

Random Error and the Role of Statistics

Statistics plays two main roles in the analysis of epidemiologic data: first, to measure variability in the data in an effort to assess the role of chance, and second, to estimate effects after correcting for biases such as confounding. This chapter concentrates on the assessment of variability. The use of statistical approaches to control confounding is discussed in Chapters 10 and 12.

An epidemiologic study can be viewed as an exercise in measurement. As in any measurement, the goal is to obtain an accurate result, with as little error as possible. Systematic error and random error can distort the measurement process. Chapter 7 describes the primary categories of systematic error. The error that remains after systematic error is eliminated is *random error*. Random error is nothing more than variability in the data that cannot be readily explained. Sometimes, random error stems from a random process, but it may not. In randomized trials, some of the variability in the data reflects the random assignment of subjects to the study groups. In most epidemiologic studies, however, there is no random assignment to study groups. For example, in a cohort study that compares the outcome of pregnancy among women who drink heavily chlorinated water with the outcome among women who drink bottled water, it is not chance but the decision making or circumstances of the women themselves that determines the cohort in which the women are grouped. The individual assignments to categories of water chlorination are not random; nevertheless, some of the variability in the outcome is considered to be random error. Much of this variation may reflect hidden biases and presumably can be accounted for by factors other than drinking water that affect the outcome of pregnancy. These factors may not have been measured among these women or perhaps not even discovered.

ESTIMATION

If an epidemiologic study is thought of as an exercise in measurement, the result of the study should be an estimate of an epidemiologic quantity. Ideally,

the analysis of data and the reporting of results should report the magnitude of that epidemiologic quantity and portray the degree of precision with which it is measured. For example, a case-control study may be undertaken to estimate the incidence rate ratio (RR) between use of cellular telephones and the occurrence of brain cancer. The report on the results of the study should present a clear estimate of the RR, such as $RR = 2.5$. When an estimate is presented as a single value, we refer to it as a *point estimate*. In this example, the point estimate of 2.5 quantifies the estimated strength of the relation between the use of cellular telephones and the occurrence of brain cancer. To indicate the precision of the point estimate, we use a *confidence interval*, which is a range of values around the point estimate. A wide confidence interval indicates low precision, and a narrow interval indicates high precision.

CHANCE

In ordinary language, the word *chance* has a dual meaning. One meaning refers to the outcome of a random process, implying an outcome that could not be predicted under any circumstances; the other refers to outcomes that cannot be predicted easily but are not necessarily random phenomena. For example, if you unexpectedly encounter your cousin on the beach at Cape Cod, you may describe it as a chance encounter. Nevertheless, there were presumably causal mechanisms that can explain why you and your cousin were on the beach at Cape Cod at that time. It may be a coincidence that the two causal mechanisms led to both of you being there together, but randomness does not necessarily play a role in explaining the encounter.

Flipping a coin is usually considered to be a randomizing event, one that is completely unpredictable. Nevertheless, the flip of a coin can be predicted with sufficient information about the initial conditions and the forces applied to the coin. The reason we consider it a randomizing event is that most of us do not have the necessary information nor the means to figure out from it what the outcome of the flip would be. Some individuals, however, have practiced flipping coins enough to predict the outcome of a given toss almost perfectly. For the rest of us, the flip of a coin appears random, despite the fact that the underlying process is not actually random. As we practice flipping or learn more about the sources of error in a body of data, we can reduce errors that may appear random at first. Physicists tell us that we will never be able to explain all components of error, but for the problems that epidemiologists address, it is reasonable to assume that much of the random error that we observe in data could be explained with better information.

POINT ESTIMATES, CONFIDENCE INTERVALS, AND P VALUES

We use confidence intervals because a point estimate, being a single value, cannot express the statistical variation, or random error, that underlies the estimate.

Random Error and the Role of Statistics

Statistics plays two main roles in the analysis of epidemiologic data: first, to measure variability in the data in an effort to assess the role of chance, and second, to estimate effects after correcting for biases such as confounding. This chapter concentrates on the assessment of variability. The use of statistical approaches to control confounding is discussed in Chapters 10 and 12.

An epidemiologic study can be viewed as an exercise in measurement. As in any measurement, the goal is to obtain an accurate result, with as little error as possible. Systematic error and random error can distort the measurement process. Chapter 7 describes the primary categories of systematic error. The error that remains after systematic error is eliminated is *random error*. Random error is nothing more than variability in the data that cannot be readily explained. Sometimes, random error stems from a random process, but it may not. In randomized trials, some of the variability in the data reflects the random assignment of subjects to the study groups. In most epidemiologic studies, however, there is no random assignment to study groups. For example, in a cohort study that compares the outcome of pregnancy among women who drink heavily chlorinated water with the outcome among women who drink bottled water, it is not chance but the decision making or circumstances of the women themselves that determines the cohort in which the women are grouped. The individual assignments to categories of water chlorination are not random; nevertheless, some of the variability in the outcome is considered to be random error. Much of this variation may reflect hidden biases and presumably can be accounted for by factors other than drinking water that affect the outcome of pregnancy. These factors may not have been measured among these women or perhaps not even discovered.

