Remember the old saw that figures don’t lie, but liars do figure? Joel Best, the author of *Damned Lies and Statistics: Untangling Numbers from the Media, Politicians and Activists* (University of California Press), would probably not be so kind to the figures. Indeed, Best, who teaches sociology and criminal justice at the University of Delaware, plainly spends a good portion of his time at the office teasing elusive truths from the witches’ brew of numbers that pervade debates over public policy. This witty, well-written excerpt – from a chapter of the book titled “mutant statistics” – is apt to leave the reader more depressed than angry. For much of the difficulty with numbers discussed therein is a product of ignorance rather than of intent. Read it and weep. Or, at the very least, learn to view statistics and the nice folks who peddle them as proof of this or that with deepest skepticism.

© University of California Press, reprinted with permission.
Not all statistics start out bad, but any statistic can be made worse. Numbers – even good numbers – can be misunderstood or misinterpreted. Their meanings can be stretched, twisted, distorted or mangled. These alterations create what we can call “mutant statistics” – distorted versions of the original figures.

Many mutant statistics have their roots in innumeracy – difficulty grasping the meaning of numbers and calculations – which is widespread. Not only is much of the general public innumerate, but also the advocates promoting social and economic problems are often not any better. They may become confused about a number’s precise meaning. They may misunderstand how a problem has been defined, how it has been measured, or what sort of sampling method has been used. At the same time, their commitment to their cause (“After all, it’s a big problem!”) may lead them to improve a statistic, to make the numbers seem more dramatic, more compelling. Some mutant statistics may be products of advocates’ cynicism, of their deliberate attempts to distort information in order to make their claims more convincing. This seems particularly likely when mutation occurs at the hands of large institutions that twist information into the form most favorable to their interests. But mutation can also be a product of sincere, albeit muddled, interpretations by innumerate advocates.

Once someone utters a mutant statistic, there’s a good chance that those who hear it will accept it and repeat it. Innumerate advocates influence their audiences: the media repeat mutant statistics and the public accepts – or does not challenge – whatever numbers the media present. A respected commentator may hear a statistic and repeat it, making the number seem even more credible.

As statistics gain wide circulation, number laundering occurs. The figures become harder to challenge because everyone has heard them and everyone assumes they must be correct. This is especially true when numbers reinforce our beliefs or interests. (“Of course that’s true!”)

Consider one widely circulated statistic about the dangers of anorexia nervosa. Anorexia usually occurs in young women, and some feminists argue that it is a response to societal pressures for women to be beautiful and to cultural standards that equate slenderness with beauty. Advocates who were seeking to draw attention to the problem estimated that 150,000 American women were anorexic, and noted that anorexia could be fatal.

At some point, some feminists began reporting that each year 150,000 women died from anorexia. (In fact, only about 70 deaths a year are attributed to anorexia.) This simple transformation – turning an estimate for the total number of anorexic women into the annual number of fatalities – produced a dramatic, memorable statistic. Advocates repeated the erroneous figure in books, in newspaper columns, on talk shows and so on. There were soon numerous sources for the mistaken number. A student searching for material for a term paper on anorexia, for instance, had a good chance of encountering – and repeating
– this wildly inaccurate statistic, and each repetition helped ensure that the mutant statistic would live on.

Yet it should have been obvious that something was wrong with this figure. Anorexia typically affects young women. In the United States, roughly 8,500 females 15 to 24 years old die from all causes each year; another 47,000 women aged 25 to 44 also die annually. What were the chances, then, that there could be 150,000 deaths from anorexia each year?

But, of course, most of us have no idea how many young women die each year. ("It must be a lot.") When we hear that anorexia kills 150,000 young women a year, we assume that whoever cites the number must know that it is true.

**Let me count the ways**

How and why does mutation occur? Here, I explore four common ways of creating mutant numbers, beginning with the most basic error – making inappropriate generalizations from a statistic. I then turn to transformations – taking a number that means one thing and interpreting it to mean something completely different. Next, I explore confusion – transformations that involve misunderstanding the meaning of more complicated statistics. Finally, I consider compound errors – the ways in which bad statistics can be linked to form chains of error. In these four ways, bad statistics not only take on lives of their own, but they do increasing damage as they persist.

