicago." In other words, the only passengers eligible for inclusion in the survey were the 8 percent who were already flying Midway. To treat the sample as representative, one would have to make the startling assumption that Midway's popularity among those who fly it was the same as among those who don't. If there was any surprise at all in the results, it was that one in six Midway passengers apparently preferred to be flying on a different airline.



#2 LOOK! A TREND!

Journalists sometimes attach great importance to random data shifts that may already be irrelevant by the time they are reported. Admittedly, it's not always easy to distinguish a mere fluctuation from the start of a meaningful trend. The effort to do so is worth making, however, and in some cases pays off quickly.

ARNOLD BARNETT is professor of operations research at MIT's Sloan School of Management, where he specializes in applied probability and statistics. Both Business Week *and* Fortune *have cited Barnett as one of the most effective business-school teachers in the United States.*

Airline Safety

In 1993, the International Airline Passengers Association (IAPA) began rating airlines in terms of safety. IAPA focused on the decade ending in 1993 and rated airlines primarily on the basis of two ratios: fatal accidents per million flights performed, and passengers killed per million passengers carried. Among large U.S. jet carriers, American, Delta, and Southwest were classified as "outstanding," while Continental, Northwest, TWA, United, and USAir received the positive but clearly lesser designation "very good." Newspapers around the country reported IAPA's investigation.

But because fatal air accidents involving U.S. jets are exceedingly rare, even airlines with the same safety record over the long run can differ in safety performance over short spans. Indeed, if a ranking of carriers by safety reflects mere fluctuations, it should be highly changeable as the observation period varies. As the table below shows, this is indeed the case. The table ranks the eight large U.S. jet carriers by the death risk for a person who randomly chose one of the airline's flights during 10-year periods ending in 1983, 1988, and 1993. The lower the numbers, the fewer the fatalities. (Airlines with no deaths at all during a period are starred; these are ranked by number of flights performed.)

DEATH-RISK RANKING FOR 10-YEAR PERIOD ENDING....

AIRLINE	12/31/93	12/31/88	12/31/83
AMERICAN	1*	6	7
CONTINENTAL	4	4	5
DELTA	5	5	1*
NORTHWEST	7	7	2*
SOUTHWEST	2*	2*	4*
TWA	3	3	8
UNITED	6	1*	6
USAIR	8	8	3*

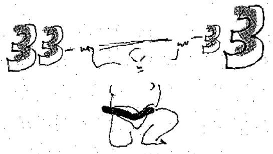
To put it delicately, the results cannot be characterized as stable. The first-ranked airline was different in all three periods and, strikingly, the airline that was best in one period always fell in the bottom half of the rankings in the other two. Southwest Airlines had a perfect record over all three periods but, because it had far fewer flights than the other carriers, was in a better position than they to avoid fatalities. The two airlines that were ranked lowest in the two most recent periods (Northwest and USAir) had no passenger deaths at all in the third. The mortality data, in short, provide a pitifully tenuous basis for putting these airlines into two distinct categories—a point that was overlooked both by IAPA's analysts and by the newspapers that publicized their results.

Foreigners in Florida

Last October, after a well-publicized murder of a German tourist in Florida, the whole world heard that in fact nine foreign tourists had been slain in that state during the preceding year. Fear of such violence has cost Florida hundreds of thousands of recent visitors, yet this response could well be an overreaction to statistical noise.

Even if homicides against foreign tourists in Florida occur at a low, constant rate over time, there are bound to be some periods when the rare events bunch together, much as there will be other periods when none occur at all. Suppose, for example, that over many years there is on average a 1 percent chance each day that a foreign tourist will be murdered somewhere in Florida. Such killings will average 3.65 per year (365 x .01), and the average interval between successive killings will be 100 days—long enough, presumably, to dispel inclinations to speak of a trend. But probabilistic calculations (not included here) also show that, over a full decade, the chance is nearly 3 in 10 that there will be *some* 12-month period with 9 or more killings; over a 20-year period, the chance of such a bloody stretch rises to roughly 1 in 2.

In the six months following October 1993, the press fell silent on the subject of murders of foreigners in Florida. Conceivably, a menacing trend was reversed because of sensible measures it provoked, such as the elimination of visible evidence that a car is rented. But it is also quite possible that there was no real trend to reverse, and that the pattern no more signaled heightened danger to foreign tourists than a year without murders would have signaled a future free of risk.



