Teaching Elementary Probability and Statistics: Some Applications in Epidemiology

Hardeo Sahai

Department of Biostatistics and Epidemiology
University of Puerto Rico
San Juan, Puerto Rico 00936

Michael R. Reesal

Social and Preventive Medicine
University of Saskatchewan
Prince Albert, Saskatchewan S6V 691
Canada

Introduction

Teaching probability and statistics with practical applications in other disciplines is one of the best pedagogical tools for heightening interest in mathematics. By introducing applications of probability and statistics in nonmathematical contexts, one can motivate students with limited initial interests in mathematics to become increasingly familiar with important mathematical tools. Many useful and often exciting applications of probability and statistics emerge by analyzing realistic world problems.

This article illustrates some common applications of probability and statistics in the field of epidemiology as they may be presented to an undergraduate class in probability and statistics.

Epidemiology is the branch of medical science that attempts to discover associations between events, patterns and, hopefully, the cause of disease in human populations. The word epidemic originally referred to outbreaks of contagious diseases in humans; it is derived from the Greek word *epi-demos* (επι–δεμος) meaning upon-people. Epidemiologists frequently work with rates, proportions, and other quantitative measures of occurrence, prevalence, and causes of disease where the concepts and principles of elementary probability and statistics can be applied. Some common problems in epidemiologic studies are described where techniques of elementary probability and statistical inference yield results of considerable interest.

Measures of Association of Disease and Risk Factors: Use of Conditional Probability

Discovering the relationship between environmental factors and the occurrence of fatal diseases is an important objective in epidemiology. During its early development, epidemiology was concerned mainly with epidemics of disease, infection rates, and problems related to the spread and control of disease. More recently the application of epidemiologic methods has resulted in many important discoveries, such as the relationship between smoking and lung cancer; between air pollution, smog, and chronic bronchitis; between fluoride deficiency in drinking water and dental care in children.

The use of epidemiologic methods can be illustrated by studying the relationship between smoking and lung cancer. To do this, consider Table 1.

Table 1

Distribution of Lung Cancer Cases in a Population of Smokers and Nonsmokers

	Smokers	Nonsmokers	Total
With lung cancer	A	B	A+B
Without lung cancer	C	D	C+D
Total	A+C	B+D	A+B+C+D

Cohort vs. Case-Control Study

In order to study a population of smokers and nonsmokers, with or without lung cancer, epidemiologists work with samples derived from a population of interest. Two approaches for taking a sample are in common use. In the first approach, a cohort study, two samples are taken: (a) one from a smoking group (A + C) and (b) another from a nonsmoking group (B + D).

In the second approach, the case-control study, one sample comes from the group with lung cancer (A + B), and the other from a group without lung cancer (C + D).

Consider Table 2 which gives the distribution of subjects in different categories in samples from the population in Table 1. In a cohort study, the marginal totals (a + c) and (b + d) are fixed at the time of sampling while in a case-control study the marginal totals (a + b) and (c + d) are fixed.

Relative Risk

An important concept in epidemiology is the relative risk. To motivate the definition of relative risk in a cohort study, the probability that any individual has a lung cancer given he or she is a smoker is:

$$P(\text{Cancer/Smoker}) = a/(a + c)$$

Table 2

Distribution of Lung Cancer Cases in Samples of Smokers and Nonsmokers

	Smokers	Nonsmokers	Total
With lung cancer	a	b	a+b
Without lung cancer	c	d	c+d
Total	a+c	b+d	a+b+c+d

Similarly, the probability that an individual has lung cancer given that he or she is a nonsmoker is:

$$P(\text{Cancer/Nonsmoker}) = b/(b + d)$$

The relative risk (RR) is defined as the ratio of the two probabilities:

$$\frac{P(\text{Cancer/Smoker})}{P(\text{Cancer/Nonsmoker})} = \frac{a/(a + c)}{b/(b + d)} = \frac{a(b + d)}{b(a + c)} \quad (1)$$

This is the relative risk of getting lung cancer for a smoker compared to a nonsmoker.

