11
Lying with statistics

We prefer the term "statistical communication," but the phrase "how to lie with statistics" is a good hook to get students thinking about the issues involved. We try throughout to dampen the natural cynicism that comes with this topic and emphasize that, to most effectively tell the truth, you must avoid lying by accident as well as on purpose. We assign readings and also discuss several ways of lying with statistics that are not covered in usual treatments of the topic.

11.1 Examples of misleading presentations of numbers

We illustrate the difficulties of statistical communication by clipping newspaper articles and discussing them in class. (The misleading is not always done by the newspaper; in many cases, newspapers report on lying done by others.) When presenting an example, we break the class into groups of four, give them two minute to discuss, and then move to a general class discussion. In this section we give several examples of articles we have used in class.

11.1.1 Fabricated or meaningless numbers

The simplest form of lying with statistics is to simply to make up a number, such as Senator McCarthy's proclaimed (but nonexistent) list of 205 Communists, or, for a more recent example, "Patriot Missile Hits Revised from 41 to 4" (see Fig. 11.1). More subtly, numerical measurements can be used dubiously, as in the article, "Survey: U.S. Kids Reading Well" displayed in Fig. 11.1: it is not at all clear if there is a reasonable way to compare reading ability in Finnish, Hungarian, English, Chinese, and so forth. Comparison may be possible, but without more detail, it is not clear how to interpret these rankings. Amusingly, the two articles in Fig. 11.1 appeared on the very same page of the newspaper (which of course reveals the possibilities of using newspaper articles as source material for statistics classes). These examples are useful in class because the students are told so much about subtle ways of misleading with statistics that it is refreshing to remind them that simple fabrication or conceptual errors are possible too.

11.1.2 Misinformation

Perhaps the most common error involving statistics is to make a claim that is contradicted by available statistical information. It is not difficult to find such ex-

Published 2017 by Oxford University Press.

A-8 Wednesday, September 30, 1992 ★★ SAN FRANCISCO EXAMINER

Survey: U.S. kids reading well

No relation between skills, economics

By Dianne Henk
ASSOCIATED PRESS

ALBANY, N.Y. — Schoolchildren in the United States rank near the top of the class in a 31-nation study of basic reading skills.

Finnish students finished first in both age groups tested, ages 9 and 14, said the International Association for the Evaluation of Educational Achievement.

The United States was second among 9-year-olds and ninth among 14-year-olds. The United States trailed by small margins among 14-year-olds, said Alan Purves, a member of the research team and an education professor at the State University of New York at Albany.

The study found reading levels in a country are closely related to economic development, health and adult literacy.

"It's not necessarily whether you're richer; it's whether the kids' families have more books or not," Purves said.

"Parents clearly can help, particularly by providing books in the home," he said. "And it looks as if they can help by encouraging kids to read and reading to kids."

The study found American schools do a good job of teaching basic reading skills, although students were weaker on reading such things as maps and charts, Purves said.

The test did not address critical reading skills, such as grasp of poetry and philosophical argument.

Girls scored consistently higher in all countries and for both ages.

The report supported findings that extensive television viewing hurts reading scores.

Viewing up to two hours of television daily did not show much of an effect, Purves said. But "there's a downward slope when you get beyond about 2½ hours," he said.

READING SKILLS

Rankings from the International Association for the Evaluation of Educational Achievement for basic reading skills of 9-year-olds and 14-year-olds. Not all countries participated in both age group studies.

9-year-olds

1. Finland
2. United States
3. Sweden
4. France
5. Italy
6. New Zealand
7. Norway
8. Iceland
9. Hong Kong
10. Singapore
11. Switzerland
12. Ireland
13. Belgium (French)
14. Greece
15. Spain
16. West Germany
17. Canada (British Columbia)
18. East Germany
19. Hungary
20. Slovenia
21. Netherlands
22. Cyprus
23. Portugal
24. Denmark
25. Trinidad-Tobago
26. Indonesia
27. Venezuela

14-year-olds

1. Finland
2. France
3. Sweden
4. New Zealand
5. Hungary
6. Iceland
7. Switzerland
8. Hong Kong
9. United States
10. Singapore
11. Slovenia
12. East Germany
13. Denmark
14. Portugal
15. Canada (British Columbia)
16. West Germany
17. Norway
18. Italy
19. Netherlands
20. Ireland
21. Greece
22. Cyprus
23. Spain
24. Belgium (French)
25. Trinidad-Tobago
26. Thailand
27. Philippines
28. Venezuela
29. Nigeria
30. Zimbabwe
31. Botswana

Patriot missile hits revised from 41 to 4

GAO downgrades weapons' success against Iraqi Scuds; congressman says U.S. was misled

By David Evans
CHICAGO TRIBUNE

WASHINGTON — Patriot missiles shot down four Iraqi Scud missiles during the Persian Gulf war, far fewer than the 41 successes in 42 engagements claimed at war's end by the Defense Department, according to the latest analysis of the missile's combat performance.

In a shot-by-shot review of each missile launch and ground damage reports, General Accounting Office investigators found that Iraqi missiles were intercepted in only 9 percent of the Patriot engagements. The GAO is the investigative watchdog for Congress, and the audit of Patriot performance was done on behalf of the House Government Operations Committee.

