INTRODUCTION TO Public Health

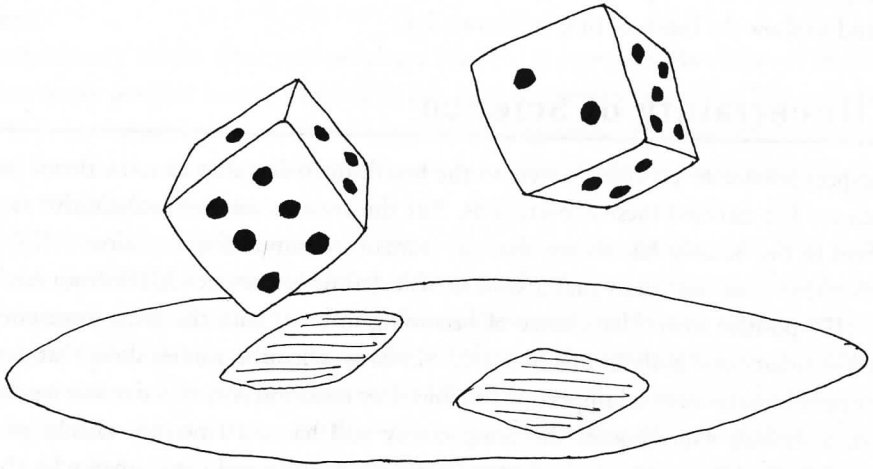
Third Edition

Mary-Jane Schneider

CHAPTER

7

Statistics: Making Sense of Uncertainty



Understanding Uncertainty

The science of epidemiology rests on statistics. In fact, all public health, because it is concerned with populations, relies on statistics to provide and interpret data. Chapter 8 discusses the kinds of data government collects to assess the need for public health programs and evaluate public health progress. The term statistics refers to both the numbers that describe the health of populations and the science that helps to interpret those numbers.

The science of statistics is a set of concepts and methods used to analyze data in order to extract information. The public health sciences discussed in this book depend on the collection of data and the use of statistics to interpret the data. Statistics makes possible the translation of data into information about causes and effects, health risks, and disease cures.

Because health is determined by many factors—genes, behavior, exposure to infectious organisms or environmental chemicals—that interact in complex ways in each individual, it is of-

ten not obvious when or whether specific factors are causing specific health effects. As Chapter 6 discussed, there are ethical and logistical limits to the kinds of studies that can be conducted on human populations, and as Chapter 9 explains, there are limits to the conclusions that can be drawn from biomedical studies. Only by systematically applying statistical concepts and methods can scientists sometimes tease out the one influence among many that may be causing a change in some people's health. Often, however, statistics indicates that an apparent health effect may be simply a random occurrence.

The problems and limits of epidemiology discussed in Chapter 6 are defined in large part by the uncertainties that are the subject of the science of statistics. This chapter discusses the science of statistics in more detail, describing how it is used to clarify conclusions from a study or a test, to put numbers into perspective so that researchers can make comparisons and discern trends, and to show the limits of human knowledge.

The Uncertainty of Science

People expect science to provide answers to the health questions that concern them. In many cases, science has satisfied these expectations. But the answers are not as definitive as people want them to be. Science has shown that the human immunodeficiency virus (HIV) causes AIDS. But that does not mean that a woman will definitely contract AIDS from having sex with an HIV-positive man. Her chance of becoming infected with the virus from one act of unprotected intercourse is about one in 1000.[1] Similarly, scientific studies show that as a treatment for early breast cancer, a lumpectomy followed by radiation is as effective as a mastectomy. However, a woman who chooses the lumpectomy still has a 10 percent chance of cancer recurrence.[2] Both the woman who had unprotected intercourse and the woman who chose the lumpectomy would dearly like to believe that they will be one of those in the majority of cases who will have a positive outcome, but science cannot promise them that. It can only say, statistically, that if 1000 women like her have unprotected sex with an HIV-positive man, 999 probably will fare well while one will not, and if 100 women with early breast cancer have a lumpectomy with radiation, 90 probably will be cancer-free after 12 years while 10 will have a recurrence.

In many cases, there are not enough data even to give us that degree of certainty, or the data that exist are too ambiguous to allow a valid conclusion. In 1995, the *New England Journal of Medicine* published a report that the Nurses' Health Study (a cohort study), which had monitored 122,000 nurses for 14 years, found a 30 percent to 70 percent increased risk of breast cancer in women who had taken hormone replacement therapy after menopause.[3] One month later, the *Journal of the American Medical Association* published the results of a case-control study

that found no increased risk from the hormones. Some 500 women who had newly diagnosed breast cancer were no more likely to have taken postmenopausal hormones than a control group of 500 healthy women.[4] In *The New York Times* article reporting on the studies, each researcher is quoted suggesting possible flaws in the other study.[5] There was little comfort in these results for women seeking certainty on whether the therapy would improve their health. According to one view, postmenopausal estrogen was clearly worth the possible risk of cancer because it appeared to decrease a woman's risk of heart disease and osteoporosis. In the opposing argument, women could achieve similar benefits without the possible risk through exercise, avoiding smoking, eating a low-fat diet, maintaining a normal weight, and taking aspirin. Now, as discussed in Chapter 6, a clinical trial has contradicted some of the findings of each of these studies; hormone replacement therapy has been found to increase cancer risk and not to benefit the heart.

Contradictory results from epidemiologic studies are common. As discussed in Chapter 6, there are many possible sources of error in this kind of research, including bias and confounding, which are factors irrelevant to the hypothesis being tested that may affect a result or conclusion. Later in this chapter, additional factors to be considered in assessing whether to believe a study's conclusions are examined.

