Statistics and Passive Smoking

Statistics has long played a prominent role in evaluating the effects of smoking on human health. From the time when doctors first suspected a link between smoking and lung cancer, an enormous number of studies have been conducted to determine (i) whether this apparent relationship was real, (ii) the nature and strength of the relationship, and (iii) whether smoking was indeed the causative mechanism behind most lung cancers among smokers. The history of this scientific research has frequently been permeated with controversy. In fact, Sir Ronald Fisher, arguably the greatest statistician of all time, objected strenuously to the conclusions of other prominent scientists regarding the effects of smoking [1].

Remarkably, smoking tobacco was once thought to be beneficial to one's health. Jean Nicot, the French ambassador to Portugal in 1558 and the source of the word *nicotine,* actually shipped tobacco from the West Indies to France for studies on its possible use as a cancer cure [2]. Today it is universally agreed that there is strong positive association between smoking and lung cancer, as well as heart disease and several other significant ailments.

Smokers have dramatically higher rates of lung cancer than do nonsmokers. Studies have found that the *relative risk* of lung cancer for a typical smoker is approximately 10 [3]. This means that ten times as many cases of lung cancer can be expected to occur in a group of regular smokers as in an equivalent group of individuals who do not smoke. Moreover, an individual's likelihood of contracting lung cancer rises roughly in proportion to the square of the number of cigarettes consumed per day and to the fourth or fifth power of the number of years the person has been smoking [4].

The major tobacco companies—who have consistently employed high quality statistical experts—have always maintained that although an *association* between smoking and lung cancer exists, a *cause and effect relationship* has never been shown. Perhaps other factors link smoking to lung cancer. For example, could it not be the case that people who tend to be under stress are more likely to become smokers than those who aren't, and that this stress may induce biological changes that lead to lung cancer.

The tobacco industry takes the position that the only definitive methodology that could rule out such possibilities is to randomly assign members of a large collection of young people to either be smokers or nonsmokers for several decades, then compare the lung cancer rates of the two groups—an obviously impractical, and unethical, experiment. Thus in their view a causative relationship can never be established. Nevertheless, the mountains of biological and epidemiological evidence that have been amassed have convinced the public health community and the great majority of the general public that smoking is indeed the *cause* of the high rates of lung cancer among smokers and exsmokers.

Passive Smoking

It is one thing to know that smoking engenders serious health risks among smokers themselves. Smoking is a personal decision and the option to quit is always available. But it is a whole new ball game if passive smoking, that is, exposure to "second-hand smoke" (also called "environmental tobacco smoke") is harmful to a nonsmoker. Now the health of many "innocent victims"—the children, spouses, and coworkers of smokers—is at stake.

For this reason, many states, cities, and counties have recently passed laws restricting the locations in which smokers can light up. Moreover, the impetus for the massive ongoing legal and governmental assault on the tobacco industry has come primarily from the passive smoking issue. Yet the scientific evidence regarding passive smoking is far more cloudy (no pun intended) than that for direct (first-hand) smoke.

Numerous studies on the effects of passive smoking have been conducted during the last several years. The watershed event, however, was the release in January 1993 of a report by the Environmental Protection Agency that concluded that second-hand smoke was indeed harmful. The EPA study is an instructive example of modern statistical analysis, its role in shaping public policy, and the controversies that sometimes result.

The EPA looked at eleven studies of nonsmoking women who lived with longtime smokers. The studies chosen were those judged to be of the highest quality.

MARK SCHILLING is Professor of Mathematics at California State University, Northridge.