In releasing the report Tuesday, committee Chairman John Conyers Jr., D-Mich., said: "We have watched the claims for this missile drop from 100 percent during the war to 96 percent in official statements to Congress, to 80, 70, 52, 25, and now we're under 10 percent and dropping.

"The Patriot may have hit only a few Scud warheads, and there are doubts about these. ... The public and Congress were misled," Conyers declared.

Maj. Peter Keating, an Army spokesman, said: "The GAO report does repudiate the critics who asserted there wasn't a single warhead kill during the war."

The Pentagon benefited enormously from the initial impression of a near-perfect missile defense. Congress increased the Patriot budget by hundreds of millions of dollars last year and pumped an additional $1 billion into the "Star Wars" global missile-defense program.

The GAO report caps a months-long debate in which independent experts have criticized the Army for inflating the Patriot's wartime performance.

The GAO report said the Army now expresses "high confidence" that 25 percent of the Patriot engagements hit Scuds, but even this claim cannot be supported by the evidence. The Army relied heavily on ground damage reports and "probable kill" messages flashed by the Patriot missiles just before they detonated.

In both cases, the data are unreliable.

The GAO investigators took a more conservative approach, giving credit for a successful intercept only if a recovered Scud contained Patriot fragments or if radar tapes showed a rapid slowdown of a Scud, indicating falling debris after an intercept.

Several of the million-dollar Patriots were fired in each engagement when it appeared that a Scud had been launched. Of the total of 158 fired during the gulf war, it now appears that half were launched at non-existent targets or at falling debris from Scuds that broke up on re-entry.

San Francisco Examiner
For delivery call toll free
1-281-EXAM

Fig. 11.1 Two articles, appearing on a single page of the *San Francisco Examiner*, illustrating potentially misleading numbers. In the top article, it is not explained how one can accurately compare reading ability for students using different languages. The bottom article discusses the possibility that the Army fabricated statistics during the Gulf War.

amples once you looking out for them. For example, we came across the following statement in the *Economist* magazine:

> Back in Vietnam days, the anti-war movement spread from the intelligentsia into the rest of the population, eventually paralyzing the country's will to fight.

This common belief is in fact false: as described in Section 3.5.2, the highly educated people in the United States were in fact *more* likely to support the Vietnam War (see Fig. 3.8 and the discussion on page 362). If the students have worked this example when covering descriptive statistics, it is enlightening for them to see a reputable magazine make the same error they made themselves earlier in the semester.

In this case, the magazine's statistical error was to make a false claim without checking against readily available information on public opinion. This sort of mistake illustrates why statistical data are gathered in the first place.

11.1.3 Ignoring the baseline

A common error, whether accidental or intentional, is to compare raw numbers without adjusting for expected baseline differences. For an obvious example, it is no surprise that California has more teachers than Arizona, given that California has many more residents. A more reasonable comparison would be teachers per person in each state, or perhaps the number of teachers divided by the number of children between 5 and 18 years of age. As this simple example illustrates, it is not clear what should be the baseline, but for most purposes it is important to make some sort of adjustment.

Not adjusting for baseline occurs all the time. For example, it is commonplace for dollar trends over time to be reported without adjusting for inflation. There is some debate over the most appropriate price adjustment (see Section 3.7.2) but it can't be right to use raw dollars. A more subtle adjustment problem appears in Fig. 11.4 on page 170.

Figure 11.2 illustrates how baselines can be ignored in a map. In this map of Berkeley, California, areas were shaded that had more than 200 thefts and 75 burglaries in the previous year. We show this map to the students and ask what is wrong here; eventually they realize that it would be more appropriate to compare crime rates per population. For example, the large shaded region on the left of the map contains relatively few people; in fact, much of the shaded area is in the San Francisco Bay.

The class can continue by discussing appropriate measures of population—that is, the denominator in the crime rate—for example, perhaps burglaries should be measured per household and robberies measured with respect to the number of pedestrians who frequent the area.

11.1.4 Arbitrary comparisons or data dredging

A more subtle error is selection of data, which is also illustrated by the map in Fig. 11.2. (Once again, there is so much educational potential from a single newspaper clipping.) Setting aside the problems with population variation, the decision about where to set the threshold for shading appears arbitrary. What if

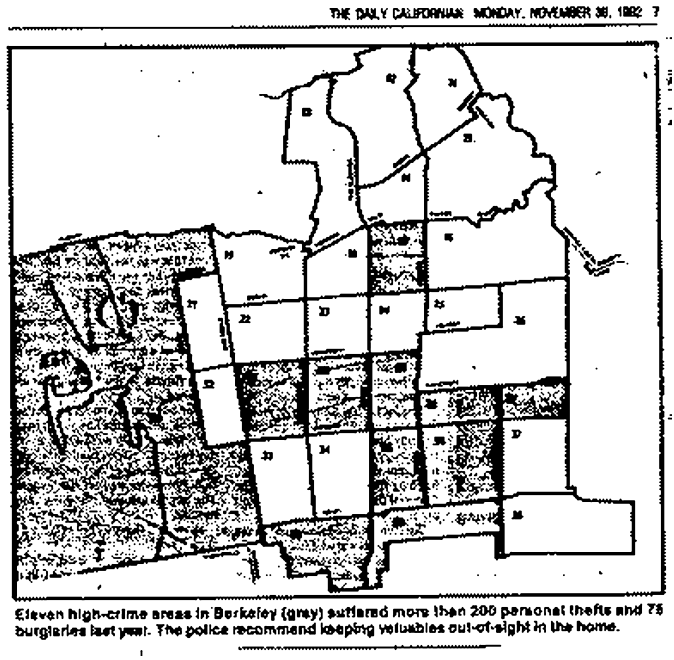


Fig. 11.2 A map, from the student newspaper of the University of California at Berkeley, illustrating several potential statistical fallacies. Areas are shaded or not shaded according to total crimes, without adjusting for possible differences in population in the different areas. In addition, the threshold—200 personal thefts and 75 burglaries—appears somewhat arbitrary, suggesting that the shading on the map might be open to manipulation.