**Generalization:**

**Elementary Forms of Error**

Generalization is an essential step in statistical reasoning. We rarely are able to count all the cases of some social problem. Instead, we collect some evidence from a sample and generalize from it. Generalization involves some basic processes: the problem must be defined, and a means of measurement and a sample must be chosen. These are elementary steps in social research. But even the most basic principles can be violated and, surprisingly often, no one notices when this happens. Mutant statistics – based on flawed definitions, poor measurements or bad samples – emerge, and often receive a surprising amount of attention.

**Questionable Definitions.** Consider the flurry of media coverage about the so-called epidemic of fires in African-American churches in the South in 1996. Various groups charged that the fires were the work of a racist conspiracy. Their claims recalled the history of racial terrorism in the South: black church-
The Milken Institute Review

es had often been targets of arson or bombing. Perhaps because 1996 was an election year, politicians – both Democrats (including President Clinton and Vice President Al Gore) and Republicans – denounced the fires, as did both the liberal National Council of Churches and the conservative Christian Coalition.

Activists (such as the anti-racist Center for Democratic Renewal) tried to document the increased number of fires, producing lists of church arsons and statistics about the number of suspicious fires as evidence that the problem was serious. However, investigations, first by journalists and later by a federal task force, called those claims into question.

While there were certainly some instances in which whites burned black churches out of racist motives, there was no evidence that a conspiracy linked the fires. Moreover, the definition of what was a racially motivated church fire proved to be unclear; the activists’ lists included fires at churches with mostly white congregations, fires known to have been set by blacks, teenage vandals or mentally disturbed people, and fires set to collect insurance.

When journalists checked the records of the insurance industry, they discovered not only that the number of fires in 1996 was not unusually high, but also that church arsons had been generally declining since at least 1980. The federal task force ultimately failed to find any evidence of either an epidemic of fires or a conspiracy.

In short, statistics attempting to demonstrate the existence of an epidemic of church arsons lacked a clear definition of what ought to count as a racially motivated church fire. By the same token, those who claimed there was an epidemic did not define the number of fires it would take to constitute one. Indeed, the absence of any clear definitions made it difficult to assess the evidence. The advocates who offered lists of fires (and asserted that each was evidence of a racist conspiracy) may have been convinced, but those who tried to identify cases using some sort of clear definition failed to find any evidence of an epidemic.

**Inadequate Measurement.** Clear, precise definitions are not enough. Whatever is defined must also be measured, and meaningless measurements will produce meaningless statistics. For instance, consider recent federal efforts to count hate crimes – crimes motivated by racial, religious or other prejudice.

In response to growing concern, the Federal Bureau of Investigation invited local law enforcement agencies to submit annual reports on hate crimes within their jurisdictions and, beginning in 1991, the bureau began issuing national hate-crime statistics.

Although the FBI had collected data on the incidence of crime from local agencies for decades, counting hate crimes posed special problems. When police record a reported crime – say, a robbery – it is a relatively straightforward process. Usually the victim comes forward and tells of being forced to surrender money to the robber. But identifying a hate crime requires something more: an assessment of the criminal’s motive. A robbery might be a hate crime if prejudice motivates the robber, but the crimes committed by robbers with other motives are plainly not hate crimes, even if the robber and the victim are of different races.

There are real disagreements about how to define and measure hate crimes. Not surprisingly, some activists favor broad, inclusive standards that will avoid false negatives. Some feminists, for example, argue that rapes should automatically be considered hate crimes on the grounds that all rape is motivated by gender prejudice. But local officials, who may be reluctant to publicize tensions
within their communities, may favor much narrower standards: a cross-burning on an African-American family's lawn may be classified as a teenage prank rather than as a hate crime.

Because there is much variation in how – and even whether – agencies measure hate crimes, hate-crime statistics have been incomplete and uneven. In 1991, the FBI collected hate-crime data from only 32 states; less than a quarter of all law enforcement agencies supplied reports. By 1996, 49 states and the District of Columbia reported some data, but many agencies still did not participate.

More important, many of the agencies that did file reports indicated that they had recorded no hate crimes during 1996. Twelve states reported fewer than 10 hate crimes apiece and Alabama's law enforcement agencies reported none. So long as many agencies refused to submit hate-crime statistics – and others used wildly different standards to classify hate crimes – the data collected and published would have little value. We might even suspect that the jurisdictions that report the most hate crimes will be those with the most liberal governments, because they are more likely to press law enforcement agencies to take such reporting seriously. This suggests that hate-crime statistics may be a better measure of local officials' political beliefs than of the incidence of hate crimes.