ESTIMATION

If an epidemiologic study is thought of as an exercise in measurement, the result of the study should be an estimate of an epidemiologic quantity. Ideally,

the analysis of data and the reporting of results should report the magnitude of that epidemiologic quantity and portray the degree of precision with which it is measured. For example, a case-control study may be undertaken to estimate the incidence rate ratio (RR) between use of cellular telephones and the occurrence of brain cancer. The report on the results of the study should present a clear estimate of the RR, such as $RR = 2.5$. When an estimate is presented as a single value, we refer to it as a *point estimate*. In this example, the point estimate of 2.5 quantifies the estimated strength of the relation between the use of cellular telephones and the occurrence of brain cancer. To indicate the precision of the point estimate, we use a *confidence interval*, which is a range of values around the point estimate. A wide confidence interval indicates low precision, and a narrow interval indicates high precision.

CHANCE

In ordinary language, the word *chance* has a dual meaning. One meaning refers to the outcome of a random process, implying an outcome that could not be predicted under any circumstances; the other refers to outcomes that cannot be predicted easily but are not necessarily random phenomena. For example, if you unexpectedly encounter your cousin on the beach at Cape Cod, you may describe it as a chance encounter. Nevertheless, there were presumably causal mechanisms that can explain why you and your cousin were on the beach at Cape Cod at that time. It may be a coincidence that the two causal mechanisms led to both of you being there together, but randomness does not necessarily play a role in explaining the encounter.

Flipping a coin is usually considered to be a randomizing event, one that is completely unpredictable. Nevertheless, the flip of a coin can be predicted with sufficient information about the initial conditions and the forces applied to the coin. The reason we consider it a randomizing event is that most of us do not have the necessary information nor the means to figure out from it what the outcome of the flip would be. Some individuals, however, have practiced flipping coins enough to predict the outcome of a given toss almost perfectly. For the rest of us, the flip of a coin appears random, despite the fact that the underlying process is not actually random. As we practice flipping or learn more about the sources of error in a body of data, we can reduce errors that may appear random at first. Physicists tell us that we will never be able to explain all components of error, but for the problems that epidemiologists address, it is reasonable to assume that much of the random error that we observe in data could be explained with better information.

POINT ESTIMATES, CONFIDENCE INTERVALS, AND P VALUES

We use confidence intervals because a point estimate, being a single value, cannot express the statistical variation, or random error, that underlies the estimate.

If a study is large, the estimation process can be comparatively precise, and there may be little random error in the estimation. A small study, however, has less precision, which means that the estimate is subject to more random error. A confidence interval indicates the amount of random error in the estimate. A given confidence interval is tied to an arbitrarily set level of confidence. Commonly, the level of confidence is set at 95% or 90%, although any level in the interval of 0% to 100% is possible. The confidence interval is defined statistically as follows: If the level of confidence is set to 95%, it means that if the data collection and analysis could be replicated many times and the study were free of bias, the confidence interval would include within it the correct value of the measure 95% of the time. This definition presumes that the only thing that would differ in these hypothetical replications of the study would be the statistical, or chance, element in the data. It also presumes that the variability in the data can be described adequately by a statistical model and that biases such as confounding are nonexistent or completely controlled. These unrealistic conditions are typically not met even in carefully designed and conducted randomized trials. In nonexperimental epidemiologic studies, the formal definition of a confidence interval is a fiction that at best provides a rough estimate of the statistical variability in a set of data. It is better not to consider a confidence interval to be a literal measure of statistical variability but rather a general guide to the amount of random error in the data.

The confidence interval is calculated from the same equations that are used to generate another commonly reported statistical measure, the *P value*, which is the statistic used for statistical hypothesis testing. The *P value* is calculated in relation to a specific hypothesis, usually the *null hypothesis*, which states that there is no relation between exposure and disease. For the *RR* measure, the null hypothesis is $RR = 1.0$. The *P value* represents the probability, assuming that the null hypothesis is true and the study is free of bias, that the data obtained in the study would demonstrate an association as far from the null hypothesis or farther than what was actually obtained. For example, suppose that a case-control study gives, as an estimate of the relative risk, $RR = 2.5$. The *P value* answers this question: What is the probability, if the true $RR = 1.0$, that a given study may give a result as far as this or farther from 1.0? The *P value* is the probability, conditional on the null hypothesis, of observing as strong an association as was observed or a stronger one.