the rule were changed to "more than 300 thefts and 50 burglaries" or "more than 100 thefts and 100 burglaries" or "more than 250 thefts or burglaries"? Unless this was some sort of preset standard, this thresholding rule seems subject to manipulation in a way similar to multiple comparisons (see Section 9.7.2).

We ask the students how they could improve the map to allay these suspicions. One suggestion has been to construct two smaller maps—one for thefts and one for burglaries—and to use four levels of shading to indicate ranges of theft or burglary rates.

For a more lighthearted example of selection, we tell the students of a comment we saw in the newspaper several years ago: "The team whose city has the tallest free-standing structure has won six of the past ten [baseball] World Series." This statement was obviously intended to be humorous, but it is interesting to debunk it. First, from the structure of the statement, we can suppose that the team with the tallest free-standing structure *lost* the World Series eleven years earlier (otherwise, the statement would presumably have been "seven of the past eleven" or "eight of the past twelve" or whatever). We thus have "six out of the

past eleven," which is as close to 50/50 as can be, given that you can't win half a World Series.

Some other strange numerical comparisons appear in the articles shown in Fig. 11.3. For example, it is sad that 13 children were being killed per day, but it is not at all clear why this should be compared to the rate at which police officers are shot. The comparison later in the article to Northern Ireland is more reasonable (although it might be even more relevant to compare all violent deaths rather than restrict to gunshots).

11.2 Selection bias

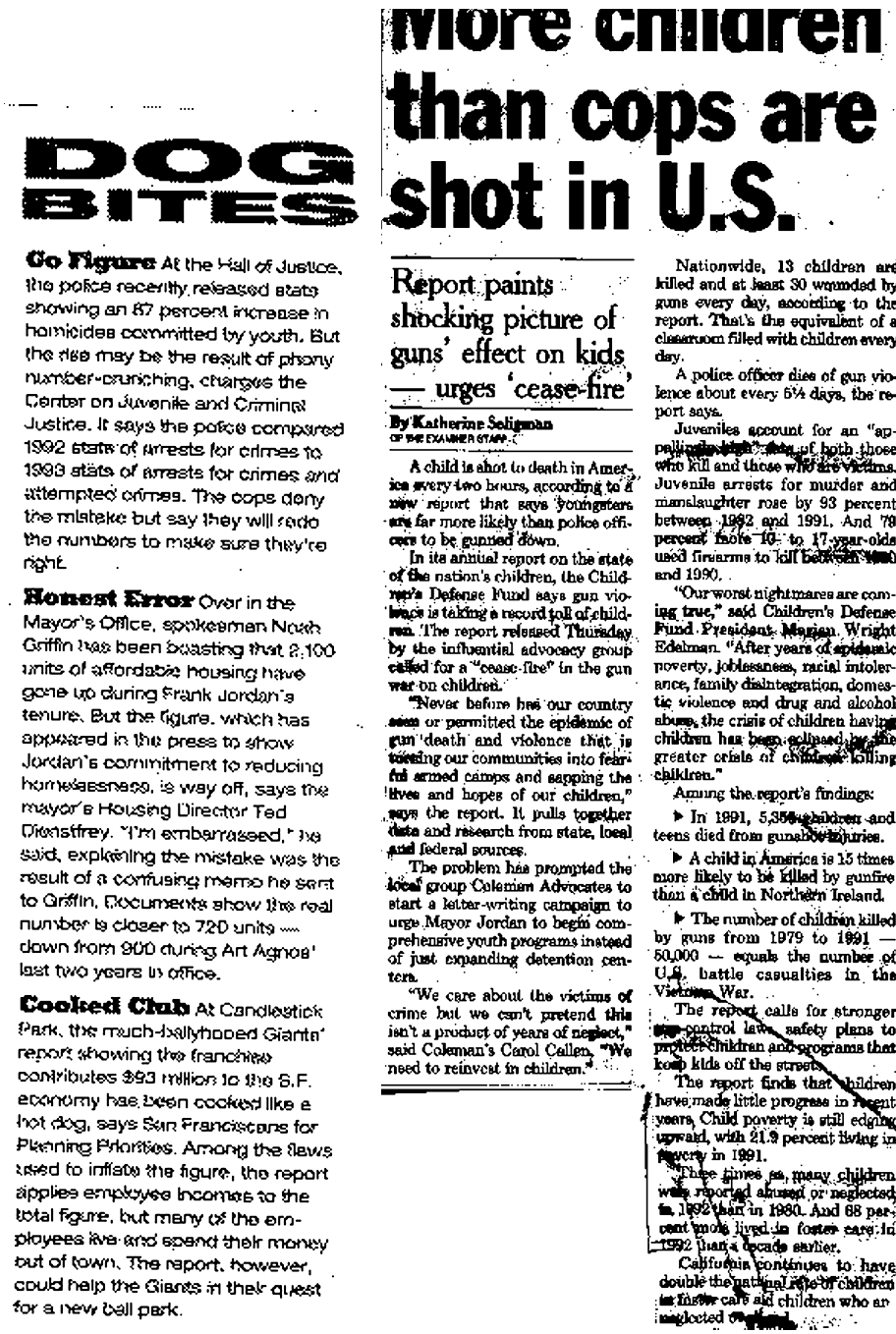
A favorite source of statistical errors is selection bias, which can generally be categorized as a sample being unrepresentative of the population because some units are much more likely than others to be represented, with the more likely units differing from the unlikely units in some important way. Before getting to this topic, we like to discuss simpler methods of lying with statistics (see above) to make it clear that no great sophistication is required to mislead.