In a case-control study, however, the marginal totals (a + c) and (b + d) are not fixed at the time of sampling. Thus, the relative risk as defined in (1) cannot be calculated. The relative risk of being exposed to smoking for cases with lung cancer compared to those without lung cancer, however, can be calculated. As earlier, this relative risk is defined by:

$$\frac{P(\text{Smoker/Cancer})}{P(\text{Smoker/No Cancer})} = \frac{a/(a + b)}{c/(c + d)} = \frac{a(c + d)}{c(a + b)} \quad (2)$$

Odds Ratio

Another important quantity commonly used to measure the association risk of a disease is the odds ratio. An odds ratio is based on the same concept as the odds of an event. In many probability applications, odds are commonly used to express the probability of an event. Rather than saying that the probability of an event is 0.3, the odds for the occurrence of the event are 3 to 7. To motivate the definition of odds ratio, note that in a cohort study, the odds for getting lung cancer for a smoker are a/(a + c) to c/(a + c). Similarly, the odds for getting lung cancer for a nonsmoker are b/(b + d) to d/b + d). The odds ratio (OR) is defined as the ratio of the two odds:

$$\frac{P(\text{Cancer/Smoker}) + P(\text{No Cancer/Smoker})}{P(\text{Cancer/Nonsmoker}) + P(\text{No Cancer/Nonsmoker})} =$$

$$\frac{(a)/(a + c) + (c)/(a + c)}{(b)/(b + d) + (d)/(b + d)} = \frac{ad}{bc} \quad (3)$$

Odds ratio is also known as the cross-product ratio since it is the ratio of the products (ad) and (bc) of entries from cells that are diagonally opposite.

As seen earlier, in a case-control study, the probability or odds of getting a disease given the smoking status of a person cannot be calculated; however, one can calculate the odds of smoking or not smoking given the status of lung cancer. Thus, the odds of smoking for a person having lung cancer are a/(a + b) to b/(a + b). Similarly, the odds of smoking for a person without lung cancer are c/(c + d) to d/(c + d). Hence, as earlier, the odds ratio is:

$$\frac{P(\text{Smoker/Cancer}) + P(\text{Nonsmoker/Cancer})}{P(\text{Smoker/No Cancer}) + P(\text{Nonsmoker/No Cancer})}$$

$$\frac{(a)/(a + b) + (b)/(a + b)}{(c)/(c+ d) + (d)/(c+ d)} = \frac{ad}{bc} \quad (4)$$

It is interesting to verify that the other odds ratio--relating the odds for having lung cancer for a smoker versus nonsmoker--is also equal to (ad)/(bc). Thus, one of the interesting properties of the odds ratio is that it is calculated the same way for both cohort and case-control studies. Further, it is readily shown that for rare diseases, when the probability that the exposed group will develop the outcome is very small, the odds ratio closely approximates the relative risk. For example, if the incidence of a particular disease in a group which is exposed to some condition is 30 per 100,000 compared with an incidence of 10 per 100,000 in a group which is not exposed to the condition, then the relative risk for exposed versus unexposed is:

$$\frac{30/100{,}000}{10/100{,}000} = 3$$

The odds ratio for exposed versus unexposed is:

$$\frac{30/99{,}970}{10/99{,}999} = 3.0006$$

This is fairly close to the relative risk. This property of odds ratio is extremely appealing to epidemiologists since they prefer to work with inexpensive case-control studies of short duration which can yield almost the same relative risk as obtained by an expensive cohort study of long duration.