#3 UNJUST LAW OF "AVERAGES"

Summary statistics about two large sets of data can invite conclusions that would not stand if the sets were examined individually, in greater detail. Comparisons of overall averages can yield particularly distorted impressions.

Underpaid Doctors?

U.S. News and World Report told us in 1983 that U.S. physicians were "growing in number but not in pay." Its chart showed that between 1970 and 1982, the number of doctors jumped from 334,000 to 480,000, but their average salary (in 1982 dollars) dropped from $103,900 to $99,950. The magazine seemed to discern in these statistics yet another application of the law of supply and demand: a relative abundance of doctors was lowering the market value of individual practitioners.

But some arithmetic raises doubts that the market's "invisible hand" was responsible for this sag. It seems reasonable to assume that perhaps 25 percent of the doctors practicing in 1970 (some 83,500) had retired by 1982, leaving about 250,000 at work. This means that roughly half the 480,000 doctors working in 1982 had begun practicing during the last 12 years. Because of this large influx, the typical physician in 1982 was probably younger than his or her 1970 counterpart. And since salaries tend to increase with age, the decline the magazine saw might well have reflected a downward shift in the age distribution among doctors rather than reduced compensation at any given age.

In fact, it is possible that the salaries of doctors in every age group actually went *up* during the period 1970-82, but that a dramatic downward trend in the age profile of physicians overall overshadowed this rise and pushed down the profession's average pay. Indeed, the minimal size of the reported drop in salary (4 percent) suggests that an age-by-age comparison might well have shown that doctors' annual pay was rising along with their numbers.

"On-Time" Airlines

In 1987, the Department of Transportation required U.S. airlines to report each month the percentage of their flights into the nation's 30 busiest airports that arrived on time. Major newspapers published these statistics, at least until the novelty wore off, and the airlines that ranked high on promptness took to stressing that point in their ads. Northwest still boasts that it is "the number one on-time airline."

Each airline's on-time score depends on its performance ratings at the 30 individual airports, but the airports the airliners serve frequently have greater effect than those it serves rarely. The averages thus naturally favor an airline that mostly flies in and out of fair-weather airports over those airlines that serve cities frequently socked in by rain or fog.

For example, America West Airlines routinely outperforms Alaska Airlines in overall on-time performance, but on further inspection this victory seems hollow. Alaska serves only five of the thirty busiest airports and, as we can see from the following table, it was prompter than America West in June 1991 at all five. But if one computes the *average* performance for flights into those five airports, America West receives a better rating. This counterintuitive result arises because a large majority (73 percent) of America West's flights into these five airports arrive at desert-sun Phoenix. Thus, America West's 92.1 percent on-time record at Phoenix dominates its five-airport statistic. Alaska Airlines scored even better in Phoenix than America West did (94.8 percent on time), but because only 6 percent of Alaska Airlines's flights go into or out of Phoenix, this result has little effect on its five-city average. By contrast, 57 percent of Alaska's flights arrive at Seattle—one of the moody weather capitals of the world—as opposed to only 4 percent of America West's. In the five-city average, in other words, America West gets to put its best foot forward and bury one of its weakest scores; Alaska Airlines is forced into the opposite position.

	Alaska Airlines		America West Airlines	
Destination	% Arrivals on Time	No. of Arrivals	% Arrivals on Time	No. of Arrivals
Los Angeles	88.9%	559	85.6%	811
Phoenix	94.8	233	92.1	5,255
San Diego	91.4	232	85.5	448
San Francisco	83.1	605	71.3	449
Seattle	85.8	2,146	76.7	262
5-Airport Total	86.7%	3,775	89.1%	7,225

Fundamental misunderstandings of statistical results can arise when two words or phrases are unwisely viewed as synonyms, or when an analyst applies a particular term inconsistently.

The Odds of Execution

A powerful example of the first problem arose in 1987, when the U.S. Supreme Court issued its controversial *McClesky v. Kemp* ruling concerning racial discrimination in the imposition of the death penalty. The Court was presented with an extensive study of Georgia death sentencing, the main finding of which was explained by the *New York Times* as follows: "Other things being as equal as statisticians can make them, someone who killed a white person in Georgia was four times as likely to receive a death sentence as someone who killed a black."