11.2.1 Distinguishing from other sorts of bias

Covering selection bias in class has two benefits: in addition to reminding the students of an important source of error, it gets them thinking systematically about sampling and probability as applied to real-world settings. Whenever we introduce a selection bias example, we like to stop and ask the students to calculate or guess the probability of selection for different units in the population. We can illustrate with many examples that we have already covered in class or homework:

- Surveys of World Wide Web users that overrepresent frequent users (Section 6.1.3)
- Counts of number of siblings in families, in which larger families are more likely to be selected because they have more children that could end up as students in the class and thus be counted (Section 6.1.6)
- Sampling from a bag of candy; larger candies are more likely to be selected (Section 9.1).

We also explain that there are all sorts of biases in statistics that are *not* selection bias. We have already considered in class: measurement error (as in the age-guessing demonstration on the first day of class (Section 2.1) or the United Nations experiment described in Section 6.4.1); lurking variables (as in the regression of earnings on height, ignoring sex, in Section 5.1.2), biased survey questions (see Section 6.1.3), inappropriate comparisons in observational studies (such as the biased estimates of the effect of coaching on SAT scores, described in Section 6.5.3), and multiple comparisons (in the study of the effects of prayer, described in Section 9.7.3). These all illustrate biases, but they are not *selection* bias because, in all these cases (with the partial exception of the SAT coaching example), the problem is not with sample selection but with the measurements or analysis performed on the units selected.



DOG BITES

Go Figure At the Hall of Justice, the police recently released stats showing an 87 percent increase in homicides committed by youth. But the rise may be the result of phony number-crunching, charges the Center on Juvenile and Criminal Justice. It says the police compared 1992 stats of arrests for crimes to 1993 stats of arrests for crimes and attempted crimes. The cops deny the mistake but say they will redo the numbers to make sure they're right.

Honest Error Over in the Mayor's Office, spokesman Nosh Griffin has been boasting that 2,100 units of affordable housing have gone up during Frank Jordan's tenure. But the figure, which has appeared in the press to show Jordan's commitment to reducing homelessness, is way off, says the mayor's Housing Director Ted Dienstfrey. "I'm embarrassed," he said, explaining the mistake was the result of a confusing memo he sent to Griffin. Documents show the real number is closer to 720 units — down from 900 during Art Agnos' last two years in office.

Cooked Club At Candlestick Park, the much-ballyhooed Giants' report showing the franchise contributes $93 million to the S.F. economy has been cooked like a hot dog, says San Franciscans for Planning Priorities. Among the flaws used to inflate the figure, the report applies employee incomes to the total figure, but many of the employees live and spend their money out of town. The report, however, could help the Giants in their quest for a new ball park.

More children than cops are shot in U.S.

Report paints shocking picture of guns' effect on kids — urges 'cease-fire'

By Katherine Seligman
OF THE EXAMINER STAFF

A child is shot to death in America every two hours, according to a new report that says youngsters are far more likely than police officers to be gunned down.

In its annual report on the state of the nation's children, the Children's Defense Fund says gun violence is taking a record toll of children. The report released Thursday by the influential advocacy group called for a "cease-fire" in the gun war on children.

"Never before has our country seen or permitted the epidemic of gun death and violence that is turning our communities into fearful armed camps and sapping the lives and hopes of our children," says the report. It pulls together data and research from state, local and federal sources.

The problem has prompted the local group Coleman Advocates to start a letter-writing campaign to urge Mayor Jordan to begin comprehensive youth programs instead of just expanding detention centers.

"We care about the victims of crime but we can't pretend this isn't a product of years of neglect," said Coleman's Carol Cullen. "We need to reinvest in children."

Nationwide, 13 children are killed and at least 30 wounded by guns every day, according to the report. That's the equivalent of a classroom filled with children every day.

A police officer dies of gun violence about every 6¼ days, the report says.

Juveniles account for an "appalling high" [illegible] of both those who kill and those who are victims. Juvenile arrests for murder and manslaughter rose by 93 percent between 1982 and 1991. And 79 percent more 10- to 17-year-olds used firearms to kill between 1985 and 1990.

"Our worst nightmares are coming true," said Children's Defense Fund President Marian Wright Edelman. "After years of epidemic poverty, joblessness, racial intolerance, family disintegration, domestic violence and drug and alcohol abuse, the crisis of children having children has been eclipsed by the greater crisis of children killing children."