People sometimes demand certainty even when science cannot provide it, as occurred in 1997 over the issue of whether women ages 40 through 49 should be screened for breast cancer using mammography. Studies had shown that routinely testing women aged 50 and over with the breast x-rays could reduce breast cancer mortality in the population. However, studies done on younger women had not demonstrated a life-saving benefit overall for this group. Routine screening of these women increases their radiation exposure, perhaps raising their risk of cancer. It also yields many false alarms, leading to unnecessary medical testing, and major expense. The follow-up testing itself may cause complications, and many of the women remain anxious even after cancer is ruled out.[6]

When Dr. Richard Klausner, the director of the National Cancer Institute (NCI), called together a panel of experts in early 1997 to advise him on the issue, the panel concluded that, for younger women, the benefit did not justify the risks and costs, and recommended that each woman make the decision in consultation with her doctor, considering her own particular medical and family history. The public and political response was heated: after a barrage of media publicity, the Senate voted 98 to 0 to endorse a nonbinding resolution that the NCI should recommend mammography for women in their 40s. A letter signed by 39 congresswomen stated that, "without definitive guidelines, the lives of too many women are at risk to permit further delay," assuming that screening could save lives despite the lack of evidence.[7(p.1104)] In the end, director Klausner, with the support of President Clinton and Secretary of Health and Human

Services Donna Shalala, recommended that women in their forties should be screened. It seems clear that pressure from politicians eager to get credit for supporting women's health led to a pretense of scientific certainty where none existed.

On this question, further analysis supported the politicians, although the benefit is weaker for the younger age group. While the "melee that followed the meeting will not qualify for a place in the history of public health's most distinguishing scientific or policy moments," in the words of one analyst, there is now a far better understanding of the issue and evidence that screening may be life-saving for some younger women.[8(p.331)] However, because the incidence of breast cancer is lower in women in their 40s, and the effectiveness of mammography is also lower in the denser breasts of the younger women, the benefit of screening is less for them. In a review of the evidence published in 2007, the conclusion seems to echo the NCI's original recommendation that individual women, in consultation with their doctor, should decide whether to be screened. The authors suggest that, "a woman 40 to 49 years old who had a lower-than-average risk for breast cancer and higher-than-average concerns about false positive results might reasonably delay screening. Measuring risks and benefits accurately enough to identify these women remains a challenge."[9(p.522)]

Remarkably, the whole political uproar was repeated in 2009, when an independent panel of experts, appointed by the Department of Health and Human Services, issued a recommendation that routine breast cancer screening begin at age 50, not 40. Because the recommendation was published in the midst of the public debate over health-care reform, conservative politicians cried "rationing." As science reporter Gina Kolata pointed out in a *New York Times* article, the dispute gives many people "a sense of déjà vu."[10] The data hadn't changed much since the earlier debate, except that new evidence was published in 2008 suggesting that some invasive breast cancers may spontaneously regress, supporting the argument that screening may lead to unnecessary treatment.

Many people concerned about how to protect their health find it frustrating when today's news seems to contradict yesterday's. As this example shows, science is a work in progress. In the words of Dr. Arnold Relman, former editor of *The New England Journal of Medicine*, "Most scientific information is of a probable nature, and we are only talking about probabilities, not certainty. What we are concluding is the best opinion at the moment, and things may be updated in the future."[11(p.11)]

Probability

Scientists quantify uncertainty by measuring probabilities. Since all events, including all experimental results, can be influenced by chance, probabilities are used to describe the variety and frequency of past outcomes under similar conditions as a way of predicting what should happen in the future. Aristotle said that, "the probable is what usually happens." Statisticians know that the improbable happens more often than most people think.[11(p.19)]

One concept scientists use to express the degree of probability or improbability of a certain result in an experiment is the *p* value. The *p* value expresses the probability that the observed result could have occurred by chance alone. A *p* value of 0.05 means that if an experiment were repeated 100 times, the same answer would result 95 of those times, while 5 times would yield a different answer. If a person tosses a coin 5 times in a row, it is improbable that it will come up the same—heads or tails—every time. However, if each student in a class of 16 conducts the experiment, it is probable that one student will get the identical result in all 5 tosses. The probability of that occurrence is 1 chance in 16, or 0.0625 ($p = 0.0625$). Thus a *p* value of 0.05 says that the probability that an experimental result occurred by chance alone is less than the probability of tossing 5 heads or 5 tails in a row. A *p* value of 0.05 or less has been arbitrarily taken as the criterion for a result to be considered statistically significant.

Another way to express the degree of certainty of an experimental result is by calculating a *confidence interval.* This is a range of values within which the true result probably falls. The narrower the confidence interval, the lower the likelihood of random error. Confidence intervals are often expressed as margins of error, as in political polling, when a politician's support might be estimated at 50±3 percent. The confidence interval would be 47 percent to 53 percent.[11]

While *p* values and confidence intervals are useful concepts in deciding how seriously to take an experimental result, it is wrong to place too much confidence in an experiment just because it yields a low *p* value or a narrow confidence interval. There may be up to 10,000 clinical trials of cancer treatment under way at any time. If a *p* value of 0.05 is taken to imply statistical significance, 5 out of every 100 ineffective treatments would appear to be beneficial, errors caused purely by chance.[11] Thus, large numbers of cancer treatments could be in clinical use that are actually not effective. Other reasons that a low-*p*-value study could lead to an erroneous conclusion could be bias or confounding, which are systematic errors. The results of the study discussed in Chapter 6 that linked coffee-drinking with pancreatic cancer were statistically significant with a *p* value of 0.001.[12] The conclusion is thought to be wrong not because of random error but because the cancer was caused by smoking rather than coffee drinking.[13]

The fact that the probable is not always what happens leads to the Law of Small Probabilities.[11] The most improbable things are bound to happen occasionally, like throwing heads five times in a row, or even—very rarely—99 times. This means, for example, that a few people with apparently fatal illnesses will inexplicably recover. They may be convinced that their recovery was caused by something they did, giving rise—if their story is publicized—to a new vogue in quack therapies. But because their recovery was merely a random deviation from the probable, other patients will not get the same benefit.

Another consequence of the Law of Small Probabilities is the phenomenon of cancer clusters. Every now and then a community will discover that it is the site of an unusual concentration of some kind of cancer, such as childhood leukemia, and everyone will be highly alarmed. Is there a carcinogen in the air or the drinking water that is causing the problem? Could the cause be electromagnetic fields, which residents blamed for the cluster of six cases of childhood cancer between 1981 and 1988 among the pupils of an elementary school in Montecito, California?[14] Under great political pressure, the local and state government will investigate, but no acceptable explanation will be found. In the case of the electromagnetic fields, it could not be proven that they were *not* responsible for the cluster, but as more studies are done the evidence is still ambiguous. Most such clusters are due to statistical variation, like an unusual run of tails in a coin toss. Such an explanation tends to be unsatisfactory to community residents, who may accuse the government of a cover-up; but after the investigation the number of new cases usually returns to more or less normal levels, and the sense of alarm subsides.