While the record-keeping may improve over time, the hate-crime statistics reported during the program's early years were nearly worthless. Practices for recording hate crimes obviously varied widely among jurisdictions, making meaningful comparisons impossible. Moreover, it should be noted that, as reporting does improve, the numbers of reported hate crimes will almost certainly increase. That is, incidents that previously would not have been counted as hate crimes will be counted and successive annual reports will show the incidence of hate crime rising. Measurement is always important, but this example illustrates why new statistical measures should be handled with special caution.

**Clear, precise definitions are not enough.** Whatever is defined must also be measured, and meaningless measurements will produce meaningless statistics.

agencies supplied reports. By 1996, 49 states and the District of Columbia reported some data, but many agencies still did not participate.

More important, many of the agencies that did file reports indicated that they had recorded no hate crimes during 1996. Twelve states reported fewer than 10 hate crimes apiece and Alabama's law enforcement agencies reported none. So long as many agencies refused to submit hate-crime statistics – and others used wildly different standards to classify hate crimes – the data collected and published would have little value. We might even suspect that the jurisdictions that report the most hate crimes will be those with the most liberal governments, because they are more likely to press law enforcement agencies to take such reporting seriously. This suggests that hate-crime statistics may be a better measure of local officials’ political beliefs than of the incidence of hate crimes.

While the record-keeping may improve over time, the hate-crime statistics reported during the program's early years were nearly worthless. Practices for recording hate crimes obviously varied widely among jurisdictions, making meaningful comparisons impossible. Moreover, it should be noted that, as reporting does improve, the numbers of reported hate crimes will almost certainly increase. That is, incidents that previously would not have been counted as hate crimes will be counted and successive annual reports will show the incidence of hate crime rising. Measurement is always important, but this example illustrates why new statistical measures should be handled with special caution.

**Bad Samples.** Earlier, I emphasized the importance of generalizing from representative samples. This is a basic principle, but one that is easily ignored.

Consider a study subtitled *A Survey of 917,410 Images, Descriptions, Short Stories, and Animations Downloaded 8.5 Million*
Times by Consumers in Over 2,000 Cities in 40 Countries, Provinces, and Territories. An undergraduate published this research in 1995 in a law review, reporting that 83.5 percent of the downloaded images were pornographic. In 1995, the Internet was still a novel phenomenon; people worried that children were frequent users, and that parents did not understand the Internet well enough to protect their children from questionable content.

An un-dergraduate published this research in 1995 in a law review, reporting that 83.5 percent of the downloaded images were pornographic. In 1995, the Internet was still a novel phenomenon; people worried that children were frequent users, and that parents did not understand the Internet well enough to protect their children from questionable content.

Claims that an extensive research project revealed that a substantial majority of Internet traffic involved pornography generated considerable concern. The huge scope of the study – 917,410 images downloaded 8.5 million times – implied that it must have been exhaustive.

But, of course, a large sample is not necessarily a good sample. In this case, the researcher did not collect a representative sample of Internet traffic. Rather, he examined postings to only 17 of some 32 Usenet groups that carried image files. In fact, his findings showed that pornographic images accounted for about 3 percent of Usenet traffic, while Usenet accounted for only about an eighth of the traffic on the Internet.

The sample was drawn from precisely that portion of the Internet where pornographic images were concentrated and was thus anything but representative. An alternative way to summarize the study’s findings was that only one-half of 1 percent of Internet traffic involved pornographic images – a markedly lower figure than 83.5 percent.

All three cases discussed above received extensive coverage from the media; all three attracted the attention of political leaders; all three involved mutant statistics. It is also true that, in all three cases, the statistics eventually drew criticism. However, critics are not always successful in influencing the public. Many people probably remain convinced that most Internet traffic is pornographic, that members of a racist conspiracy set many church fires, and so on. The impact of mutant statistics is often long-lived.

Transformation: Changing the Meaning of Statistics

Another common statistic mutation involves transforming a number’s meaning. Usually, this takes place when someone tries to repeat a number, but manages to say something different; recall that 150,000 people with anorexia became 150,000 annual deaths from anorexia.
Of course, not all transformations are as obvious as equating having a disease with dying from it. Often transformations involve more subtle misunderstandings or logical leaps.

Consider the evolution of one social commentator’s estimate that 6 percent of America’s 52,000 Roman Catholic priests “are at some point in their adult lives sexually preoccupied with minors.” This estimate originated with a psychologist and former priest who treated disturbed clergymen and derived the figure from his observations. It was, in short, an educated guess. Still, his claim was often repeated and, in the process, transformed in at least four important ways.