Among the report's findings:

- In 1991, 5,356 children and teens died from gunshot injuries.
- A child in America is 15 times more likely to be killed by gunfire than a child in Northern Ireland.
- The number of children killed by guns from 1979 to 1991 — 50,000 — equals the number of U.S. battle casualties in the Vietnam War.

The report calls for stronger gun control laws, safety plans to protect children and programs that keep kids off the streets.

The report finds that children have made little progress in recent years. Child poverty is still edging upward, with 21.9 percent living in poverty in 1991.

Three times as many children were reported abused or neglected in 1992 than in 1980. And 68 percent more lived in foster care in 1992 than a decade earlier.

California continues to have double the national rate of children in foster care and children who are neglected or [illegible]

Fig. 11.3 Newspaper clippings illustrating arbitrary and perhaps meaningless comparisons. The article on the left (from the *San Francisco Weekly*) shows several examples of official reports with questionable numbers. The article on the right (from the *San Francisco Examiner*) makes us wonder whether we should be happy or sad that more children than cops are shot.

11.2.2 Some examples presented as puzzles

Having given students examples of selection bias and clarified the concept, we are ready with several more examples, which we have drawn from the statistical and scientific literature. We present these as puzzles: for each, we briefly describe the phenomenon; the class then tries to figure out the explanation. (Answers are given on page 368.)

1. *The most dangerous profession.* In a tabulation in 1835 of the ages and professions of deceased people, it was found that the profession with the lowest average age of death was "student." Why does being a student appear to be so dangerous?
2. *Age and palm lines.* A study of 100 recently deceased people found a strong positive correlation between the age of death and the length of the longest line on the palm. Does this provide support for the claim that a long line on the palm predicts a long life?
3. *The clinician's illusion.* When asked to judge the severity of a syndrome among their patients, clinical psychiatrists tend to characterize the syndromes as much more serious and long-term, on average, than are estimated by surveys of patients who have the syndrome.
4. *Your friends are (probably) more popular than you are.* Sociologists have conducted surveys in which they select random people and ask for a list of the people they know, and then they contact a sample of the friends and repeat the survey. The people sampled at the second stage have, on average, many more friends than do the people in the original sample. This suggests that, on average, your friends are more popular than you are.
5. *Barroom brawls.* A study of fights in bars in which someone was killed found that, in 90% of the cases, the person who started the fight was the one who died.

11.2.3 Avoiding over-skepticism

When concluding the discussion, we find it useful to remind the students that the existence of statistical biases does not necessarily invalidate a finding. For example, several years ago a widely publicized study found that the average age of death of left-handers was about 8 years less than that of right-handers. (The researchers started with a sample of death certificates and then interviewed close relatives or friends to find out the handedness of the deceased.) The authors of the study went on to speculate about reasons why left-handers may be more likely to die at a younger age.

This research was criticized as biased because the frequency of left-handedness may have changed appreciably over time: if many people born 60 or more years ago were forced to use their right hands, then the probability of being left-handed would be higher for younger people, and this would result in a lower average age of death for left-handers, even in the absence of any greater risk of death for the individuals.

This claim of bias is potentially reasonable—however, the researchers on the original study did other work that suggests this bias is small, and that left-handers really do die younger, on average, than right-handers. The difference of 8 years in the raw data is probably an overestimate, but we should be wary about simply discarding the research—it would be better to estimate the magnitude of the bias and try to correct for it. We do not claim to have the answer here; it is better to keep an open mind and consider different ways of studying the problem.

As always, students must learn to be both skeptical and constructive.

11.3 Reviewing the semester's material

We cover statistical communication in the last week of class, and it provides a useful framework for reviewing all that came before.

11.3.1 Classroom discussion

It is possible to lie (or to make mistakes) by ignoring some key statistical principles. These are typically covered in detail during the semester, and when we cover "lying with statistics," we find it useful to mention these in class, asking students to quickly make up examples of each.

- Correlation does not imply causation (see Section 5.2.2).
- "Statistically significant" does not necessarily mean "important" (we can illustrate this with a hypothetical example of a very small effect that can be discovered with a very large sample size).
- Not "statistically significant" is not the same as zero (recall the example of Section 9.7.1, where 20 shots each are not enough to identify large differences in abilities between basketball shooters).
- Response bias in sample surveys (see Section 6.1.3).
- Misleading extrapolation (for example, the world record times in the mile run (Fig. 3.1 on page 20), or the regression of earnings on height in Section 5.1.2).
- Problems with observational studies (for example, the comparison of SAT scores before and after coaching, described in Section 6.5.3).
- Regression fallacy (as illustrated with several examples in Section 5.3).
- Aggregation (for example, the way the regression of earnings on height changes when sex is included in the model; see Section 10.1.2).

We illustrate many of these examples with newspaper clippings—it is fun and not difficult to get your own, possibly with the help of your students. The point here, however, is not to slam the press. In fact, as we discuss in Chapter 7 in the context of the statistical literacy assignments, we are generally happy about how the newspapers report on scientific and technical issues. But they are not perfect, and they can often let their guards down when their news is not coming from a scientific source.