If a cluster is very large, it is likely not to be a random variation—just as in coin tossing, fifty heads in a row is a much less likely outcome than five heads unless there is something wrong with the coin. A large number of cases is said to confer *power* on a study. Power is the probability of finding an effect if there is, in fact, an effect. Thus, an epidemiologic study that includes large numbers of subjects is more powerful than a small study, and the results are more likely to be valid, although systematic errors due to bias or confounding can be present in even the largest studies.

In designing studies of any kind, statisticians can calculate the size of the study population necessary to find an effect of a certain size if it exists. Studies with low power are likely to produce *false-negative* results (i.e., to find no effect when there actually is one). *False-positive* results occur when the study finds an effect that is not real (e.g.., when a random variation appears to be a true effect). In a study of epidemiologic studies, a statistician examined the power of each of 71 clinical trials that reported no effect. He concluded that 70 percent of the studies did not have enough patients to detect a 25 percent difference in outcome between the experimental group and the control group. Even a 50 percent difference in outcome would have been undetectable in half of the studies.[11] This common weakness in epidemiologic studies is probably one reason for the contradictory results so often reported in the news.

In the review of high dose chemotherapy and bone marrow transplant for advanced breast cancer, described in Chapter 6, the authors addressed the question of whether the studies had enough power to detect a 15 percent improvement in survival for the treated women. They concluded that the individual studies did not have sufficient power, but that the systematic review of all studies combined did have the power to detect a 15 percent difference if it existed. No difference was detected at a 95 percent confidence level.[15] They noted, however, that a smaller difference in outcomes might still be possible and raised the question of how much difference would be clinically relevant. Would it be acceptable for a woman to undergo the arduous treatment if her chance of survival was only 10 percent better? That is a question that cannot be answered by statisticians.

The Statistics of Screening Tests

In public health's mission to prevent disease and disability, secondary prevention—early detection and treatment—plays an important role. When the causes of a disease are not well understood, as in breast cancer, little is known about primary prevention. The best public health measure is to screen the population at risk so as to detect the disease early, when it is most treatable. Screening is also an important component of programs to control HIV/AIDS by identifying HIV-infected individuals so that they can be treated and counseled about how to avoid spreading the virus to others. As discussed in Chapter 12, newborn babies are routinely screened for certain congenital diseases that can be treated before permanent damage is done to the infants' developing brains and bodies.

While laboratory tests to be used in screening programs should ideally be highly accurate, most are likely to yield either false positives or false negatives. Tests may be highly *sensitive*, meaning that they yield few false negatives, or they may be highly *specific*, meaning that they yield few false positives. Many highly sensitive tests are not very specific and vice versa. For most public health screening programs, sensitive tests are desirable in order to avoid missing any individual with a serious disease who could be helped by some intervention. However, inexpensive, sensitive tests chosen to encourage testing of as many at-risk individuals as possible are often not very specific. When a positive result is found, more specific tests are then conducted to determine if the first finding was accurate. For example, if a sensitive mammogram finds a suspicious spot in a woman's breast, the test is usually followed up with a biopsy to determine whether the spot is indeed cancerous.

When screening is done for rare conditions, the rate of false positives may be as high as or higher than the number of true positives, leading to a lot of follow-up testing on perfectly normal people. Such a situation occurred in 1987 when the states of Illinois and Louisiana mandated premarital screening for HIV.[16] With the rate of HIV infection in the general,

heterosexual population quite low, a great many healthy people were unnecessarily alarmed and subjected to further tests, while very few HIV-positive people were identified. Some couples went to neighboring states to marry to avoid the nuisance. The programs were discontinued within a year. The problem of false positives is also the reason why mammography screening is questionable for women in their forties, as discussed earlier

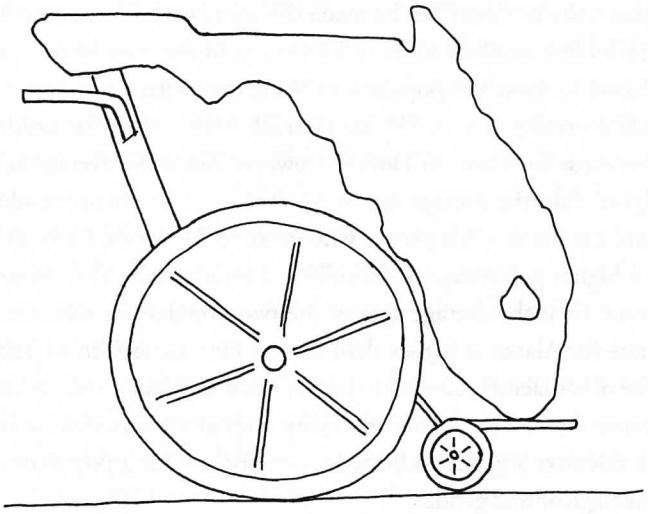
There are other conditions for which screening may not be as beneficial as expected. One of these is prostate cancer, discussed in Chapter 27. Another is lung cancer screening of smokers. Lung cancer is usually a fatal diagnosis; by the time most patients suffer symptoms, it is too late for medicine or surgery to make a difference. The idea of screening smokers so that cancers can be detected and treated earlier in the course of the disease has been around since the 1970s and 1980s. However, at that time, the only method of screening was to use chest x-rays, and it turned out that cancers detected by x-ray screening were almost always too far advanced to be treatable.