First, some of those who repeated the figure forgot that it was an estimate, and referred to the number as though it were a well-established fact – presumably a finding from a survey of priests. Second, while the psychologist’s estimate was based on a sample of priests who had sought psychological treatment (and therefore might well be especially likely to have experienced inappropriate attractions to young people), he generalized to all priests. Third, although the original estimate referred to sexual attraction rather than actual behavior, those who repeated the number often suggested that 6 percent of all priests had had sexual contacts with young people. Fourth, those young people became redefined as children; commentators charged that 6 percent of priests were pedophiles. Although the original estimate in fact suggested that twice as many priests were attracted to adolescents as to younger children, this subtlety was lost.

Thus, an estimate that perhaps 6 percent of priests in treatment were at some point sexually attracted to young people was transformed into the so-called fact that 6 percent of all priests had had sex with children. Not everyone who repeated the statistic made all four transformations, but the number’s original meaning soon became lost in a chorus of claims linking “pedophile priests” to the 6 percent figure.

This example suggests that it is impossible to predict all the ways a number might be misunderstood and given an entirely new meaning. While it may be especially easy to transform estimates and guesses because the language of guessing is often vague, more precisely defined statistics can also undergo transformation.

Homicide statistics offer an example. In addition to gathering reports of homicides to calculate crime rates, the FBI also tries to collect more detailed information for its supplementary homicide reports. The FBI encourages law enforcement agencies to complete a brief form for each homicide that asks, for example, about the victim’s age, gender and ethnicity, the relationship between victim and offender, and the circumstances of the homicide – whether the death occurred during the course of a robbery, during an argument, and so on.

These reports are inevitably incomplete. When the police find the body of a homicide victim, they ordinarily can identify the victim’s age and gender, and they often – but not always – can specify the circumstances of the homicide. But unless they identify the offender, they usually cannot know the nature of the victim-offender relationship. In such cases, the relationship is coded “unknown.” Roughly 15 to 20 percent of the reports to the FBI reports list the circumstances as unknown; nearly 40 percent indicate that the victim-offender relationship is unknown.

Completing the paperwork for the FBI is a byproduct of police work and may receive a relatively low priority. Agencies are supposed to submit these reports within five days of the
end of the month in which the homicide becomes known. While the FBI asks for updated reports when additional information becomes available, many agencies don’t bother to report changes. Thus, a homicide initially reported as involving unknown circumstances or an unknown victim-offender relationship may later be solved, but the police do not necessarily report what more they have learned to the FBI.

A classification of “unknown” means just that – at the time the report was completed, the police didn’t have the information. However, people sometimes make assumptions about the nature of the unknown circumstances or unknown victim-offender relationships reported to the FBI. In the early 1980s, the FBI drew attention to the problem of serial murderers. There had been several prominent serial murder cases in the news, and the press argued that this was, if not a new crime problem, at least one that was more common than ever before. The FBI estimated that there might be as many as 35 serial murderers active at any one time, and the media claimed that serial murderers might account for as many as 4,000 or 5,000 deaths per year. Some commentators mangled these numbers further, reporting that there were 4,000 to 5,000 active serial killers.

It should have been apparent that there was something wrong with these statistics. They implied that each killer murdered more than 100 victims per year – an improbably high average. How did the analysts arrive at the figure of 4,000 victims? Simple: they assumed that all – or at least a large share – of the reported homicides involving unknown circumstances or an unknown victim-offender relationship were serial murders. Serial murderers often kill victims unknown to them; therefore, those inclined to sensationalize the numbers assumed that cases in which the victim-offender relationship was unknown were probably serial murders.

Recent claims blaming most homicides on strangers use similar logic. Reports to the FBI classify about 15 percent of victims and offenders as strangers, but nearly 40 percent of victim-offender relationships as unknown. Some interpretations assumed that any unknown relationship must involve strangers. So they added 15 percent and 40 percent and concluded that strangers commit most (55 percent) of all murders.

Both the serial murder and the stranger-homicide claims transformed the meaning of “unknown” by assuming that, if the police can’t classify the victim-offender relationship, then the homicide must be the work of a stranger – or even a serial killer. This is an unwarranted logical leap. Researchers who have conducted more careful studies (for example, examining officials’ final classifications for homicides) have concluded that strangers account for only 20 to 25 percent of all homicides, not more than half, and that serial murderers kill perhaps 400 victims a year, not 4,000.