The Supreme Court understood the study the same way. Its majority opinion noted that "even after taking account of 39 nonracial variables, defendants charged with killing white victims were 4.3 times as likely to receive a death sentence as defendants charged with killing blacks."

But the Supreme Court, the *New York Times*, and countless other newspapers and commentators were laboring under a major misconception. In fact, the statistical study in *McClesky v. Kemp* never reached the "factor of four" conclusion so widely attributed to it. What the analyst did conclude was that the *odds* of a death sentence in a white-victim case were 4.3 times the odds in a black-victim case. The difference between "likelihood" and "odds" (defined as the likelihood that an event will happen divided by the likelihood that it will not) might seem like a semantic quibble, but it is of major importance in understanding the results.

The likelihood, or probability, of drawing a diamond from a deck of cards, for instance, is 1 in 4, or 0.25. The odds are, by definition, 0.25/0.75, or 0.33. Now consider the likelihood of drawing any red card (heart or diamond) from the deck. This probability is 0.5, which corresponds to an odds ratio of 0.5/0.5, or 1.0. In other words, a doubling of probability from 0.25 to 0.5 results in a tripling of the odds.

The death penalty analysis suffered from a similar, but much more serious, distortion. Consider an extremely aggravated homicide, such as the torture and killing of a kidnapped stranger by a prison escapee. Represent as *PW* the probability that a guilty defendant would be sentenced to death if the victim were white, and as *PB* the probability that the defendant would receive the death sentence if the victim were black. Under the "4.3 times as likely" interpretation of the study, the two values would be related by the equation:

$$PW = 4.3\,PB$$

If, in this extreme killing, the probability of a death sentence is very high, such that *PW* = 0.99 (that is, 99 percent), then it would follow that *PB* = 0.99/4.3 = 0.23. In other words, even the hideous murder of a black would be unlikely to evoke a death sentence. Such a disparity would rightly be considered extremely troubling.

But under the "4.3 times the odds" rule that reflects the study's actual findings, the discrepancy between *PW* and *PB* would be far less alarming. This yields the equation:

$$\frac{PW}{1-PW} = 4.3\left[\frac{PB}{1-PB}\right]$$

If PW = 0.99, the odds ratio in a white-victim case is 0.99/0.01; in other words, a death sentence is 99 times as likely as the alternative. But even after being cut by a factor of 4.3, the odds ratio in the case of a black victim would take the revised value of 99/4.3 = 23, meaning that the perpetrator would be 23 times as likely as not to be sentenced to death. That is:

$$\frac{PB}{1-PB} = 23$$

Work out the algebra and you find that *PB* = 0.96. In other words, while a death sentence is almost inevitable when the murder victim is white, it is also so when the victim is black—a result that few readers of the "four times as likely" statistic would infer. While not all Georgia killings are so aggravated that *PW* = 0.99, the quoted study found that the heavy majority of capital verdicts came up in circumstances when *PW*, and thus *PB*, is very high.

None of this is to deny that there is some evidence of race-of-victim disparity in sentencing. The point is that the improper interchange of two apparently similar words greatly exaggerated the general understanding of the degree of disparity. Blame for the confusion should presumably be shared by the judges and the journalists who made the mistake and the researchers who did too little to prevent it.

(Despite its uncritical acceptance of an overstated racial disparity, the Supreme Court's *McClesky v. Kemp* decision upheld Georgia's death penalty. The court concluded that a defendant must show race prejudice in his or her own case to have the death sentence countermanded as discriminatory.)

THE SKYLAB IS FALLING! THE SKYLAB IS FALLING!

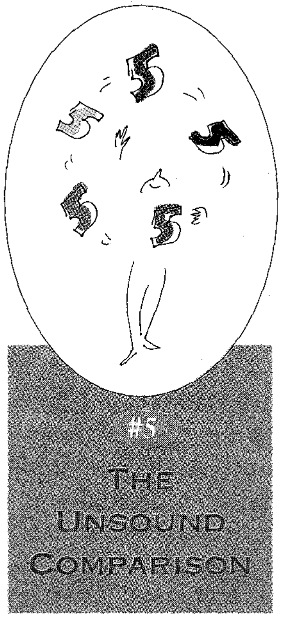
In 1979, the National Aeronautics and Space Administration (NASA) decided to warn people that the Skylab space station had dropped out of orbit and was headed toward the earth, where its debris could scatter on a populated area. To make its announcement less frightening, NASA administrator Robert A. Frosch offered an accompanying risk assessment widely repeated in the press:

1) The probability that falling debris from the Skylab will hit someone on earth—anyone at all—is 1 in 150; and

2) Because there are 4 billion people on the planet, the chance that any given person will be hit by Skylab debris is (1/150) x (1/4 billion), or 1 in 600 billion—in other words, negligible.