In fall 2006, a paper published in the *New England Journal of Medicine* reported that screening with spiral CT scans (a kind of three-dimensional x-ray) could detect lung cancers early enough that treatment allowed 80 percent of patients to survive for ten years, compared to a 10 percent survival rate for patients who had been diagnosed the usual way.[17] A few months later, the *Journal of the American Medical Association* published another study, concluding that spiral CT scanning does not save lives and may actually cause more harm than good.[18] An analysis of the findings of the first trial revealed two sources of bias: lead-time bias and overdiagnosis bias.[19] The former may occur in all cancer screening and must be taken into consideration before concluding that screening saves lives. Lead-time bias occurs when increased survival time after diagnosis is counted as an indicator of success. If early detection of a cancer does not lead to a cure, the only result of early diagnosis is that patients will live longer with the knowledge that they are sick before dying at the same time they would have died anyway. This appears to be the case in the *New England Journal of Medicine* study of lung cancer screening. In fact, the effects of the additional diagnostic tests and surgeries that follow the early diagnosis may hasten the patients' death.

Overdiagnosis bias occurs when the tumors that are detected by the screening are not likely to progress to the stage that they cause symptoms and be life-threatening. Such small tumors had also been found in the earlier lung cancer screening trials using x-rays. Overdiagnosis bias is also a problem with prostate cancer screening, as discussed in Chapter 27, and perhaps with breast cancer screening, as discussed earlier in this chapter. The only way to be sure that screening actually saves lives is to conduct randomized controlled trials, comparing mortality among patients who are screened with that of patients who are not screened. Such trials, together with data showing that breast cancer mortality overall has fallen in the United States by 24 percent since 1990, have shown that mammography does save lives.[8,9,20]

Rates and Other Calculated Statistics

As seen in Chapter 5, epidemiology makes extensive use of rates in studies of disease distribution and determinants. Rates put the raw numbers into perspective by relating them to the size of the population being considered. As Chapter 8 discusses, vast quantities of health-related data are collected on the American population, data that are used to assess the people's health and to evaluate the effectiveness of public health programs. For these purposes too, the raw numbers are subjected to statistical adjustments that yield various rates useful in making comparisons and identifying trends.



Florida's Population is Older than Average

For example, knowing that a city has 500 deaths per year is not very informative unless the population of the city is known. Death rates are generally expressed as the number of deaths per 1000 people. Thus, 500 deaths per year is a low number for a city of 100,000, while it is high for a city of 50,000. The overall death rate in the United States was 8.0 per 1000 people in 2007.[21] The same data may yield different rates depending on the population referred to. As discussed in Chapter 5, rates are usually calculated using the population at risk for the denominator. In the case of death rates, the whole population is at risk. Birth rates are an exception; like the death rate, the birth rate is defined as the number of live births per 1000 people. The fertility rate, by contrast, does use the population at risk, giving the number of live births per 1000 women ages 15 to 44. Two communities with the same fertility rate may have quite dif-

ferent birth rates if one contains many young women and the other is older with a higher proportion of men. Both rates start with the same raw number—the number of live births—but use a different population for reference. In 2006, the birth of 4,265,555 babies in the United States led to a birth rate of 14.2 per 1000 people overall. The fertility rate ranged from 49.3 per 1000 women of Cuban origin to 109.0 per 1000 women of Mexican origin.[22]

Other rates commonly used as indicators of a community's health are the infant mortality rate and the maternal mortality rate, as discussed in Chapter 18. The infant mortality rate is the number of infants that die before their first birthday in a year, divided by the number of live births in that year. The maternal mortality rate is the number of deaths among women associated with pregnancy and delivery in a year, divided by the number of live births in that year.

For some purposes, the numbers can be made still more useful by converting crude rates into adjusted rates. Death rates are often adjusted for the age of the population. The adjustment uses a statistical calculation to make the populations being examined equivalent to one another. For example, the crude mortality rate in Florida is much higher than the crude mortality rate in Alaska. There is no cause for alarm in Florida, however. Since the average age of the Floridians is significantly higher than the average age of Alaskans—in fact many residents of other states retire to Florida and die there, while people who move to Alaska are likely to be young—it is to be expected that a higher percentage of Floridians die each year. After adjusting the mortality rate to what it would be if the average ages of the two populations were the same, the age-adjusted mortality rate for Alaska is higher than that in Florida, as seen in **Table 7-1**. Rates may also be adjusted for other factors relevant to health, such as gender, race, ethnicity, and so forth. For example, because males have higher mortality rates at all ages than females, it may sometimes be useful to calculate a gender-adjusted mortality rate for a population that has a higher proportion than average of one gender.

Rates are also calculated on a group-specific basis. Researchers may calculate rates for males alone or females alone, blacks, whites, Hispanics, members of other racial or ethnic groups, and people in defined age groups. This kind of data informs us, for example, that males have higher mortality rates than females in the same age group, and that blacks have higher mortality rates than whites of the same sex and age. It is common to break down death rates from various causes by age group, revealing that different age groups are more likely to die of different causes. For example, death rates from cancer, stroke, and heart disease increase steadily with age. Death rates from AIDS, however, are highest for the 44 to 54 year age group and fall to almost zero for those older than 75 years.[23(Table 42)] Death rates from firearms injuries and motor vehicle injuries are highest in the 20 to 24 age group, although death rates from motor vehicle injuries are almost as high for people over 75.[23(Tables 44,47)]

Table 7-1 Age-Adjusted Mortality Rates for Florida and Alaska, 2006

	Florida	Alaska
Crude death rate per 100,000	940.1	500.6
Age-adjusted death rate per 100,000	711.3	775.9

Source: Data from *National Vital Statistics Reports* 58, No. 1, "Deaths: Final Data for 2006." National Center for Health Statistics (2009).

Further calculations can be done using age-specific death rates to yield life expectancies, data that is intuitively meaningful in describing the health of a population. Life expectancy is the average number of years of life remaining to people at a particular age, and it reflects the mortality conditions of the period when the calculation is made. Life expectancies may be determined by race, sex, or other characteristics using age-specific death rates for the population with that characteristic. The most common figure used in comparing the health of various populations is the life expectancy at birth. As seen in **Table** 7-2, life expectancies at birth in the United States have been increasing since 1900. In Russia, however, as the Prologue discusses, life expectancies have declined since the fall of the Soviet Union, reflecting many societal ills that have led to poorer health of the population there. **Table** 7-3 shows the life expectancy at birth for males and females of selected countries.