The lesson from the misinterpretation of these statistics, then, concerns transformations created by careless inferences about the meaning of official statistics. In these cases, observers assumed that they knew what had actually happened in cases that the police labeled “unknown.” They produced dramatic, frightening figures that exaggerated the deaths caused by strangers or serial murderers, and those transformed figures were widely circulated.

Transformations involve shifts in meaning; advocates for some cause convert a statistic about X into a statistic about Y. This is an obvious error. Sometimes transformations are inadvertent, reflecting nothing more than sloppy language. In such cases, people try to
repeat a statistic, but they accidentally reword a claim in a way that creates a whole new meaning. Of course, other transformations may be deliberate efforts to mislead in order to advance the advocates’ cause.

Certainly, transformations often increase a claim’s impact by making it more dramatic: the number of anorexics becomes a body count; priests attracted to adolescents become priests having sex with children; homicides of unknown circumstances become serial killings. Such statistics get repeated precisely because they are dramatic, compelling numbers. A transformation that makes a statistic seem less dramatic is likely to be forgotten, but a more dramatic number stands a good chance of being repeated. It is a statistical version of Gresham’s Law: bad statistics drive out good ones.

There is another lesson here: transformation errors often reflect innumeracy. It should have been obvious that anorexia could not kill 150,000 women per year. Similarly, the people who asserted that there were 35 active serial murderers killing 4,000 victims each year should have realized that those two figures made no sense when combined, that both could not be correct.

Advocates are not the only ones to blame. The reporters who wrote stories about all those deaths from anorexia or serial murder should have asked themselves whether those numbers were plausible; they might even have investigated the claims before repeating them. Yet, in each case, these numbers received wide circulation for years – and continued to be repeated even after they were called into question. After all, the mutant statistics were now readily available; people could easily find them on the Internet or printed in many sources.

Transformation errors are thus easy to make, but difficult to set right. Transformation only requires that one person with influence over the media misunderstand a statistic and repeat the number in a way that gives it a new meaning. Even if someone recognizes the mistake and calls attention to it, the error is likely to live on.

Confusion: Garbling Complex Statistics

The examples discussed so far involve misunderstanding simple, straightforward statistics. But some statistics get mangled because they are difficult to grasp, and are therefore easily confused.

Consider Work Force 2000, a 1987 report commissioned by the United States Department of Labor that projected changes in the

Some statistics get mangled because they are difficult to grasp, and are therefore easily confused.
American work force. Work-force demographics are gradually changing for several reasons; most important, a growing proportion of women work, so females account for a growing percentage of workers. In addition, the percentage of workers who are nonwhite is growing, a fact that reflects several developments, including immigration patterns and ethnic differences in birth rates. The combined effect of these changes is gradually reducing the proportion of white males in the work force. In 1988, when *Work Force 2000* appeared, white males accounted for 47.9 percent of all workers, and the report projected that, by 2000, this percentage would fall to 44.8 percent.

However, rather than describing the change in such easily understood terms, the report’s authors chose to speak of “net additions to the work force.” The report predicted the populations of workers that would enter and leave the work force (because of death, retirement and so on) between 1988 and 2000. For example, the authors estimated that 13.5 million white males would join the work force and 11.3 million would leave. The difference—2.2 million—would be white males’ “net addition to the work force.” Because the numbers of female and nonwhite workers are growing faster than those of white males, white males made up a relatively small share—less than 15 percent—of the anticipated total net addition to the work force.

Rather than describing the gradual decline in white males’ proportion of the work force in terms of a straightforward percentage (47.9 percent in 1988, falling to 44.8 percent in 2000), the authors of *Work Force 2000* chose to use a more obscure measure (net additions to the work force). That was an unfortunate choice, because it invited confusion. In fact, it even confused the people who prepared the report. *Work Force 2000* came with an executive summary for those too busy to read the entire document. The summary mangled the report’s findings by claiming that “only 15 percent of the new entrants to the labor force over the next 13 years will be native white males, compared to 47 percent in that category today.” That sentence was wrong for two reasons. First, it confused net additions to the labor force (expected to be roughly 15 percent white males) with all new entrants to the labor force (white males were expected to be about 32 percent of all those entering the labor force); and, second, it made a meaningless comparison between the percentage of white males among net work force entrants and white males’ percentage in the existing labor force (roughly 47 percent). The statistical comparison seemed dramatic, but it was pointless.