Numerical Example

Consider Table 3 which classifies the frequency of smoking and lung cancer in a sample of 100,000 people in a community. The relative risk of lung cancer for smokers compared to nonsmokers is:

$$\frac{120/20{,}000}{10/80{,}000} = 48.00$$

Table 3

Distribution of Lung Cancer Cases in a Sample of 100,000 People in a Community

	Smokers	Nonsmokers	Total
With lung cancer	120	10	130
Without lung cancer	19,880	79,990	98,870
Total	20,000	80,000	100,000

Thus, a smoker has a 48 times greater risk of lung cancer than a nonsmoker. The odds ratio of lung cancer for smokers compared to nonsmokers is:

$$\frac{(120/20{,}000) + (19{,}880/20{,}000)}{(10/80{,}000) + (79{,}990/80{,}000} =$$

$$\frac{120 \times 79{,}990}{10 \times 19{,}880} = 48.28$$

Thus, the odds for lung cancer are 48.28 times greater for a smoker that for a nonsmoker. In general, RR or OR need not be in such close agreemen;, though, as seen earlier, for rare diseases, OR closely approximates RR.

It should be noted that the results in Table 3 could have arisen either from a cohort or a case-control study. In either case, the table would look the same. If the results were obtained from a case-control study, the relative risk of lung cancer for smokers to nonsmokers could not be calculated. Since in that case, the marginal totals 20,000 and 80,000 would not have been fixed and would vary from sample to sample. The odds ratio, however, would be calculated in the same way for both cohort and case-control studies. For a detailed discussion of the subject including results on statistical estimation and real-life applications to AIDS, see Gastwirth (1987).

Prevalence and Incidence: Probability as Relative Frequency

In epidemiology, the terms prevalence and incidence are used to denote probabilities in a special context. The prevalence of a disease is the probability of currently having the disease regardless of the duration of time one has had the disease. To illustrate the computation of prevalence, let D be the event that a randomly chosen person in a certain population has the disease. Then, using the relative frequency definition of probability, the prevalence is calculated as:

$$\frac{\text{The number of people with the disease}}{\text{Total number of people in the population}}$$

The incidence of a disease is the probability of developing a new case of the disease during some specified time period among all people who did not have the disease at the beginning of the time interval. To illustrate the computation of the incidence, let I be the event that a randomly chosen person in a certain population develops a new case of the disease. Then using the relative frequency definition of probability, the incidence is calculated as:

$$\frac{\text{The number of new cases of the disease}}{\text{Total number of people at the beginning of the period who are free from the disease}}$$

Suppose in a community of 5,000 persons, 10 cases of hepatitis were found in a given year. Then the prevalence of the hepatitis is:

$$10/5{,}000 = 0.002$$

Thus, the probability that a randomly chosen person in the population has the case of hepatitis is 0.002.

Now, suppose that the number of new cases of hepatitis in the same community in a given time interval were 2. Then the incidence of hepatitis is:

$$2/5{,}000 = 0.0004$$

Thus the probability that a randomly chosen person develops a new case of hepatitis is 0.0004.

Screening Test: An Application of Bayes' Rule

In many epidemiologic studies, a common diagnostic procedure is to administer a screening test for presence or absence of a disease. Unfortunately, many screening tests are not definitive. A false negative is defined as a diseased person whose test is negative. A false positive is defined as a non-diseased person whose test is positive. The sensitivity of a test is the probability that the test is positive given that the person has the disease. The specificity of a test is the probability that the test is negative given that the person does not have the disease.

Predictive Accuracy of a Test

Epidemiologists are often interested in measuring the predictive accuracy of a test. The predictive accuracy positive (PA^+) of a screening test is the probability that a person has disease given that the test is positive. Similarly, the predictive accuracy negative (PA^-) is the probability that the person does not have disease given that the test is negative. The higher the predictive accuracy, the more valuable is the test. To accurately diagnose the disease for each patient, the ideal is to have tests

such that both PA^+ and PA^- are 1.

Application of Bayes' Rule

Unfortunately, the predictive accuracy of a screening test often cannot be determined directly; however, if the prevalence rate of a disease in a population is known, the sensitivity and specificity of the test (which are the quantities the physicians can estimate) can be used to compute the predictive accuracy by the use of a well-known result in probability--the Bayes' rule.