P values can be calculated using statistical models that correspond to the type of data that have been collected (see Chapter 9). In practice, the variability of collected data is unlikely to conform precisely to any given statistical model. For example, most statistical models assume that the observations are independent of one another. Many epidemiologic studies, however, are based on observations that are not independent. Data also may be influenced by systematic errors that increase variation beyond that expected from a simple statistical model. Because the theoretical requirements are seldom met, a *P value* usually cannot be taken as a meaningful probability value. Instead, it can be viewed as something less technical: a measure of relative consistency between the null hypothesis and the data in hand. A large *P value* indicates that the data are highly consistent with the null hypothesis, and a low *P value* indicates that the data are not very consistent with the null hypothesis. More specifically, if a *P value* were as small as .01, it would mean that the data were not very consistent with the null hypothesis, but a *P value* as large as .5 would indicate that the data were reasonably consistent

with the null hypothesis. Neither of these *P values* should be interpreted as a strict probability. Neither tells us whether the null hypothesis is correct or not. The ultimate judgment about the correctness of the null hypothesis will depend on the existence of other data and the relative plausibility of the null hypothesis and its alternatives.

WHAT IS THE PROBABILITY THAT THE NULL HYPOTHESIS IS CORRECT?

Some people interpret a *P value* as a probability statement about the correctness of the null hypothesis, but that interpretation cannot be defended. First, the null hypothesis, like any hypothesis, should be regarded as true or false but not as having a probability of being true. A probability would not be assigned to the truth of any hypothesis except in a subjective sense, as in describing betting odds. Even in framing a subjective interpretation or in assigning betting odds, the *P value* should not be considered to be equivalent to the probability that the null hypothesis is correct.

It is true that the *P value* is a probability measure. When the data are very discrepant with the null hypothesis, the *P value* is small, and when the data are concordant with the null hypothesis, the *P value* is large. Nonetheless, the *P value* is not the probability that the null hypothesis is correct. It is calculated only after assuming that the null hypothesis is correct, and it refers to the probability that the association observed in the data, divided by its standard error, would deviate from the null value as much as it did or more. It can thus be viewed as a measure of consistency between the data and the null hypothesis, but it does not address whether the null hypothesis is correct. Suppose you buy a ticket for a lottery. Under the null hypothesis that the drawing is random, your chance of winning is slim. If you win, the *P value* evaluating the null hypothesis (that you won by chance) is tiny, because your winning is not a likely outcome in a random lottery with many tickets sold. Nevertheless, someone must win. If you did win, does that constitute evidence that the lottery was not random? Should you reject the null hypothesis because you calculated a very low *P value*? The answer is that even with a very low *P value*, the tenability of the null hypothesis depends on what alternative theories you have. One woman who twice won the New Jersey state lottery said she would stop buying lottery tickets to be fair to others. The more reasonable interpretation is that her two wins were chance events. The point is that the null hypothesis may be the most reasonable hypothesis for the data even if the *P value* is low. Similarly, the null hypothesis may be implausible or just incorrect even if the *P value* is high.

STATISTICAL HYPOTHESIS TESTING VERSUS ESTIMATION

Often, a *P value* is used to determine the presence or absence of statistical significance. Statistical significance is a term that appears laden with meaning,

although it tells nothing more than whether the P value is less than some arbitrary value, almost always .05. The term *statistically significant* and the statement " $P < .05$ " (or whatever level is taken as the threshold for statistical significance) are equivalent. Neither is a good description of the information in the data.

Statistical hypothesis testing is a term used to describe the process of deciding whether to reject or not to reject a specific hypothesis, usually the null hypothesis. Statistical hypothesis testing is predicated on statistical significance as determined from the P value. Typically, if an analysis gives a result that is statistically significant, the null hypothesis is rejected as false. If a result is not statistically significant, it means that the null hypothesis cannot be rejected. It does not mean that the null hypothesis is correct. No data analysis can determine definitively whether the null hypothesis or any hypothesis is true or false. Nevertheless, it is unfortunately often the case that a statistical significance test is interpreted to mean that the null hypothesis is false or true according to whether the statistical test of the relation between exposure and disease is or is not statistically significant. In practice, a statistical test, accompanied by its declaration of "significant" or "not significant," is often mistakenly used as a forced decision on the truth of the null hypothesis.