Unfortunately, the dramatic number captured people’s attention. The press fixed on the decline in white male workers as the report’s major finding, and began to repeat the error. Officials at the Department of Labor
tried to clarify its meaning, but the mutant statistic took on a life of its own. Politicians, labor and business leaders, and activists all warned that the workplace was about to undergo a sudden change – that white males were an endangered species.

The mangled statistic was itself remangled; for example, one official testified before Congress that by “the year 2000, nearly 65 percent of the total work force will be women,” yet no one asked how or why that might occur.

It is easy to see why people repeated these claims. The notion that white males would soon become a small proportion of all workers offered support for a variety of political ideologies. Liberals saw the coming change as proof that more needed to be done to help women and minorities – who, after all, would be the workers of the future. Liberal proposals based on Work Force 2000 called for expanded job training for nontraditional (that is, nonwhite or female) workers, additional programs to educate management and workers about the need for diversity in the workplace, and so on.

In contrast, conservatives viewed the changing work force as further evidence that immigration, feminism and other developments threatened traditional social arrangements. In response to claims that white male workers were disappearing, a wide range of people found it easier to agree (“We knew it! We told you so!”) than to ask critical questions about the statistical claims.

The reaction to Work Force 2000 teaches a disturbing lesson: complex statistics are prime candidates for mutation. Not that the statistics in Work Force 2000 were all that complex. But the meaning of “net additions to the work force” was not obvious, and when people tried to put it in simpler language – such as “new workers” – they mangled the concept.

The report’s authors ought to have realized that most people would not grasp this relatively complicated idea. And, of course, the people who interpreted the report (beginning with the authors of the executive summary) unintentionally mangled it to produce figures with new, wildly distorted meanings. Thus, a correct but difficult-to-understand statistic became an easy-to-understand but completely wrong number.

Similar confusion characterized press coverage of a medical study supposedly showing that doctors referred blacks and women for cardiac catheterization less often than whites and men. In the study, researchers gave doctors information (e.g., descriptions of chest pain and results of stress tests) about fictional patients who were described as either black or white, female or male. The doctors were asked how they would treat these patients, and the researchers examined which kinds of patients were referred for cardiac catheterization. Interestingly, white women, black men, and white men were equally likely to receive referrals for the procedure: catheterization was recommended for 90.6 percent of the patients in each group. In contrast, only 78.8 percent of black women were referred for catheterization.

This study attracted considerable press coverage when the media summarized the results as showing that blacks and women were 40 percent less likely to receive cardiac testing. How could the press produce this mangled statistic from these data?

The answer lies in the researchers’ decision to report their results in terms of odds ratios. Producing this statistic involved a two-stage calculation. First, the researchers calculated the odds of people in different groups being referred for catheterization. Remember that 90.6 percent of white women received referrals; this means that, among 1,000 white
women, 906 would get referrals, and 94 would not; therefore the odds of a white woman being referred were 9.6 to 1 (906 referrals/94 non-referrals = 9.6). That is, for every white woman not referred, 9.6 would be referred.

Black men and white men had the exactly same percentages of being referred, and therefore the same odds of referral. But the odds of black women being referred were only 3.7 to 1. Among 1,000 black women, there would be 788 referrals and 212 non-referrals. Thus for every black woman not referred, 3.7 would be referred.

Notice that when white men and white women are grouped together, both sexes had the same rates of referral and the same aggregate odds – 9.6 to 1. However, when black men (who had the same rate of referral as whites) were combined with black women (who had a lower rate of referral), the overall odds for all blacks were lower – 5.5 to 1. Similarly, the aggregate odds for all men (black and white) were higher than the aggregate odds for all women.

So far, we have been talking about the odds of being referred for cardiac catheterization. But the researchers reported the odds ratios. This slightly more complicated statistic involves a second stage of calculation. For example, the ratio of the odds of men being referred (9.6) to the odds of women being referred (5.5) is 1 to 0.6 (5.5/9.6 = 0.6). Similarly, the ratio of the odds of whites being referred (9.6) to the odds of blacks being referred (5.5) is 1 to 0.6. Odds ratios, like net additions to the work force, are statistics that lack intuitive meaning. Most people don’t think in terms of odds ratios, nor do they understand what the term means.

Certainly, this was true for the reporters who announced that blacks and women were only 60 percent as likely to receive heart testing as whites and men were. They misunder-stood the odds ratio (0.6) to mean the relative likelihood of receiving the procedure. The correct comparison would have involved calculating the risk ratios – that is, figuring the relative chance, or risk, of being referred for testing.