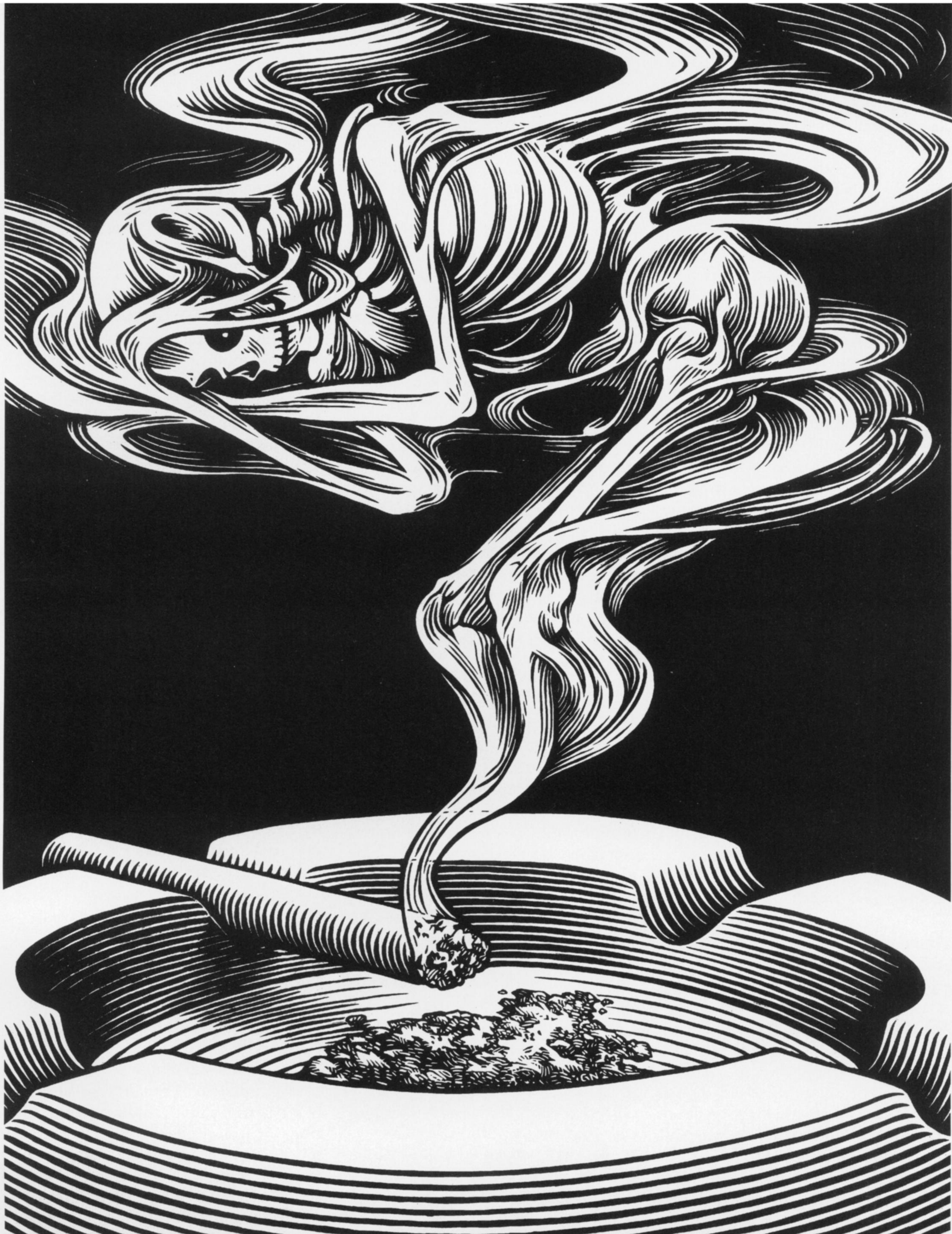
Illustration by Greg Nemec

Now when a large number of scientific studies of the same question are carried out, it is almost inevitable that their results will differ, often greatly. Reasons for this include variations in the populations and experimental designs used, as well as the normal variability that occurs when small to moderate sized samples are used.

Some of the eleven studies the EPA examined showed a higher than expected incidence of lung cancer among the nonsmoking spouses of smokers, while others did not. To make an overall determination concerning the effects of passive smoking, the EPA combined the eleven studies by means of a *meta analysis,* a recently developed type of statistical procedure that gives a different weight to each study reflecting the study's reliability and relevance. Since a meta analysis has a larger effective sample than any individual study, the results of the meta analysis are potentially more accurate.

From the meta analysis the EPA estimated that nonsmoking women who live with smokers have, on average, a relative risk of 1.19 compared to comparable women living in a smoke-free home. That is, living with a smoker results in a 19% greater chance of contracting lung cancer than living with a nonsmoker. This translates into more than 3000 U.S. citizens dying each year of lung cancer caused by breathing another person's cigarette smoke.