26 CHICAGO SUN-TIMES, THURSDAY, MARCH 31, 1994

RICH CHAPMAN/SUN-TIMES

Deputy Fire Commissioner John Ormond (from left), Walgreens district manager Kermit Crawford, Fire Commissioner Raymond Orozco and Deputy District Chief John Schneidwind announce Wednesday that Walgreens is donating 1,000 smoke detectors for areas plagued by fires.

Walgreens Donates Smoke Alarms

By Phillip J. O'Connor
Staff Writer

Eighteen people have been killed in Chicago fires so far in March and there were either no smoke detectors or non-working detectors in all of them, Fire Cmdr. Raymond E. Orozco said Wednesday.

So far in 1994, there have been 38 fire deaths—one more than during the comparable period last year—and in 29 of them there were either no smoke detectors or non-working detectors, Orozco said.

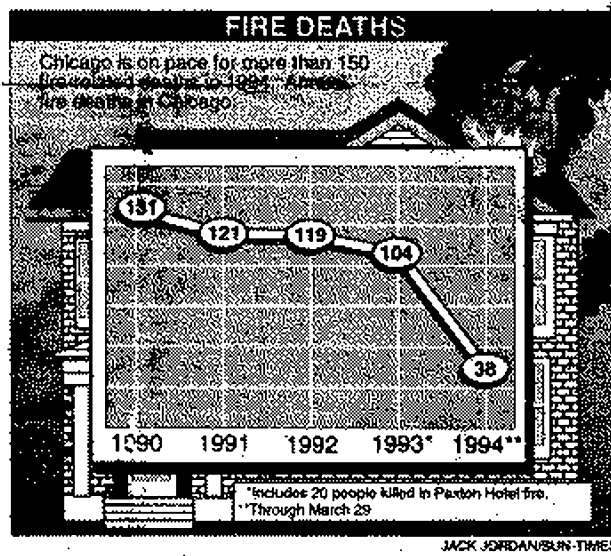
The March toll included three recent fires that killed a total of 13 people, he said.

Orozco cited the statistics at a news conference at which he announced Walgreens Drug Stores has donated 1,000 smoke detectors to the Fire Department, which will distribute them in areas where fire deaths have occurred.

Orozco plans to have members of fire companies that actually fought the deadly blazes go door-to-door and distribute them to citizens whose homes are not protected by detectors.

"Smoke detectors save lives," Orozco said. He urged residents to check batteries in their home smoke detectors over Easter weekend when they change clocks to daylight savings time, which begins early Sunday.

He also urged citizens to help prevent injury and death by developing a home escape plan in the event of a fire and to practice good fire safety. "A fire, prevented, cannot hurt anybody," he said.



JACK JORDAN/SUN-TIMES

Fig. 11.4 Find the two most important statistical errors in this newspaper article. Answers appear on pages 369–369.

11.3.2 Assignments: Find the lie or create the lie

As a homework problem, we assign the problem of finding the two most important statistical errors in the article shown in Fig. 11.4. This is a challenging problem—in fact, many of our colleagues cannot readily find the errors.

In addition, we give the students the following homework assignment:

Do one of the following:

- Find an article in a recent newspaper or magazine that lies with statistics in some way. Explain what the "lie" is and how you would correct it.
- Find an article in a recent newspaper or magazine with numerical information that does *not* lie or mislead. Using the data in the article, create your own misleading "lie."

Include a photocopy of the article with your homework solutions. You get double credit for this problem if your article is *not* used by any other student in the class.

Other possibilities include students working in groups to find the most outrageous statistical "lie" or to produce the most outrageous "lie" themselves.

11.4 1 in 2 marriages end in divorce?

Class discussions can be structured around commonly-quoted, but not necessarily well-understood, numbers. FOr example, we hear statistics like "1 in 2 marriages end in divorce," but how can you really estimate that? It's tricky. If you look at the number of marriages in 1995 per 1000 adults, that number is roughly twice the number of divorces in that year. But, those getting divorces in 1995 are by and large not the same people as those getting married that year. Since so many marriages are ongoing, do you have to wait until one member of a married couple dies before you can count that as a nondivorced marriage? For example, only about 1 in 7 women who married in the early 1940s eventually divorced. But it is not very satisfying to make statements about people who married sixty years ago. Often assumptions are made that the divorce rate in a current year continues indefinitely into the future. A careful study of marriage longevity would need to consider life expectancy and control for age at marriage and length of marriage.

11.5 Ethics and statistics

Ethics is a topic that we find interesting, but students in statistics classes seem to be wary of it. We recommend preparing for a class discussion about ethics by gathering some relevant newspaper clippings; here we give examples of several areas in which ethical issues arise in statistical data collection and analysis. As always, we set up each problem and then ask the students to discuss in pairs as a prelude to general class discussion.

11.5.1 Cutting corners in a medical study

On March 15, 1994, the *New York Times* reported on a Federal investigation of a Canadian researcher at St. Luc's Hospital in Montreal. The investigation found violations of the scientific guidelines that govern the way the study was carried out. According to the *Times* article:

A Federal investigation found that Dr. Poisson had falsified data in his part of the study that helped change the way breast cancer is treated. That influential study concluded that full mastectomies were not necessary to prevent the spread of early forms of the disease in many women.

. . . While insisting that he did little more than tell "white lies" that he believed would not change the conclusions of the study, Dr. Poisson signed an agreement with the Federal Food and Drug Administration acknowledging that he had falsified results of other studies.