If 90.6 percent of whites and 84.7 percent of blacks are referred, then blacks are 93 percent as likely (84.7/90.6 = .93) to get referrals. That is, blacks were 93 percent as likely to be referred as whites, and women 93 percent as likely as men. Blacks and women, then, were not 40 percent less likely to receive referrals; they were 7 percent less likely to be referred.

As in the case of Work Force 2000, the misinterpretation began by mangling a poorly understood statistic. Reporters tried to translate the unfamiliar notion of odds ratios into more familiar statements of probability, and their resulting claims were simply wrong.

Two other aspects of this case deserve mention. First, the researchers’ decision to compare grouped data for men and women (and blacks with whites) distorted their findings. Remember that white men, black men, and white women had exactly the same rates of referrals. The use of aggregate comparisons obscured the real pattern (that only black women were referred at lower rates). Rather than suggesting that all women or all blacks were less likely to receive referrals, the researchers should have emphasized that black women received different recommendations from all other patients.

Second, we might wonder about the significance of receiving those referrals. The press reports simply assumed that referrals for cardiac catheterization (an invasive medical procedure that carries its own risks) were always appropriate, in effect implying that every patient should have received a referral. But this may be wrong. Perhaps the study showed that physicians were too quick to refer men...
and white women for a risky procedure. But the press reports never considered this possibility. The press also tended to forget that the doctors in this study were examining fictitious files, not treating real patients.

The ease with which somewhat complex statistics can produce confusion is troubling because we live in a world in which complex numbers are becoming more common. Simple statistical ideas – fractions, percentages, rates – are reasonably well understood by many people. But many social problems involve complex chains of cause and effect that can be understood only through complicated models. Thus, current understandings for why some people develop heart disease or cancer assume that heredity plays a part, that various behaviors (diet, exercise, smoking and so on) play roles, and that the environment has an influence. Sorting out the interconnected causes of these problems requires relatively complicated statistical ideas – net additions, odds ratios and the like. If we have an imperfect understanding of these ideas, and if the reporters and other people who relay the statistics to us share our confusion, the chances are good that we'll soon be hearing – and perhaps making decisions on the basis of – mutated statistics.

**Compound Errors:**
**Creating Chains of Bad Statistics**

Bad statistics often take on a life of their own. Rarely criticized, they gain widespread acceptance, and they are repeated over and over. Each repetition makes the number seem more credible. And, of course, bad statistics can become worse through mutation. But that's not the end of the process. Bad statistics can have additional ramifications when they become the basis for calculating still more statistics.

We can think about this process as compounding errors into a chain of bad statistics;
and his colleagues conducted lengthy interviews with several thousand people about their sexual experiences. These interviews became the basis for two books, *Sexual Behavior in the Human Male* (1948) and *Sexual Behavior in the Human Female* (1953), that are popularly known as the Kinsey Reports.

The books challenged the polite fiction that most sex was confined to marriage; they revealed that many people had experience with a wide range of sexual behaviors, such as masturbation and premarital sex. However, the Kinsey data could not provide accurate estimates for the incidence of different sexual behaviors. While the interviews constituted a large sample, that sample was not representative, let alone random. It contained far more college-educated people than there were in the general population and, in an effort to explore a broad range of sexual experiences, Kinsey deliberately arranged interviews with a substantial number of active homosexuals, as well as many people who had been imprisoned.

Nonetheless, commentators sometimes treat the Kinsey findings as though they offer an authoritative, representative portrait of the American population. For example, gay and lesbian activists sometimes argue that one-tenth of the population is homosexual, and they refer to the Kinsey Reports to support that claim.

Rather than define heterosexual and homosexual as a simple dichotomy, the Kinsey Reports described a continuum that ranged from individuals who had never had a homosexual experience to those who had some incidental homosexual experiences to those whose sexual experiences had been exclusively homosexual. Still, the male report estimated that “10 percent of the males are more or less exclusively homosexual... for at least three years between the ages of 16 and 55.” Later surveys, based on more-representative samples, have concluded that 3 to 6 percent of males (and a lower percentage of females) have had significant homosexual experience, and that the incidence of homosexuality among adults is lower – between 1 and 3 percent. However, gay and lesbian activists often dispute these lower estimates; they prefer the one-in-ten figure because it suggests that homosexuals are a substantial minority – roughly equal in number to African-Americans in the United States – who form a group too large to be ignored. Thus, the 10 percent figure lives on, and it is often used in calculating other new statistics about
gays and lesbians. Consider, for example, claims that one-third of teenage suicides – or roughly 1,500 deaths a year – involve gay or lesbian adolescents. Gay activists invoke this statistic to portray the hardships gay and lesbian youth confront. It suggests that stigma and social isolation are severe enough to drive many adolescents to kill themselves.