But NASA's risk description was ambiguous. What does it mean to say that there is a 1 in 150 chance that "someone" will be hit by debris? Clearly, the implication is that there is a 149 in 150 chance that no one will be hit, but how many people are hit given that "someone" is? The answer to that question is crucial to determining the level of individual risk. If the number of people struck by Skylab debris cannot exceed 1, then (and only then) does an individual have a 1 in 4 billion chance of being victimized, given that someone is struck. But why is it certain that debris could hit at most 1 person? If Skylab landed on a crowded bus or a busy marketplace (or, much worse, on a 747 seven miles above the earth), dozens or even hundreds of people could be simultaneously injured or killed. NASA's estimate completely ruled out such events.

Fortunately, the debris fell harmlessly in a remote part of Australia. But the lesson is that an elusive word like "someone" is not useful in describing an event. When a word can be construed in different ways, the reader and even the data analyst can unintentionally jump from one interpretation to another, as presumably NASA did when it first equated "someone" to "at least one" but then shifted to "exactly one" in the middle of its calculation.



#5 THE UNSOUND COMPARISON

Press accounts of scientific studies sometimes invite readers to reach conclusions by comparing a reported statistic with some other that supposedly represents a natural baseline. But the proposed baseline may be anything but natural.

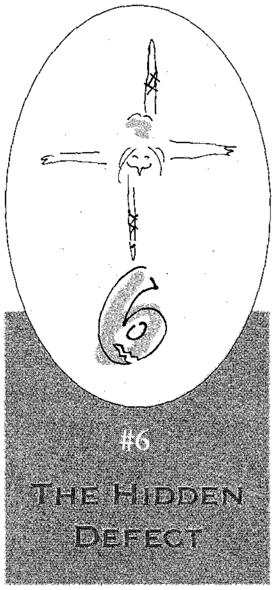
MURDER CITIES

Early in 1992, the *New York Times* reported that record numbers of killings occurred the previous year in four of the nation's ten largest cities: Los Angeles, San Diego, Dallas, and Phoenix. The implication was that even one all-time high among such cities was unusual, let alone four. The report failed to point out, however, that all four of these cities also reached new highs in population in 1991; thus, even if their per capita murder rates had not changed since Cain slew Abel, their absolute 1991 murder tolls would have set new records. Indeed, because six cities in the top ten set population records in 1991, the fact that only four of them set new highs in numbers of homicides might in itself be reassuring.

SHARING THE WEALTH?

In 1982, *National Review* magazine announced that it had "good news for all egalitarians and redistributionists." Apparently, the Internal Revenue Service had found even more effective ways to soak the rich. The principal piece of evidence for this conclusion was that in 1980, the top 10 percent of U.S. earners paid 52 percent of all federal income tax, up from 49 percent five years earlier. The article encouraged the reader to assume that had not the IRS tinkered with its tax codes, the share of taxes the wealthiest 10 percent paid in 1980 would have remained at its 1975 level.

But the magazine advanced this point with incomplete evidence. In fact, it was perfectly conceivable that the top 10 percent was paying a growing share of the nation's taxes simply because this group's share of the nation's income was going up. In this particular case, the faulty assumption was not fatal because the unchecked data about earnings among the wealthy supported the story's claim: the share of income amassed by the wealthiest 10 percent of Americans changed very little from 1975 to 1980. But the glib comparison between the two years was unsound, and invoking the same "top tenth" argument for the 1980s—when the Census Bureau reports that U.S. income inequality did indeed rise sharply—would produce a quite misleading result.



#6 THE HIDDEN DEFECT

Questionable analyses in the first five categories can be spotted by anyone with a little knowledge of statistical methods. A more insidious type of misinformation unravels only when a reader probes the numbers and looks at their source.

HOME PREGNANCY TESTING

An advertisement for a home pregnancy test boasted that, under laboratory conditions, the test was 99.5 percent accurate. There is no reason to disbelieve that statistic, although a single error rate seems unnecessarily vague. There are, after all, two kinds of error such a test can make: it can tell a pregnant woman that she is not and it can tell a woman not pregnant that she is.