Another calculated concept that is sometimes used as a measure of premature mortality is years of potential life lost (YPLL). It gives greater weight to deaths of young people, appropriate to the priorities of public health, which has the goal not of eliminating death entirely but of enabling people to live out their natural lifespan with a minimum of illness and disability. Calculation of YPLL arbitrarily chooses 75 as the age before which a death is considered premature (age 65 was used before 1996). As an example, the death of a person 15 to 24 years of age counts as 55.5 YPLL before age 75. Unintentional injuries rank relatively high in YPLL because they are likely to kill young people, who have more years to lose. **Table** 7-4 shows a comparison of the leading causes of death in the United States with the leading causes of YPLL.

Risk Assessment and Risk Perception

While some statistical concepts may seem difficult and confusing, people have an intuitive understanding of statistics affecting their everyday lives. They understand that the future is full of uncertainties, and they intuitively try to minimize risks or at least weigh risks against expected benefits. Their intuitive judgment of risks, however, often does not coincide with the more scientific estimates of statisticians. It turns out that while judgments of risk by the

Table 7-2 Life Expectancy at Birth According to Race and Sex in the United States, Selected Years

Specified age and year	All races			White			Black or African American[1]		
	Both sexes	Male	Female	Both sexes	Male	Female	Both sexes	Male	Female
At birth	Remaining life expectancy in years								
1900[2,3]	47.3	46.3	48.3	47.6	46.6	48.7	33.0	32.5	33.5
1950[3]	68.2	65.6	71.1	69.1	66.5	72.2	60.8	59.1	62.9
1960[3]	69.7	66.6	73.1	70.6	67.4	74.1	63.6	61.1	66.3
1970	70.8	67.1	74.7	71.7	68.0	75.6	64.1	60.0	68.3
1980	73.7	70.0	77.4	74.4	70.7	78.1	68.1	63.8	72.5
1990	75.4	71.8	78.8	76.1	72.7	79.4	69.1	64.5	73.6
1995	75.8	72.5	78.9	76.5	73.4	79.6	69.6	65.2	73.9
1997	76.5	73.6	79.4	77.1	74.3	79.9	71.1	67.2	74.7
1998	76.7	73.8	79.5	77.3	74.5	80.0	71.3	67.6	74.8
1999	76.7	73.9	79.4	77.3	74.6	79.9	71.4	67.8	74.7
2000	77.0	74.3	79.7	77.6	74.9	80.1	71.9	68.3	75.2
2001	77.2	74.4	79.8	77.7	75.0	80.2	72.2	68.6	75.5
2002	77.3	74.5	79.9	77.7	75.1	80.3	72.3	68.8	75.6
2003	77.4	74.7	80.0	77.9	75.3	80.4	72.6	68.9	75.9
2004	77.8	75.2	80.4	78.3	75.7	80.8	73.1	69.5	76.3
2005	77.8	75.2	80.4	78.3	75.7	80.8	73.2	69.5	76.5

[1]Data shown for 1900–1960 are for the nonwhite population.
[2]Death registration area only. The death registration area increased from 10 states and the District of Columbia (DC) in 1900 to the coterminous United States in 1933.
[3]Includes deaths of persons who were not residents of the 50 states and DC.

Source: U.S. Centers for Disease Control and Prevention, *Health, United States*, 2008. Table 26

Table 7-3 Life Expectancy at Birth for Males and Females of Selected Countries

	Male			Female		
Country	**1980**	**2004**	**Rank**	**1980**	**2004**	**Rank**
Australia	71.0	78.1	5	78.1	83.0	6
Austria	69.0	76.4	16	76.1	82.1	13
Belgium	70.0	75.6	20	76.8	81.6	16
Bulgaria	68.5	69.1	32	73.9	76.3	33
Canada	71.7	77.8	7	78.9	82.6	8
Chile	—	74.0	28	—	80.6	24
Costa Rica	71.9	76.4	16	77.0	80.7	3
Cuba	72.2	75.4	21	—	79.8	27
Czech Republic	66.8	72.6	29	73.9	79.0	30
Denmark	71.2	75.2	23	77.3	79.9	26
England and Wales	70.8	76.8	13	78.8	81.1	20

(continues)

Table 7-3 *(Continued)*

Finland	69.2	75.3	22	77.6	82.3	9
France	70.2	76.7	14	78.4	83.8	3
Germany	69.6	75.7	18	76.1	81.4	18
Greece	72.2	76.6	15	76.8	81.5	16
Hong Kong	71.6	79.0	1	77.9	84.7	2
Hungary	65.5	68.6	33	72.7	76.9	32
Ireland	70.1	75.2*	—	75.6	80.3*	—
Israel	72.2	77.9	6	75.8	82.2	12
Italy	70.6	76.8*	—	77.4	82.9*	—
Japan	73.4	78.6	2	78.8	85.6	1
Netherlands	72.8	76.9	12	79.2	81.4	18
New Zealand	70.0	77.5	8	76.3	81.7	15
Northern Ireland	68.3	76.0	18	75.0	80.8	22
Norway	72.3	77.5	8	79.2	82.3	9
Poland	66.0	70.7	30	74.4	79.2	29
Portugal	67.7	74.5	25	75.2	81.0	21
Puerto Rico	70.8	74.1	27	76.9	82.3	9
Romania	66.6	68.3	34	71.9	75.	34
Russian Federation	61.4	59.1	35	73.0	72.4	35
Scotland	69.0	74.2	26	75.2	79.3	28
Singapore	69.8	77.1	11	74.7	82.0	14
Slovakia	66.8	70.3	31	74.3	77.8	31
Spain	72.5	77.2	10	78.6	83.7	4
Sweden	72.8	78.4	4	78.8	82.7	7
Switzerland	72.8	78.6	2	79.6	83.7	4
United States	70.0	75.2	23	77.4	80.4	25

*2002 Data

Source: Data from U.S. Centers for Disease Control and Prevention, *Health, United States*, 2008. Table 25

average person include statistical estimates that are often fairly accurate, they are also influenced by psychological factors that should perhaps be taken into consideration by the public health professionals.