A declaration of statistical significance offers less information than the P value, because the P value is a number, whereas statistical significance is just a dichotomous description. There is no reason that the numeric P value must be degraded into this less-informative dichotomy. Even the more quantitative P value has a problem, however, because it confounds two important aspects of the data, the strength of the relation between exposure and disease and the precision with which that relation is measured. To have a clear interpretation of data, it is important to be able to separate the information on strength of relation and precision, which is the job that estimation does for us.

P-VALUE (CONFIDENCE INTERVAL) FUNCTIONS

To illustrate how estimation does a better job of expressing strength of relation and precision, we describe a curve that is often called a *P-value function* but is also referred to as a *confidence interval function*. The P -value function enlarges on the concept of the P value. The P value is a statistic that can be viewed as a measure of the compatibility between the data in hand and the null hypothesis. We can enlarge on this concept by imagining that instead of testing just the null hypothesis, we also calculate a P value for a range of other hypotheses. Consider the rate ratio measure, which can range from 0 to infinity and equals 1.0 if the null hypothesis is correct. The ordinary P value is a measure of the consistency between the data and the hypothesis that $RR = 1.0$. Mathematically, however, we are not constrained to test only the hypothesis that $RR = 1.0$. For any set of data, we can in principle calculate a P value that measures the compatibility between those data and any value of RR . We can even calculate an infinite number of P values that test every possible value of RR . If we did so and plotted the results, we end up with the P -value function. An example of a P -value function is given in Figure 8-1, which is based on the data in Table 8-1 describing a case-control study of drug exposure during pregnancy and congenital heart disease.¹

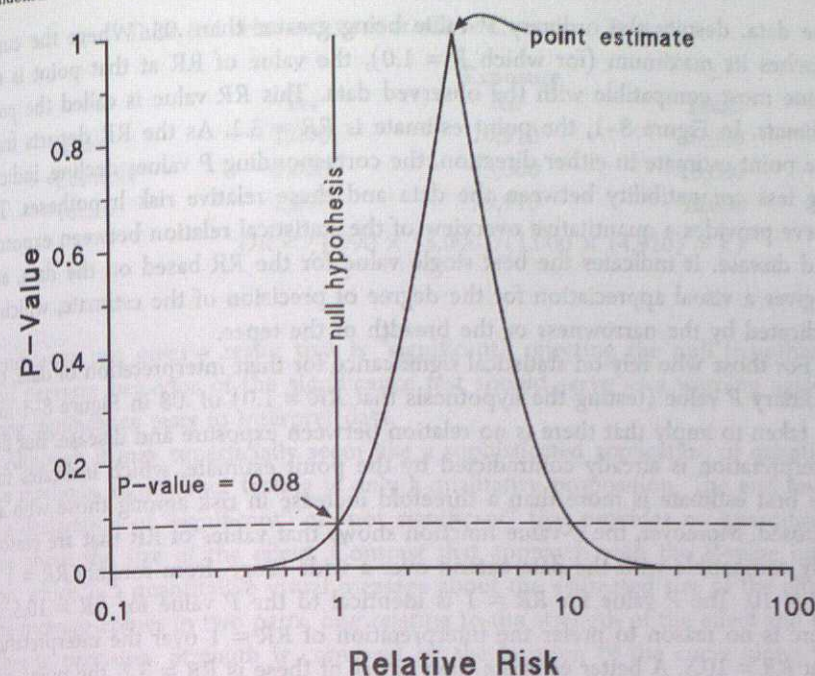


Figure 8-1 P -value function for the case-control data in Table 8-1.

The curve in Figure 8-1, which resembles a tepee, plots the P value that tests the compatibility of the data in Table 8-1 with every possible value of RR . When $RR = 1.0$, the curve gives the P value testing the hypothesis that $RR = 1.0$; this is the usual P value testing the null hypothesis. For the data depicted in Figure 8-1, the ordinary P value is .08. This value would be described by many observers as not significant, because the P value is greater than .05. To many people, *not significant* implies that there is no relation between exposure and disease in the data. It is a fallacy, however, to infer a lack of association from a P value. The curve also gives the P values testing every other possible value of the RR , thus indicating the degree of compatibility between the data and every possible value of RR . The full P -value function in Figure 8-1 makes it clear that there is a strong association in

Table 8-1 CASE-CONTROL DATA FOR CONGENITAL HEART DISEASE AND CHLORDIAZEPOXIDE USE IN EARLY PREGNANCY

	Chlordiazepoxide Use		
	Yes	No	Total
Cases	4	386	390
Controls	4	1250	1254
Total	8	1636	1644

$$OR = (4 \times 1250) / (4 \times 386) = 3.2$$

Data from Rothman et al.¹

the data, despite the ordinary P value being greater than .05. Where the curve reaches its maximum (for which $P = 1.0$), the value of RR at that point is the value most compatible with the observed data. This RR value is called the *point estimate*. In Figure 8-1, the point estimate is $RR = 3.2$. As the RR departs from the point estimate in either direction, the corresponding P values decline, indicating less compatibility between the data and these relative risk hypotheses. The curve provides a quantitative overview of the statistical relation between exposure and disease. It indicates the best single value for the RR based on the data, and it gives a visual appreciation for the degree of precision of the estimate, which is indicated by the narrowness or the breadth of the tepee.