A brochure put out separately by the manufacturer of this test kit, however, was extremely disturbing. It showed that the 99.5 percent estimate was based on data summarized in the table that follows. The table does indicate only 1 error in 200 assessments, but it raises

	ACTUALLY PREGNANT	ACTUALLY NOT PREGNANT
TEST SAYS PREGNANT	197	0
TEST SAYS NOT PREGNANT	1	2
TOTAL	198	2

two questions. Why were 99 percent of the women tested—198 out of 200—pregnant? And, even more strangely, why was the accuracy of the test for nonpregnant women estimated from a sample size of *two*?

Things got worse as the brochure went on. The 2-for-2 accuracy statistic about nonpregnant women was based on an analysis of the test results by laboratory technicians. But the main advantage of a home pregnancy test is that women can use it themselves. The brochure took account of this issue by reporting what happened when the women interpreted results on their own: of 101 such women who were not pregnant, 8 mistakenly concluded that they were.

In other words, the manufacturer had two accuracy results about nonpregnant women. One, based on a (presumably) representative sample of the product's users, showed an error rate of 8 percent in 101 trials. The other, based on a "sample" of laboratory technicians, obtained a 0 percent error rate over 2 trials. In its advertising, the manufacturer applied the 0 percent rate in the small expert sample and ignored the 8 percent rate in the large, unbiased one.

Safe Traveling

It is widely known that auto-safety statistics are grim while air safety data are greatly reassuring. Yet many people feel safer driving than flying, largely because they think themselves such good drivers that the fatality rates don't really apply to them. Such flattering self-assessments received apparent support from a 1991 study described in the journal *Risk Analysis*. The study's primary finding (which was reported in both the *Wall Street Journal* and the *Washington Post)* was that a prototypical safe U.S. driver—age 40, sober, belted into the seat of a heavier-than-average car—actually suffers "slightly less" mortality risk on a 600-mile trip than a person who takes the same trip by air.

The analysis began with the overall death rate per mile driven on rural interstate highways, the main thoroughfares for intercity auto trips. The researchers then revised this initial risk estimate using multipliers that reflected various characteristics of cars and drivers. Having a heavier-than-average car multiplied the risk estimate by 0.77 (that is, reduced it by 23 percent), while having a 40-year-old driver multiplied the estimate by 0.68. The final risk factor for a particular combination of factors was the product of the individual adjustments.

Unfortunately for those who prefer to drive, this analysis greatly exaggerates the safety of driving because the risk-reduction factors are not truly independent: Part of the *reason* 40-year-olds die less frequently in car crashes than 18-year-olds is that the middle-aged motorists tend to drive heavier cars, wear seat belts, and stay off the road when intoxicated. Taking credit for each of these factors separately, as the study did, amounts to quadruple-counting and greatly overstates the safety of driving versus flying.

The study exacerbated this error by failing to distinguish between the safety records of different types of aircraft. In their risk calculations for 600-mile flights, the researchers worked with merged accident data for all types of aircraft. But a flight of 600 miles is almost always performed by a jet, and jets have far better safety records than propeller planes. The peculiar approximations of this study led it to conclude that the mortality risk from driving 600 miles was comparable to that of flying 600 miles. A more fair and logical analysis would show that flying is safer by a factor of at least five.

Toward Statistical Literacy

When Miss Marple, Agatha Christie's famous detective, was asked why she always believed the worst about human nature, she responded that "the worst is so often true." Similarly, statistical reports in the media involve flaws regularly enough that some initial skepticism is well deserved.

The most cautious general course for the reader is to treat such reports more as public announcements that studies have been done than as clear guides to their content or reliability. Readers might not only look for evidence that researchers, reporters, or advertisers have committed one or more of the six deadly sins but also cultivate a general awareness that statistics can yield highly divergent interpretations. When a particular interpretation of the reported data pattern is advanced, have the analysts reasonably excluded other possibilities, or failed even to recognize them?

Ultimately, should the conclusions really matter to the reader, then there is no avoiding the arduous task of finding the study and reading it. And contacting the author for further details is both wise and legitimate.

For the alert individual, statistical humbug should be no harder to ferret out than other forms of illogical argument. It just takes practice and time.■