Public health's mission to protect the population from disease and injury requires governments to minimize risks or at least weigh risks against expected benefits, just as individuals do in their own lives. The formal process of risk assessment identifies events and exposures that may be harmful to humans and estimates the probabilities of their occurrence as well as the extent of harm they may cause.

Risk assessment is often done on the basis of historical data: for example, one may predict that the number of motor vehicle crashes next year will be similar to the number this year, increasing or decreasing according to the trend established over the past several years. Risks that certain chemicals cause cancer in humans are usually estimated by analogy with data obtained from animal studies. For many situations, however, there is little basis on which to make comparisons. In such cases, assessing risks involves making many assumptions, some of which may be little better than guesses. To estimate the probability of a mishap in a new technology, various possible chains of events are considered, and a risk for something going wrong is estimated for each step, perhaps by analogy with conventional technology. Risks of the individual steps are then added or multiplied to obtain a risk for the whole. This approach was used, for exam-

Table 7-4 Years of Potential Life Lost (YPLL) before Age 75 by Cause of Death and Rank, 2005

Cause of Death	YPLL	Rank by YPLL	Rank by No. of Deaths
Cancer	1525.2	1	2
Unintentional injuries	1132.7	2	5
Heart disease	1110.4	3	1
Suicide	347.3	4	11
Homicide	276.8	5	15
Cerebrovascular disease	193.5	6	3
Chronic lower respiratory disease	181.2	7	4
Diabetes	179.9	8	6
Chronic liver disease and cirrhosis	152.6	9	12
HIV disease	133.6	10	17
Influenza and pneumonia	83.6	11	8

Source: U.S. Centers for Disease Control and Prevention, *Health, United States*, 2008, Table 29 and National Vital Statistics Report: Final Data for 2005.

ple, when nuclear power plants were first introduced, and it helped engineers to identify what kind of safety devices should be incorporated to reduce the probability of failure.[24] Still, the assessment appears to have underestimated the risk at Three Mile Island, as discussed later in this section.

Using such methods, scientists calculate probabilities that various injurious events will occur and rank them in order, as shown in **Table** 7-5. According to an analysis published in 1987, experts said that the most risky activities and technologies were motor vehicles, smoking, alcoholic beverages, handguns, and undergoing surgery. When the representatives of the general public were asked for their perceptions of risks, however, they headed their list with nuclear power, which was ranked 20th by the experts. Other risks that people tend to rank higher than the experts do are electromagnetic fields, genetic engineering, and radioactive waste.[25]

As a result of the apparent irrationality of the public in response to risks that the experts estimated to be small, a field of study has developed concerning risk perception. While experts assess risk on the basis of expected mortality as predicted from historical data, the general public includes other considerations in its assessments. When these additional criteria are analyzed, it appears that the public's perception may not be so irrational after all.

Risk perception researchers have found that people's concern about a risk is affected by certain associated factors. For example, familiar risks are more acceptable than unfamiliar ones. Risks that people perceive they have control over are more acceptable than those that are uncontrollable. A risk with potentially catastrophic consequences is unacceptable, even if it is highly unlikely to occur. People are more likely to accept a risk from an activity that is perceived as beneficial, but they want the risks and benefits to be distributed equitably.

Risk perception researchers classify risks on two scales: dread and knowability. The more dreaded the risk, the less acceptable it is; similarly, unknown risks are less acceptable than known risks. **Figure** 7-1 maps various risks according to the concern they evoke on the two scales. Thus although driving an automobile is, statistically, one of the most risky activities, it does not arouse great anxiety because it is neither dreaded nor unknown. Moreover, people perceive that they have control when they are driving, and the benefit is obvious to them. Conversely, a nuclear reactor accident is highly dreaded, and thus is perceived by the public as more risky than the experts believe it to be. People perceive that they lack control over nuclear reactors, and the benefits of nuclear power may not be clear to people who live in their vicinity.

The public's perception about nuclear power gained credibility after the 1979 accident at the Three Mile Island nuclear reactor in Pennsylvania. According to the experts, numerous safeguards were in place to prevent an accident, and the chance of a serious breakdown was remote. In fact, the safety systems worked to the extent that there was no disaster; no one was killed, and there was no significant radiation leak. Nevertheless, the fact that the breakdown occurred at all sent a signal that the experts may have underestimated the risks. Public opposition to nu-

Table 7-5 Ordering of Perceived Risk for 30 Activities and Technologies

Activity or Technology	League of Women Voters	College Students	Experts
Nuclear power	1	1	20
Motor vehicles	2	5	1
Handguns	3	2	4
Smoking	4	3	2
Motorcycles	5	6	6
Alcoholic beverages	6	7	3
General (private) aviation	7	15	12
Police work	8	8	17
Pesticides	9	4	8
Surgery	10	11	5
Fire fighting	11	10	18
Large construction	12	14	13
Hunting	13	18	23
Spray cans	14	13	26
Mountain climbing	15	22	29
Bicycles	16	24	15
Commercial aviation	17	16	16
Electric power (nonnuclear)	18	19	9
Swimming	19	30	10
Contraceptives	20	9	11
Skiing	21	25	30
X-rays	22	17	7
High school and college football	23	26	27
Railroads	24	23	19
Food preservatives	25	12	14
Food coloring	26	20	21
Power mowers	27	28	28
Prescription antibiotics	28	21	24
Home appliances	29	27	22
Vaccinations	30	29	25

The ordering is based on the geometric mean risk ratings within each group. Rank 1 represents the most risky activity or technology.

Source: Reprinted with permission from P. Slovic, "Perception of Risk," *Science* 236: 281. Copyright 1987, AAAS.

clear power increased dramatically, and stricter requirements for reactor safety were imposed, raising construction and operating costs. The coup de grace for nuclear power came with the 1986 reactor meltdown at Chernobyl, in the Ukraine, which did cause lost lives and widespread radioactive contamination of the environment. Since then, no new nuclear reactors have opened in the United States.[25] With the recent concerns about fossil fuels' causing climate change, perceptions of the risk–benefit balance for nuclear power may change.