For those who rely on statistical significance for their interpretation of data, the ordinary P value (testing the hypothesis that $RR = 1.0$) of .08 in Figure 8-1 may be taken to imply that there is no relation between exposure and disease. But that interpretation is already contradicted by the point estimate, which indicates that the best estimate is more than a threefold increase in risk among those who are exposed. Moreover, the P -value function shows that values of RR that are reasonably compatible with the data extend over a wide range, from roughly $RR = 1$ to $RR = 10$. The P value for $RR = 1$ is identical to the P value for $RR = 10.5$, so there is no reason to prefer the interpretation of $RR = 1$ over the interpretation that $RR = 10.5$. A better estimate than either of these is $RR = 3.2$, the point estimate. The main lesson here is how misleading it can be to try to base an inference on a test of statistical significance, or, for that matter, on a P value.

The lesson is reinforced when we consider another P -value function that describes a set of hypothetical data given in Table 8-2. These hypothetical data lead to a narrow P -value function that reaches a peak slightly above the null value, $RR = 1$. Figure 8-2 contrasts the P -value function for the data in Table 8-2 with the P -value function given earlier for the data in Table 8-1. The narrowness of the second P -value function reflects the larger size of the second set of data. Large size translates to better precision, for which the visual counterpart is the narrow P -value function.

There is a striking contrast in messages from these two P -value functions. The first function suggests that the data are imprecise but reflect an association that is strong; the data are readily compatible with a wide range of effects, from very little or nothing to more than a 10-fold increase in risk. The first set of data thus raises the possibility that the exposure is a strong risk factor. Although the data do not permit a precise estimate of effect, the range of effect values consistent with the data includes mostly strong effects that would warrant concern about the exposure. This concern comes from data that give a "nonsignificant" result for a test of the null hypothesis. In contrast, the other set of data, from Table 8-2, gives a precise estimate of an effect that is close to the null. The data are not very compatible with a strong effect and, indeed, may be interpreted as reassuring about the absence of a strong effect. Despite this reassurance, the P value testing the null hypothesis is .04; a test of the null hypothesis would give a "statistically significant" result, rejecting the null hypothesis. In both cases, reliance on the significance test would be misleading and conducive to an incorrect interpretation. In the first case, the association is "not significant," but the study is properly interpreted as raising concern about the effect of the exposure. In the second case, the study provides reassurance about the absence of a strong effect, but the

Table 8-2 HYPOTHETICAL CASE-CONTROL DATA

	Exposure		Total
	Yes	No	
Cases	1,090	14,910	16,000
Controls	1,000	15,000	16,000
Total	2,090	29,910	32,000

$$OR = (1,090 \times 15,000) / (1,000 \times 14,910) = 1.1$$

significance test gives a result that is "significant," rejecting the null hypothesis. This perverse behavior of the significance test should serve as a warning against using significance tests to interpret data.

Although it may superficially seem like a sophisticated application of quantitative methods, significance testing is only a qualitative proposition. The end result is a declaration of "significant" or "not significant" that provides no quantitative clue about the size of the effect. Contrast that approach with the P -value function, which is a quantitative visual message about the estimated size of the effect. The message comes in two parts, one relating to the strength of the effect and the other to precision. Strength is conveyed by the location of the curve along the horizontal axis and precision by the amount of spread of the function around the point estimate.

Because the P value is only one number, it cannot convey two separate quantitative messages. To get the message about both strength of effect and precision,