To illustrate the computation, the following events are defined:

T: The screening test is positive
D: The person has disease
$\bar{T}$: The screening test is negative
$\bar{D}$: The person does not have disease

Now, by the earlier discussion,

$$\text{Sensitivity} = P(T \mid D)$$
$$\text{Specificity} = P(\bar{T} \mid \bar{D})$$
$$PA^+ = P(D \mid T)$$
$$PA^- = P(\bar{D} \mid \bar{T})$$

Let P(D) denote the prevalence rate of the disease in the general population, i.e., the probability that a randomly chosen person in the population has the disease. Now, the predictive accuracy of the test can be determined by the following formulas using Bayes' rule:

$$P(D \mid T) = \frac{P(T \mid D) \cdot P(D)}{P(T \mid D) \cdot P(D) + P(T \mid \bar{D})\, P(\bar{D})}$$

$$PA^+ = \frac{(\text{Sensitivity})(\text{Prevalence})}{(\text{Sensitivity})(\text{Prevalence}) + (1-\text{Specificity})(1-\text{Prevalence})} \quad (5)$$

and

$$P(\bar{T} \mid \bar{D}) = \frac{P(\bar{T} \mid \bar{D})\, P(\bar{D})}{P(\bar{T} \mid \bar{D}) \cdot P(\bar{D}) + P(\bar{T} \mid D) \cdot P(D)}$$

$$PA^- = \frac{(\text{Specificity})(1-\text{Prevalence})}{(\text{Specificity})(1-\text{Prevalence}) + (1-\text{Sensitivity})(\text{Prevalence})} \quad (6)$$

As a screening test for the infection with human immunodeficiency virus (HIV), the AIDS virus, acquired immunodeficiency syndrome (AIDS) testing for the presence of antibody in blood is used. The most widely used test for detecting AIDS antibody is the enzyme-linked immunoabsorbent assay (ELISA). It has been found that ELISA has a false negative rate of 0.002% and a false positive rate of 2.5%. (Taylor & Przybyszewski, 1988). That is, out of 100,000 persons tested, only 2 results were falsely negative. ELISA is almost 100% sensitive for detecting persons with AIDS antibody. The test, however, is less efficient in its specificity, being 95.24% specific. Another way to express this is to say that a negative result virtually eliminates the possibility of having antibody; while, at least 5 out of every 100 results obtained from persons without the antibody may be falsely positive.

As seen before, another quantity which influences the predictive accuracy of a test is the probability which says that a positive result denotes the presence of the antibody. Prevalence says that the negative result denotes the absence of the antibody. To illustrate the computation of the predictive accuracy of the ELISA, consider a population with the prevalence, or antibody frequency, of 10 per 100,000 cases (0.01%):

Sensitivity=0.99998
Specifity = 0.9524
Prevalance = 0.0001

Then, by the use of the formulas (5) and (6),

$$PA^+ = \frac{(0.9998)(0.0001)}{(0.99998)(0.0001) + (1-0.9524)(1-0.0001)} = 2.0962 \times 10^{-3}$$

and

$$PA^- = \frac{(0.9524)(1-0.0001)}{(0.9524)(1-0.0001) + (1-0.99998)(0.0001)} = 1.0$$

Thus ,a negative result from ELISA is very predictive since it is almost 100% sure that such a person is free of the AIDS antibody; however, a positive result is not predictive at all.

Using the sensitivity and the specificity properties of the ELISA test given above, Table 4 illustrates the values of the

Table 4

Probability of True Positive and True Negative at Different Levels of Prevalence for Elisa (100% Sensitive; 95% Specific)

Population	Prevalence/100,000	%	False Positive	True Positive	PA^+	PA^-
A	10	0.01	5,000	10	3.9988×10^{-4}	1.0
B	50	0.05	4,998	50	1.9970×10^{-3}	1.0
C	50,000	50	2,500	50,000	0.80	1.0
D	75,000	75	1,250	75,000	0.9230	1.0

predictive accuracy positive and predictive accuracy negative at different levels of disease prevalence in the population.