In a written reply to the United States Public Health Service's Office of Research Integrity, which spearheaded the Federal investigation, he stated, "I always feel sorry for a nice case to be denied the right to enter a good protocol just on account of trivial details: a difference of a few days in the date of surgery because the patient took a long time to decide."

"When I lost a patient who did not wish to participate," he said, "I always took it as a personal defeat, knowing that the best protocol with the best biostatisticians is useless unless enough patients are registered" . . .

Dr. Poisson said he should have read the fine print of the study design more carefully to avoid the irregularities.

We provide students with a copy of this excerpt from the newspaper article, and because the discussion can get heated, we ask them to write their answer to the following question: On the basis of Dr. Poisson's remarks, why do you think the federal investigators were concerned about the experimental justification for the claim that "full mastectomies were not necessary to prevent the spread of early forms of the disease in many women"?

We give them five minutes to work in pairs to write a short answer and then we collect their written responses and lead a discussion on the topic. We use these written answers to express opinions of students who might be too shy to enter the discussion. In our discussion, we quote from Dr. Poisson's letter to the *New England Journal of Medicine*, where he stated, "My sole concern at all times was with the health of my patients . . . For me, it was difficult to tell a woman with breast cancer that she was ineligible to receive the best available treatments because she did not meet 1 criterion of 22, when I knew this criterion had little or no intrinsic oncologic importance." We also provide the opinion of Dr. Broder, Director of the National Cancer Institute, which funded the study. In his testimony before the House Subcommittee on Oversight and Investigations, he said, "we consider the entire data-set from St. Luc to be a total loss to the American taxpayer." These are delicate issues.

11.5.2 Searching for statistical significance

As another example, a colleague who works at a university statistical consulting service reported the following story. A company wanted to get a drug approved, but their study appeared to have no statistically significant results. (See Section 9.7.2 for a classroom demonstration of multiple comparisons.) The researchers at the company broke up the data into subgroups in about 15 or 20 ways, and then they found something significant. Is this data manipulation? What should the statistician do? In this case, the company reported the results and their stock went up 50%.

11.5.3 Controversies about randomized experiments

A fundamental ethical problem in statistics arises in experimentation, for example in the context of studies of experimental drugs for treating AIDS. On one side, organizations such as the National Institutes of Health insist on randomly assigning treatments (for example, by flipping a coin for each patient to decide which treatment to assign). The advantage of randomized experiments is that they allow reliable conclusions without the need to worry about lurking variables. However, some groups of AIDS patients have opposed randomization, instead making the argument that each patient should be assigned the best available treatment (or, to be more precise, whatever treatment is currently believed to be the best). The ethical dilemma is to balance the benefits to the patients in the study (who would like the opportunity to choose among available treatments) with future patients (who would be served by learning as soon as possible about the effectiveness of the competing treatments).

The issue is complicated. On one hand, the randomized study is most trustworthy if all the patients in the study participate; if they are not treated respectfully, the patients might go outside the study and try other drugs, which could bias the estimates of treatment effects. On the other hand, the patients might be benefiting from being in an experimental study: even if the treatment is randomized, the patients are getting close medical attention from the researchers. Current best practice is to design studies so that all subjects will be expected to benefit in some way, but still keeping the randomized element. For example, a study can compare two potentially beneficial experimental treatments, rather than comparing a treatment to an inert "control." But there will always be conflicts of interest between the patients in the study, the scientists conducting it, and the public at large.

11.5.4 How important is blindness?

Other ethical issues arise in the blindness or double-blindness of experiments. In order to achieve blindness, studies in psychology often use deception. For example, in the United Nations experiment described in Section 6.4.1, the students were not told that the anchoring values of "10" and "65" were an experimental manipulation.

For another example (among many), we bring up in class the topic of stereotyping in job interviews. Are some applicants discriminated against because of their race, ethnicity, or gender? Once students have given some opinions on the topic, we steer the discussion toward the question of how this sort of bias could be measured statistically. One approach is to compare the success of people of different groups in job interviews. This would be an observational study, and we ask the students what sort of lurking variables would need to be controlled for, and how this could be done (see Section 6.5 for examples in which we discuss these issues).

Another way that job discrimination has been studied is to run a randomized experiment in which the experimental subjects are people who might be in a position to evaluate job applicants. We ask students what might be the treat-

ments: for example, the subjects might be given hypothetical resumes, identical in all respects except that the ethnicity of the job applicant is selected at random (similarly to how the anchoring values were assigned at random in the United Nations experiment). This experiment must be done blindly, which necessarily involves deceiving the experimental subjects. (If they are told ahead of time that this is a study of ethnic stereotyping, their heightened awareness may affect their judgments of the job applicants.) Once the experiment is over, the subjects can be told of the true purpose of the experiment, but the practice of misleading them is still controversial and, some would say, unethical. On the other hand, it is important for society to learn the extent of problems such as racial bias, and certain information can be gathered using deception that would be difficult to gather any other way.