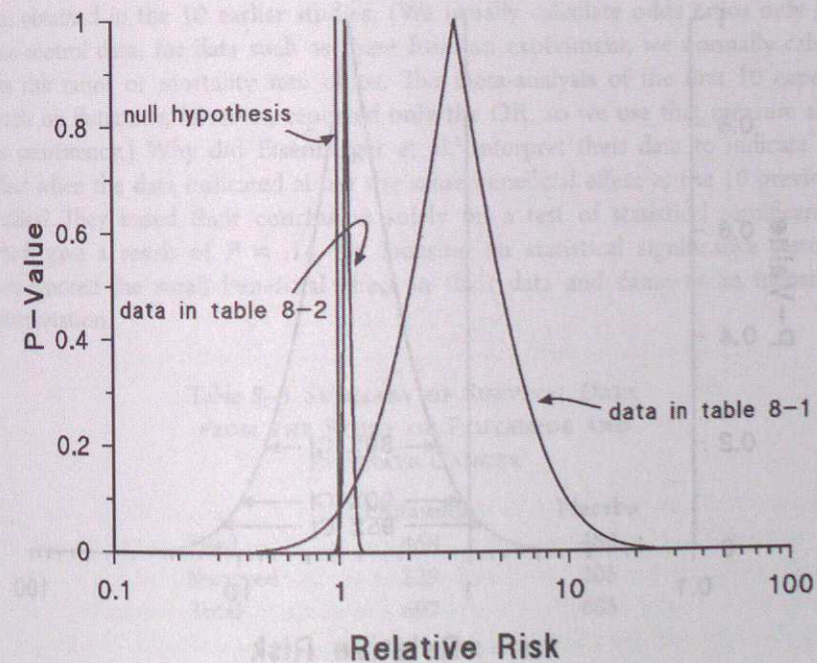


Figure 8-2 P -value function for the data in Table 8-1 and the hypothetical case-control data in Table 8-2.

at least two numbers are required. Perhaps the most straightforward way to get both messages is from the upper and lower confidence limits, the two numbers that form the boundaries to a confidence interval. The P -value function is closely related to the set of all confidence intervals for a given estimate. This relation is depicted in Figure 8-3, which shows three different confidence intervals for the data in Figure 8-1. These three confidence intervals differ only in the arbitrary level of confidence that determines the width of the interval. In Figure 8-3, the 95% confidence interval can be read from the curve along the horizontal line where $P = .05$ and the 90% and 80% intervals along the lines where $P = .1$ and $.2$, respectively. The different confidence intervals in Figure 8-3 reflect the same degree of precision but differ in their width only because the level of confidence for each is arbitrarily different. The three confidence intervals depicted in Figure 8-3 are described as *nested* confidence intervals. The P -value function is a graph of all possible nested confidence intervals for a given estimate, reflecting all possible levels of confidence between 0% and 100%. It is this ability to find all possible confidence intervals from a P -value function that leads to its description as either a P -value function or a confidence interval function.

It is common to see confidence intervals reported for an epidemiologic measure, but it is uncommon to see a full P -value function or confidence interval function. Fortunately, it is not necessary to calculate and display a full P -value function to infer the two quantitative messages, strength of relation and precision, for an estimate. A single confidence interval is sufficient, because the upper and lower confidence bounds from a single interval are sufficient to determine the entire P -value function. If we know the lower and upper limit to the confidence interval, we know the location of the P -value function along the horizontal axis

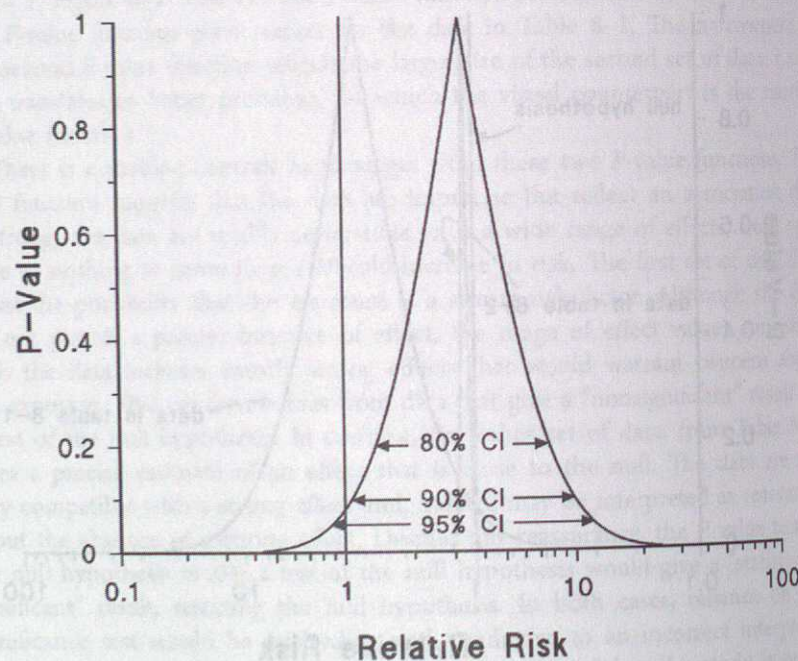


Figure 8-3 P -value function for the data from Table 8-1, showing how nested confidence intervals can be read from the curve.

and the spread of the function. Thus, from a single confidence interval, we can construct an entire P -value function. We do not need to go through the labor of calculating this function if we can visualize the two messages that it can convey directly from the confidence interval.