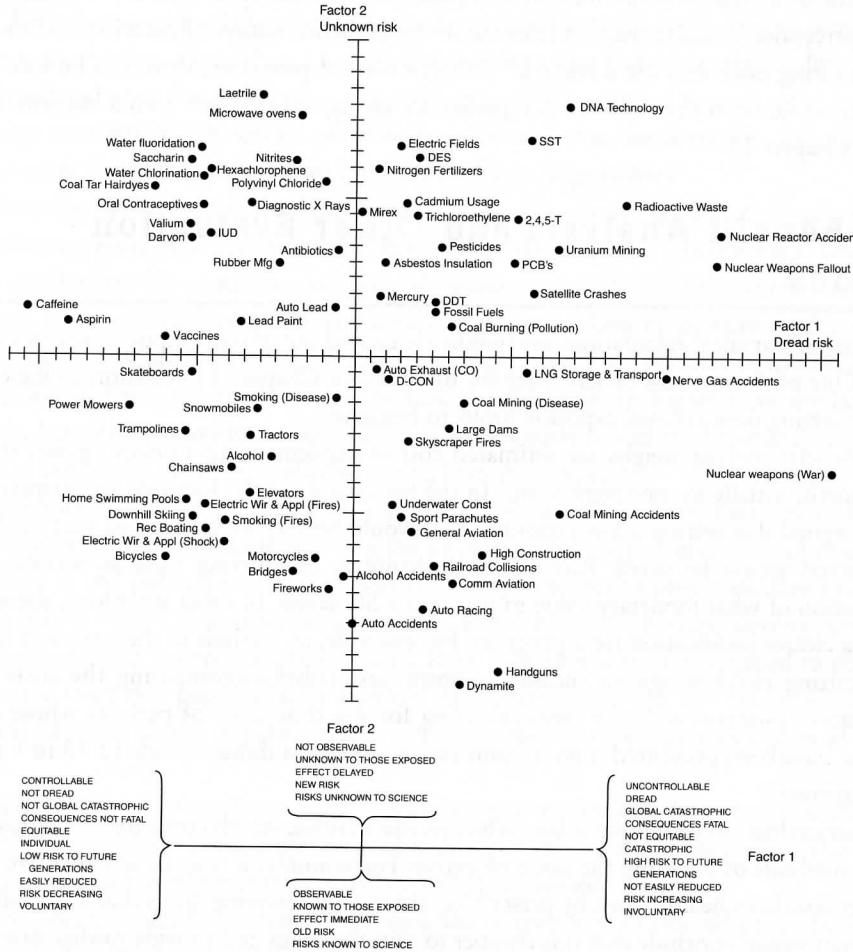


FIGURE 7-1 Location of 81 Hazards on Factors 1 and 2. *Source:* P. Slovic et al. in *Perilous Progress: Managing the Hazards of Technology*, (p. 108) by R. W. Kates et al., ed.

An interesting example of anomalous risk perception—one that is of great relevance to public health—is the paradox that adolescents so often engage in activities that they "know" to be dangerous, such as smoking, drunk driving, drug use, and unprotected sex. Studies aimed at understanding why teens engage in health-threatening behaviors can help to design interventions to prevent such behaviors. In the case of smoking, for example, surveys have shown that teenagers can fairly accurately predict the probability that smokers will die of lung cancer and other diseases. However, the same surveys have found that teenage smokers perceive themselves to be at little or no risk. It turns out that they plan to quit smoking in the next few years, an inaccurate perception because they underestimate the addictive nature of nicotine and the difficulty of quitting once they are addicted.[26] Tobacco control programs, then, can be focused on convincing adolescents that tobacco companies are trying to lure them into addiction, as discussed in Chapter 15.

Cost–Benefit Analysis and Other Evaluation Methods

Other types of statistical calculations are frequently carried out as part of public health decision making. One of these is cost benefit analysis, discussed in Chapter 3 in relation to the controversy over setting occupational exposure limits to benzene.

Cost–benefit analysis weighs the estimated cost of implementing a policy against the estimated benefit, usually in monetary terms. In the benzene example discussed in Chapter 3, the industry argued that setting a low exposure limit would be very expensive and that the benefit in lives saved would be small. Part of the difficulty in conducting such an analysis is the determination of what monetary value to place on a life saved. In other situations, the analysis provides a clearer justification for a program. For example, an analysis of the costs and benefits of immunizing children against measles, mumps, and rubella—comparing the costs of the immunization program with the costs of caring for the thousands of patients whose disease would not have been prevented if no immunizations had been done—yielded a 13 to 1 ratio of benefits to costs.[27]

Another evaluation technique is cost-effectiveness analysis, which compares the efficiency of different methods of attaining the same objective. For example, it may be so expensive to prevent heart attacks in healthy men by prescribing cholesterol-lowering drugs that a cost-effectiveness analysis would conclude that it is cheaper to skip the drugs and provide cardiac care for the men who do suffer an attack. Cost–benefit analysis and cost-effectiveness analysis "cannot serve as the sole or primary determinant of a health care decision," according to a congressional report, but the process of identifying and considering all the relevant costs and benefits can improve decision making.[28(p.211)]

Conclusion

The world is full of uncertainty. Science may not always be able to provide answers to people's questions. Statistics is a way to learn, at least, how certain people can be about what they think they know.

Statistics is a tool widely used in public health. Most epidemiologic studies and most studies in the other public health sciences depend on statistics to analyze data and interpret findings. Statistical analyses can establish the probability that what was observed occurred by chance alone. A measure commonly used to indicate the probability that a study finding is the result of chance is the *p* value. Even when a low *p* value indicates that a result is statistically significant, there is still a chance that the result is not valid, even if all sources of bias are ruled out. Studies with large numbers of subjects are more likely to be valid than small studies, although sources of error other than random variation are still possible in large studies.

Knowledge of statistics is also important in evaluating screening tests, used as a secondary prevention approach to detect diseases so that they can be treated at an early stage. Tests that are highly sensitive tend to yield false positives, while tests that are highly specific tend to yield false negatives. Most screening programs use sensitive tests and follow up positive results with more expensive tests that are both highly sensitive and highly specific. For conditions that are rare in the population being screened, the rate of false positives may be higher than the rate of true positives. Screening programs are also subject to biases, such as lead-time bias and overdiagnosis bias, that may make them less useful for saving lives than expected.