Population A might represent the prevalence in a premarital test population; B, a hospital clinic population; and C and D, intravenous drug addicts at different cities in the United States. The obvious inference is that a positive result from A or B needs to be confirmed by using a more specific test but a positive result from C or D provides a sufficiently reliable index of the disease without the same imperative for second testing.

This example, using a test which is 100% sensitive and 95% specific, illustrates the usefulness of carrying out probability calculations in planning screening procedures for monitoring the health status of large populations. Although difficult to believe, it is important to understand that the extent to which certainty exists as a probability of being true or false depends on the prevalence of a disease in the tested population. The same test applied to two populations with different prevalence rates will yield different predictive values for positive and negative results.

Screening Results: A Tabular Representation

The concepts related to the computation of various probabilities in screening tests can be summarized more succinctly in Table 5. Suppose there are N individuals in the population classified according to the presence or absence of the disease, and the results of screening tests produced the results in Table 5.

Table 5

Distribution of Screening Results in Diseased and Nondiseased Population

	With Disease	Without Disease	Total
Screened positive	a	b	a+b
Screened negative	c	d	c+d
Total	a+c	b+d	a+b+c+d

The various probabilities considered earlier are given as follows:

Prevalence = $P(D) = (a + c)/N$
Sensitivity = $P(\bar{T} \mid \bar{D}) = a/(a + c)$
Specificity = $P(\bar{T} \mid \bar{D}) = d/(b + d)$
False positive rate = $P(\bar{T}$ and $\bar{D}) = b/N$
False negative rate = $P(\bar{T}$ and $D) = c/N$
Predictive value = $P(D \mid T) = a/(a + b)$
positive (PV^+)
Predictive value = $P(\bar{D} \mid \bar{T}) = d/(c + d)$
negative (PV^-)

The formulation of Bayes's rule given in equations (5) and (6) can be readily verified in terms of the aforementioned probabilities.

Suppose the screening tests by ELISA for HIV infection on 100,000 individuals yield the results shown in Table 6.

Table 6

Results of Screening Tests by ELISA for HIV Infection on 100 Individuals

	With AIDS	Without AIDS	Total
Screened positive	47,500	2,500	50,000
Screened negative	2	49,998	50,000
Total	47,502	52,498	100,000

Using these data one can determine prevalence, sensitivity, specificity, false positive (rate), false negative (rate), predictive value positive, and predictive value negative. The required probabilities are readily calculated as follows:

Prevalence = 47,502/100,000 = 47.0%
Sensitivity = 47,500/47,502 = 99.996%
Specificity = 49,998/52,498 = 95.238%
False positive (rate) = 2,500/100,000 = 2.5%
False negative (rate) = 2/100,000 = 0.002%
Predictive value positive = 47,500/50,000 = 95.0%
Predictive value negative = 49,998/50,000 = 99.996%

Conclusion

This article has illustrated some simple applications of elementary probability and statistics to a branch of medical science called epidemiology. Medical science is particularly concerned with statistical inference. Probability is widely used in clinical diagnosis where probabilistic statements are the terms in which a competent prognosis is given. There are many interesting and useful applications of probability and statistics to medical science, and teachers are encouraged to look for such applications from other sources. Some excellent references for applications of probability and statistics to medicine are Colton (1974), Ingelfinger, Mosteller, Thilodeau, and Ware (1987), and Murphy (1979).

References

Colton, T. (1974). *Statistics in medicine*. Boston, MA: Little Brown and Co.

Gastwirth, (1987). Statistical precision of medical screening procedures. *Statistical Science, 2*, 213-238.

Ingelfinger, J. A., Mosteller, F., Thilodeau, L. A., & Ware, J. H. (1987). *Biostatistics in clinical medicine* (2nd ed.). New York: MacMillan.

Murphy, E. A. (1979). *Probability in medicine*. Baltimore, MD: John Hopkins University Press.

Taylor, R. N., & Prgybyszewski, V. A. (1988). Summary of the Center for Disease Control human immunodeficiency virus (HIV) performance evaluation on surveys for 1985 and 1986. *American Journal of Clinical Pathology, 89*, 1-13.