Regrettably, confidence intervals are too often not interpreted with the image of the corresponding P -value function in mind. A confidence interval can unfortunately be used as a surrogate test of statistical significance: a confidence interval that contains the null value within it corresponds to a significance test that is "not significant," and a confidence interval that excludes the null value corresponds to a significance test that is "significant." The allure of significance testing is so strong that many people use a confidence interval merely to determine "significance" and thereby ignore the potentially useful quantitative information that the confidence interval provides.

Example: Is Flutamide Effective in Treating Prostate Cancer?

In a randomized trial of flutamide, which is used to treat prostate cancer, Eisenberger et al.² reported that patients who received flutamide fared no better than those who received placebo. Their interpretation that flutamide was ineffective contradicted the results of 10 previous studies, which collectively had pointed to a modest benefit. The 10 previous studies, on aggregate, indicated about an 11% survival advantage for patients receiving flutamide [odds ratio (OR) = 0.89]. The actual data reported by Eisenberger et al. are given in Table 8-3. From these data, we can calculate an OR of 0.87, almost the same result (slightly better) as was obtained in the 10 earlier studies. (We usually calculate odds ratios only for case-control data; for data such as these from an experiment, we normally calculate risk ratios or mortality rate ratios. The meta-analysis of the first 10 experiments on flutamide, however, reported only the OR, so we use that measure also for consistency.) Why did Eisenberger et al.² interpret their data to indicate no effect when the data indicated about the same beneficial effect as the 10 previous studies? They based their conclusion solely on a test of statistical significance, which gave a result of $P = .14$. By focusing on statistical significance testing, they ignored the small beneficial effect in their data and came to an incorrect interpretation.

Table 8-3 SUMMARY OF SURVIVAL DATA FROM THE STUDY OF FLUTAMIDE AND PROSTATE CANCER

	Flutamide	Placebo
Died	468	480
Survived	229	205
Total	697	685

OR = 0.87

95% CI: 0.70-1.10

Data from Eisenberger et al.²

The original 10 studies on flutamide were published in a review that summarized the results.³ It is helpful to examine the *P*-value function from these 10 studies and to compare it with the *P*-value function after adding the study of Eisenberger et al.² to the earlier studies (Fig. 8-4).⁴ The only change apparent from adding the data of Eisenberger et al.² is a slightly improved precision of the estimated benefit of flutamide in reducing the risk of dying from prostate cancer.

Example: Is St. John's Wort Effective in Relieving Major Depression?

Extracts of St. John's Wort (*Hypericum perforatum*), a small, flowering weed, have long been used as a folk remedy. It is a popular herbal treatment for depression. Shelton et al.⁵ reported the results of a randomized trial of 200 patients with major depression who were randomly assigned to receive either St. John's Wort or placebo. Of 98 who received St. John's Wort, 26 responded positively, whereas 19 of the 102 who received placebo responded positively. Among those whose depression was relatively less severe at entry into the study (a group that the investigators thought might be more likely to show an effect of St. John's Wort), the

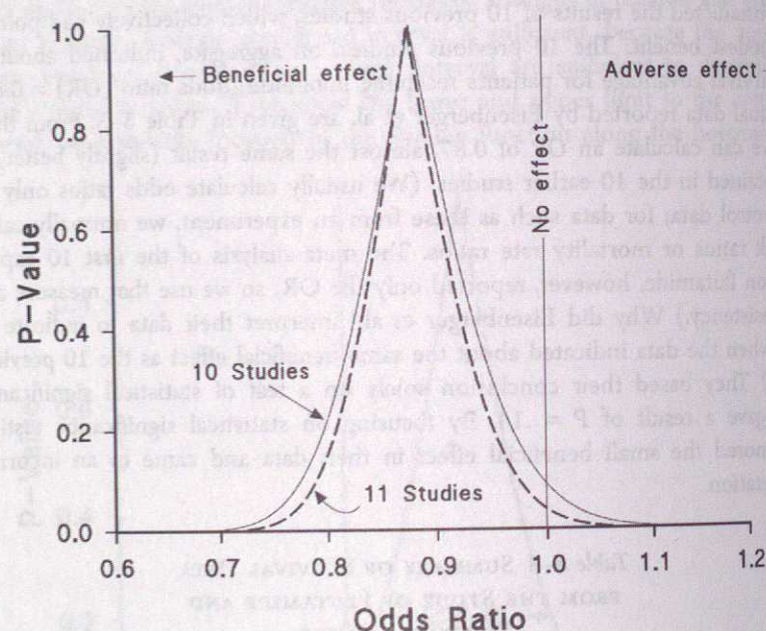


Figure 8-4 *P*-value functions for the first 10 studies of flutamide and prostate cancer survival (solid line)³ and for the first 11 studies (dashed line) after adding the study by Eisenberger et al.² The study by Eisenberger et al. did not shift the overall findings toward the null value but instead shifted the overall findings a minuscule step away from the null value. Nevertheless, because of an inappropriate reliance on statistical significance testing, the data were incorrectly interpreted as refuting earlier studies and indicating no effect of flutamide, despite the fact that the findings replicated previous results. (Reproduced with permission from Rothman et al.⁴)