To put numbers into perspective, they are often converted into rates. Rates are useful in epidemiology and as a way of understanding the importance of the vast quantities of data used for assessment of the public's health and evaluation of public health programs. Rates commonly used as public health indicators are mortality (death) rates, birth rates, fertility rates, infant mortality rates, and maternal mortality rates. Rates may be statistically adjusted to make them comparable from one population to another. Age-specific rates can also be calculated. Other statistical concepts useful as public health indicators are life expectancy and years of potential life lost.

Public health's efforts to protect the population may require calculations of risk. Risk assessment is a formal process of calculating probabilities of various injurious events. The scientific assessment of risk sometimes conflicts with people's perception of risk.

Public health is based on science, including the science of statistics, which is the science of uncertainty. To paraphrase statistician and author Robert Hooke, scientific studies are often the only way to answer people's questions, but the studies do not produce "unassailable, universal truths that should be carved on stone tablets." Instead, they produce statistics, which must be interpreted.[11(p.64)]

References

1. R. Royce et al., "Sexual Transmission of HIV," *New England Journal of Medicine* 336 (1997): 1072–1078.
2. B. Fisher et al., "Reanalysis and Results after 12 Years of Follow-Up in a Randomized Clinical Trial Comparing Total Mastectomy with Lumpectomy with or without Irradiation in the Treatment of Breast Cancer," *New England Journal of Medicine* 333 (1995): 1456–1467.
3. G. A. Colditz et al., "The Use of Estrogens and Progestins and the Risk of Breast Cancer in Postmenopausal Women," *New England Journal of Medicine* 332 (1995): 1589–1593.
4. J. L. Stanford et al., "Combined Estrogen and Progestin Hormone Replacement Therapy in Relation to Risk of Breast Cancer in Middle-Aged Women," *Journal of the American Medical Association* 274 (1995): 137–142.
5. G. Kolata, "Cancer Link Contradicted by New Hormone Study," *The New York Times*, July 12, 1995.
6. H. C. Sox, "Benefit and Harm Associated with Screening for Breast Cancer," *New England Journal of Medicine* 338 (1998): 1145–1146.
7. V. L. Ernster, "Mammography Screening for Women Aged 40 through 49—A Guidelines Saga and a Clarion Call for Informed Decision Making," *American Journal of Public Health* 87 (1997): 1103–1106.
8. R. A. Smith, "Breast Cancer Screening among Women Younger than 50: A Current Assessment of the Issues," *CA-A Cancer Journal for Clinicians* 50 (2000): 312–336.
9. K. Armstrong et al., "Screening Mammography in Women 40 to 49 Years of Age: A Systematic Review for the American College of Physicians," *Annals of Internal Medicine* 146 (2007): 516–526.
10. G. Kolata, "Get a Test, No Don't. Repeat," *The New York Times*, November 22, 2009.
11. V. Cohn and L. Cope, *News and Numbers: A Guide to Reporting Statistical Claims and Controversies in Health and Other Fields*, 2nd ed. (Ames: Iowa State University Press, 2001), 11.
12. B. MacMahon et al., "Coffee and Cancer of the Pancreas," *New England Journal of Medicine* 304 (1981): 630–633.
13. G. Taubes, "Epidemiology Faces Its Limits," *Science* 269 (1995): 164–169.
14. P. Brodeur, *The Great Power-Line Cover-Up: How the Utilities and the Government Are Trying to Hide the Cancer Hazards Posed by Electromagnetic Fields* (Boston: Little, Brown, 1993).
15. C. Farquhar et al., "High-Dose Chemotherapy and Autologous Bone Marrow or Stem Cell Transplantation versus Conventional Chemotherapy for Women with Metastatic Breast Cancer," *Cochrane Database of Systematic Reviews* 2005, Issue 3, www.cochrane.org/reviews/en/ab003142.html, last assessed as up-to-date March 16, 2007, accessed January 31, 2010.
16. R. Bayer, *Private Acts, Social Consequences: AIDS and the Politics of Public Health* (New York: The Free Press, 1989).
17. International Early Lung Cancer Action Program Investigators, "Survival of Patients with Stage I Lung Cancer Detected on CT Screening," *New England Journal of Medicine* 355 (2006): 1763–1771.

18. P. B. Bach et al., "Computed Tomography Screening and Lung Cancer Outcomes," *Journal of the American Medical Association* 297 (2007) 953-961.
19. H.G. Welch et al., "How Two Studies on Cancer Screening Led to Two Results," *The New York Times*, March 13, 2007.
20. D.A. Berry et al., "Effect of Screening and Adjuvant Therapy on Mortality from Breast Cancer," *New England Journal of Medicine* 353 (2005): 1784–1792.
21. U.S. Centers for Disease Control and Prevention, "Births, Marriages, Divorces, and Deaths: Provisional Data for 2007," *National Vital Statistics Report* 56, No. 21, www.cdc.gov/nchs-data/nvsr/nvsr56/nvsr56_21.htm, last modified October 15, 2008, accessed October 18, 2008.
22. J. A. Martin et al., "Births: Final Data for 2006," *National Vital Statistics Report* 57, No. 7, www.cdc.gov/nchs/data/nvsr/nvsr57/nvsr57_07.pdf, January 7, 2009, accessed January 31, 2010.
23. U.S. Centers for Disease Control and Prevention, *Health, United States*, 2007, www.cdc.gov/nchs/data/hus/hus07.pdf, November 2007, accessed January 31, 2010.
24. R. Wilson and E. A. C. Crouch, "Risk Assessment and Comparisons: An Introduction," *Science* 236 (1987): 267–270.
25. P. Slovic, "Perception of Risk," *Science* 236 (1987): 280–285.
26. P. Slovic, "Do Adolescent Smokers Know the Risks?" in P. Slovic, *The Perception of Risk* (London: Earthscan Publications Ltd., 2000), 363–371.
27. C. C. White et al., "Benefits, Risks and Costs of Immunization for Measles, Mumps and Rubella," *American Journal of Public Health* 75 (1995): 739.
28. G. Pickett and J. J. Hanlon, *Public Health: Administration and Practice* (St. Louis, Missouri: Times Mirror/Mosby, 1990), 211.