In other settings, blindness can pose medical risks. For example, in early studies of heart bypass operations, the treatment was compared to a control regimen of medical (non-surgical) intervention. This posed difficulties both for blindness and double-blindness (a patient knows if he or she has had heart surgery, and so does a doctor making subsequent evaluations). The solution chosen was to perform a "sham operation" on the patients receiving the control treatment—that is, to open up their chests, do nothing, and sew them back up—so they would be externally indistinguishable from the patients who received actual surgery. This certainly seems to be an ethically questionable way to achieve blindness. On the other hand, if it allows doctors to learn more about the effectiveness of heart bypass surgery, maybe it is worth it? This is a topic for student discussion.

11.5.5 Use of information in statistical inferences

Criminal justice

Another sort of ethical issue concerns the acceptable uses of information in making inferences and decisions. Members of ethnic minorities are more likely to be stopped and questioned by police. In what sense is this a reasonable policy given crime statistics? This has been a controversial area, both in stopping people on the street and in their cars on highways. Using the information about ethnicity is generally considered unethical, partly because of its obvious unfairness. There is general agreement that upholding the principle of equal treatment under the law is more important than potential short-term efficiency of police procedures.

In a criminal trial, a jury is not supposed to use knowledge of prior arrests or convictions to draw conclusions about the guilt of a suspect. From a statistical standpoint, if there is evidence that a person has committed previous crimes, he or she is probably more likely to have committed the crime in question—but legally and ethically, it is not considered acceptable to use this information. Similar issues exist in many settings.

Background information and course grades

For another example, consider a course in which a "pre-test" is given at the beginning of the semester to assess general background knowledge. The students are told that the pre-test will not count in their grade, but they are requested to

try their best. Then, during the semester, there are some scheduling difficulties, and the midterm exam for the course has to be canceled. At the end of the semester, the instructor considers using the pre-test score in place of the missing midterm grades. Is this fair? Students generally think not, and in this setting, most instructors would not use the pre-test. However, ignoring this information probably means that the students' final grades will be less accurate indicators of student abilities: if a student did better on the pre-test, he or she is probably a better student, even after controlling for the grades in other aspects of the course. The following two goals conflict: (a) keeping the promise to the students (or, even more generally, in using information before the course began to determine grades) and (b) producing final grades that best reflect students' abilities.

The issue becomes even more complicated when the background information being used is not under the control of the student; examples that can be predictive of grades include ethnicity and parents' education levels.

Models for guessing on multiple-choice exams

Consider a test with several true/false questions. If all the students answer a question correctly, then presumably they all know the answer. Now suppose that half the students get a certain question correct. Then, how many students do you think knew the correct answer (we ask the class)? 50%? One possibility is that *none* of the students knew the correct answer, and they were all guessing.

Now consider a question that is answered correctly by 80% of the students. If a student chosen at random knows the correct answer with probability p, or guesses with probability $1-p$, then we can write, approximately, $p+0.5(1-p) = 0.8$, which yields the estimate $p = 0.6$. Thus, a reasonable guess is that 60% of the students actually know the correct answer and 40% were guessing. The conditional probability of a student knowing the correct answer, given that he or she answered the question correctly, is $60\%/80\% = 0.75$.

Where is the ethical dilemma here? Consider now the task of giving each student a total grade for the exams. The reasoning above suggests that they should get *no* credit for correctly answering the question that 50% of the students answered correctly (since the evidence is that they were all guessing), and they should get 0.75 credit for answering the question that 80% answered correctly, and so forth. Thus, the amount of points that a question is worth should depend on the probability that a student's correct answer was not due to guessing. The ethical question is of the fairness of deciding the grading system after the exam has been taken by the students. Is it fair for two students to get the same number of questions correct on the exam but different total scores, because the "circumstantial evidence" suggests that one of the students was more likely than the other to be guessing?

Misleading information

In other settings, it might not be wise to gather information if there is a suspicion that it will used inappropriately. For example the class activity of Section 8.5.2 demonstrated a scenario in which a person can fail a lie detector test and still

have a greater than 50% chance of being honest. (This happens whenever the probability of error for the lie detector is greater than the probability that a person is a liar.) A similar problem arises in imperfect medical tests of a rare disease: it is possible that most people who test positive actually do not have the disease.

This information can still be useful if used correctly: for example, a failure on a lie detector test can be grounds for further investigation, and a positive test for a disease can motivate further testing. However, if the tests are (inappropriately) assumed to be perfect, then they can result in innocent persons being accused, or healthy persons being unduly alarmed (or even given inappropriate medical procedures). The ethical dilemma is whether to gather information that can be useful in the right hands but misleading if interpreted crudely.

Privacy and confidentiality

Tradeoffs between individual and social benefits occur in many settings where data are gathered on a population. For example, medical researchers can use data on the personal histories of disease sufferers to study factors associated with disease incidence, potentially saving future lives if they discover important relationships. But such a study might require the use of private medical records, which is open to abuse. For example, if the research were funded by an insurance company, it would be important to make sure that the confidential records could not be used to deny people health insurance. Similar issues have arisen when studying HIV and AIDS: it is important for public health authorities to be aware of trends in prevalence of these conditions, but rules requiring doctors to report new cases to the authorities can backfire, by causing some people at risk to avoid testing.

More generally, any study of the general population can place burdens on the persons being studied. Participation rates in public opinion polls have been declining for decades, partly because market researchers and others have saturated people with unsolicited telephone calls. When we talk about sample surveys in a statistics class, we should remember that the respondents are not simply objects of study, but participants (who are often giving their time for free) in a research project.