A Closer Look

The numbers given in the previous paragraph are known in statistics as *point estimates,* and the information they provide is incomplete in an important way. Associated with any point estimate is a *margin of error,* which provides a statement about the probable accuracy of the estimate. The margin of error is a measure of the extent to which the results of the statistical analysis would vary if it were redone many times (with different experimental subjects each time). Typically this is found by estimating, based on sampling theory, the standard deviation of the collection of different values the point estimate would take on if the study were replicated many times.

Often the (actual) point estimate and the margin of error are combined to form a *confidence interval,* an interval that is highly likely to contain the correct value of the true effect. The point estimate is frequently the exact center of this confidence interval. The EPA's confidence interval for the relative risk of lung cancer in women married to smokers ranged from 1.04 to 1.35. This means that the data from the meta analysis are consistent with an increased risk of lung cancer falling somewhere between 4% and 35%.

In order to determine a confidence interval, however, a statistician first has to select a *level of confidence.* The EPA chose a 90% confidence level. To understand what this means, consider a statistician who calculates a large number of 90% confidence intervals in many different situations in the course of his or her work. In the long run, assuming the statistician makes no statistical design errors or computational mistakes, about 90% of the time such an interval will contain the true value that he or she is trying to estimate. In the remaining cases the interval will miss because of an "unlucky" sample that does not accurately represent the population being studied.

The EPA's choice of a 90% confidence level caused the tobacco companies to cry "foul." As most statistics students know, by far the most common choice for a confidence level is 95% (this is a historical tradition initiated by Ronald Fisher). Now, there is an inverse relationship between the confidence level and the width of the confidence interval. This makes sense since achieving a higher level of confidence should require one to estimate more conservatively. Thus a 95% confidence interval will be wider than a 90% confidence interval.

Why is this important? Because if the EPA's interval is replaced with one only slightly wider, the lower limit of the interval that estimates the relative could fall below 1.0. This would then indicate that the hypothesis that passive smoke has *no effect* on the spouse (relative risk = 1.0) is compatible with the data. In fact, the inclusion in the confidence interval of values *below* 1.0 would allow the possibility that the effect of second-hand is beneficial. Thus the tobacco companies accused the EPA of using a 90% confidence level because they *assumed* that passive smoking could not reduce the risk of lung cancer, and intended all along to show that environmental tobacco smoke is harmful. Surprisingly, the EPA statistical consultants essentially admitted this [5].

The tobacco industry also brought up the point once again that association is not the same as causation. Since the eleven studies the EPA used were observational, rather than being experiments, other "lurking" variables could conceivably account for the difference in lung cancer rates of women married to smokers and those married to nonsmokers. They raise the possibility that the diets of the latter group may have been healthier than those of the former group, offering them some protective effect against contracting cancer. They also point out that the data on the amount of exposure to cigarette smoke of those who died of lung cancer was based on the later recall of relatives, which could be erroneous or even biased.

On July 18, 1998, after five years of court pleadings and deliberations, US Federal District Judge Thomas Osteen ultimately agreed with the industry's arguments. The judge ruled that the EPA had wrongly declared second-hand smoke to be a dangerous carcinogen. Although the administration will almost certainly appeal the decision, the blow to the EPA report may give additional support to the opponents of indoor smoking bans across the nation.

Conclusions

Scientific truths, especially those involving human health, are often quite difficult to establish. Such is the case with passive smoking. It is clear that the deleterious effects of environmental tobacco smoke, whatever they may be, are orders of magnitude smaller than those caused by first-hand smoke, as the concentrations of the chemicals involved are much lower when smoke diffuses into a room or building than when it is drawn directly into the lungs through a 1/4" tube. Thus the case against second-hand smoke is necessarily much harder to establish than against smoking itself.

See Smoking *on the bottom of p. 29.*

gent segment at any point $(t', e^{at'})$ in Figure 4(a) is moved to Figure 4(b) by a horizontal translation of $t - t'$, and a vertical translation of $e^{at} - e^{at'}$. Because the subtangent at each point has constant length $1/a$, the left endpoint of each translated tangent segment in Figure 4(b) lies on a vertical line as shown. As the tangent segments in Figure 4(a) sweep out the region under the curve, the translated segments in Figure 4(b) sweep out the upper half of the rectangle in Figure 4(b). So we see that the area of the region swept out by all the tangents in Figure 4(a) is equal to the area of the corresponding region swept out in Figure 4(b).