But how could anyone hope to measure gay teenage suicides accurately? Many gays and lesbians try to conceal their sexual orientation, and some teenagers might have been driven to suicide because keeping that secret was becoming a burden. But, given this secrecy, how could anyone know which teenagers who commit suicide are gay or lesbian?

So how did advocates arrive at the statistic? They constructed a chain of bad statistics. They began with the familiar, Kinsey-based claim that one-tenth of the population – including, presumably, one-tenth of teenagers – is homosexual. Roughly 4,500 teenage deaths are attributed to suicide each year; on average, then, 10 percent of those – 450 suicides – should involve gay or lesbian teens. (Note that we have already incorporated our first dubious statistic: that 10 percent of the population is gay or lesbian.)

Next, advocates drew upon various studies that suggested that homosexuals attempted suicide at two to three times the rate of heterosexuals. Note that this figure presumes knowledge about the rates of an often-secretive behavior in two populations – one itself often hidden. Multiplying 10 percent by three led to an estimate that gays and lesbians accounted for 30 percent of suicides – and this figure was in turn rounded up to one-third. Thus, one-third of 4,500 teenage suicides – 1,500 deaths – involve gay or lesbian youths.

Notice how the final figure depends on the proponents’ assumptions. If the proportion of homosexuals among all teenagers is estimated at 3 percent, or 6 percent, the number of gay teenage suicides falls. If the rate at which homosexual teens commit suicides is only twice that of heterosexuals, the number falls.

This example offers two important lessons. The first is a reminder that bad statistics can live on. Most social scientists consider Kinsey’s 10 percent estimate for homosexuality too high. Yet some gay and lesbian activists continue to cite it precisely because it is the largest available number. In turn, 10 percent often figures into other calculations – not just about gay teenage suicides, but estimates of the number of gay voters, the size of the gay population at risk of AIDS, and so on.

The second lesson is perhaps harder to learn. Any claim about the number of gay teenage suicides should set off alarm bells. Given the difficulties in learning which deaths are suicides and which teenagers are gay, it obviously must be hard to learn the number of gay teenage suicides. It is not unreasonable to ask how the advocates arrived at that number, and which assumptions lay behind their calculations. But, of course, such examinations are the exception, not the rule.

Compound errors can begin with any of the standard sorts of bad statistics – a guess, a poor sample, an inadvertent transformation, perhaps confusion over the meaning of a complex statistic. People inevitably want to put statistics to use, to explore their implications. An estimate for the number of homeless people can help us predict the costs of social services for the homeless, just as an estimate of the proportion of the population that is homosexual lets us predict the number of gay and lesbian teenagers who may attempt suicide. But when the original numbers are bad, compound errors can result. Assessing such statistics requires another level of critical
thinking; one must ask both how advocates produced the statistic at hand (1,500 gay teenage suicides), and whether they based their calculations on earlier numbers that are themselves questionable (e.g., 10 percent of the population is homosexual).

THE ROOTS OF MUTANT STATISTICS
Detecting mutant statistics often requires tracing the history of a number, and learning how its meaning or use changes over time. Mutant statistics don’t always start out bad. Although a bad statistic often provides an excellent basis for mutation, even good statistics can be mangled into bad mutations.

Mutant statistics are not necessarily evidence of dishonesty. Many advocates are perfectly sincere, yet innumerate. Their conviction that the problem is serious and that they need to make that seriousness clear to others, coupled with their misunderstanding of what the numbers actually mean, provides the foundation for many of the sorts of errors detailed here.

However, there is also deliberate manipulation – conscious attempts to turn statistical information to particular uses. Data can be presented in ways that convey different impressions, and it is not uncommon for advocates to selectively choose which numbers they report, and to pick the words they use to describe the figures with care. That is, the numbers are selected because they promise to persuade, to support the advocates’ positions.

Whether they are sincere or cynical, advocates prefer dramatic statistics, numbers that make the problem seem as serious, the need as urgent, as possible. If, through transformation or confusion or compound error, they devise a mutant statistic that happens to be more dramatic than some original figure, there is a good chance that the mutation will spread. Drama ensures repetition, while innumeracy discourages critical thinking.

Drama ensures repetition, while innumeracy discourages critical thinking.