We conclude that area A_t is equal to the area of the rectangle shown in Figure 4(b). This rectangle has base $1/a$ and altitude e^{at}, so $A_t = \frac{1}{a}e^{at}$, the product of base and altitude. In the language of integral calculus, we have shown by an intuitive argument that

$$\int_{-\infty}^{t} e^{ax}dx = \frac{1}{a}e^{at}.$$

The intuitive method works here because the subtangents of the exponential have constant length. A similar method based on tangent segments of constant length was used in a paper entitled "Annular Rings of Equal Area" which appeared in the November 1997 issue of *Math Horizons*. These methods form the basis of a new approach to finding areas of certain regions that was developed by one of the authors (Mamikon Mnatsakanian) while he was an undergraduate at Yerevan University in Armenia. Because the definition of the area of a curved figure ultimately depends on a limit process, there is a limit process implicit in the intuitive methods described above. The details of this limit process will be expounded in an English-language text based on these methods that is being developed by the authors under the auspices of *Project MATHEMATICS!*

5. Characterization of translation invariant families

This section shows that the family of exponential curves $y = ce^{ax}$ is the only family that has the property of horizontal translation invariance described in Section 2.

Theorem. *Let f be a real differentiable function defined everywhere on the real line, and let $y = cf(x)$ be the cartesian equation of a family of curves generated by multiplying f by an arbitrary positive constant c. Assume the family is invariant under horizontal translation. That is, assume that for each t there is a constant $c > 0$ (which may depend on t) such that*

(2) $\quad f(x+t) = cf(x) \quad$ *for all x.*

Then we have:

(a) *If $f(0) = 0$, then $f(t) = 0$ for all t.*

(b) *If $f(0) \neq 0$, then $f(t) = f(0)e^{at}$, where $a = f'(0)/f(0)$.*

Proof. Taking $x = 0$ in (2) we find $f(t) = cf(0)$. (a) If $f(0) = 0$ this implies $f(t) = 0$ for all t. (b) If $f(0) \neq 0$ we have $c = f(t)/f(0)$ and (2) becomes $f(x+t) = f(x)f(t)/f(0)$. Differentiating each member with respect to x we find $f'(x+t) = f'(x)f(t)/f(0)$. When $x = 0$ this becomes

$$f'(t) = f'(0)f(t)/f(0) = af(t)$$

where $a = f'(0)/f(0)$. The only solution of this differential equation is $f(t) = f(0)e^{at}$. ■

Smoking (continued from p. 22)

Since experiments that use human subjects are not feasible, evidence must be built up gradually over time through carefully designed observational studies and extensive replication of results. Indeed, such studies continue. The World Health Organization (WHO) has recently completed yet another study of passive smoking [6], looking again at the effects on the nonsmoking spouses of smokers. In this study, which involved approximately 2200 couples, the point estimate of the relative risk for spouses of smokers as compared to spouses of nonsmokers is 1.16, with a 95% confidence interval ranging from 0.93 to 1.44.

Suppose that you worked for the EPA or a public health agency. How would you use the results of the WHO study to argue that it supports the conclusions of the EPA? Then imagine that you worked for the tobacco industry. How would you use this latest study to contend that passive smoking has still not been conclusively shown to be harmful? ■

References

1. Joan Fisher Box, R. A. Fisher: *The Life and Times of a Scientist,* Wiley and Sons, 1978.
2. Edward L. Koven, Smoking: *The Story Behind the Haze,* Nova Science Publishers, Inc., 1996.
3. U.S. Department of Health and Human Services, *Smoking and Health: A National Status Report, 2nd Edition, A Report to Congress,* 1990.
4. U.S. Department of Health and Human Services, *Preventing Tobacco Use Among Young People, A Report of the Surgeon General,* 1994.
5. *The Wall Street Journal,* July 28, 1993.
6. *The Sunday Telegraph,* March 8 & 15, 1998.