Table 8-4 REMISSIONS AMONG PATIENTS WITH LESS SEVERE DEPRESSION

	St. John's Wort	Placebo
Remission	12	5
No remission	47	45
Total	59	50

RR = 2.0
90% CI: 0.90-4.6

Data from Shelton et al.⁵

proportion of patients who had remission of disease was twice as great among the 59 patients who received St. John's Wort as among the 50 who received a placebo (Table 8-4).

In Table 8-4, *risk ratio* refers to the "risk" of having a remission in symptoms, which is an improvement, so any increase above 1.0 indicates a beneficial effect of St. John's Wort; the RR of 2.0 indicates that the probability of a remission was twice as great for those receiving St. John's Wort. Despite these and other encouraging findings in the data, the investigators based their interpretation on a lack of statistical significance and concluded that St. John's Wort was not effective. A look at the *P*-value function that corresponds to the data in Table 8-4 is instructive (Fig. 8-5).

Figure 8-5 shows that the data regarding remissions among the less severely affected patients hardly support the theory that St. John's Wort is ineffective. The data for other outcomes were also generally favorable for St. John's Wort but, for

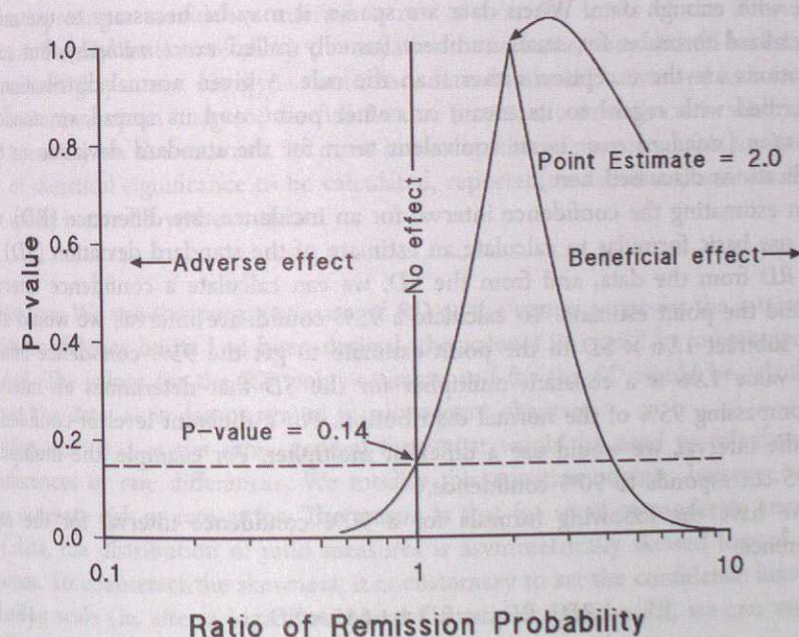


Figure 8-5 *P*-value function for the effect of St. John's Wort on remission from major depression among relatively less severely affected patients. (Data from Shelton et al.⁵)

almost all comparisons, not statistically significant. Instead of concluding, as they should have, that these data are readily compatible with moderate and even strong beneficial effects of St. John's Wort, the investigators drew the wrong conclusion, based on the lack of statistical significance in the data. Although the P value from this study is not statistically significant, the P value for the null hypothesis has the same magnitude as the P value testing the hypothesis that the $RR = 4.1$ (on the graph, the dashed line intersects the P -value function at $RR = 1.0$ and $RR = 4.1$). Although the investigators interpreted the data as supporting the hypothesis that $RR = 1.0$, the data are equally compatible with values of 1.0 or 4.1. Furthermore, it is not necessary to construct the P -value function in Figure 8-5 to reach this interpretation. An investigator need look no farther than the confidence interval given in Table 8-4 to appreciate the location and the spread of the underlying P -value function.

SIMPLE APPROACHES TO CALCULATING CONFIDENCE INTERVALS

The following chapters present basic methods for analyzing epidemiologic data. The focus is on estimating epidemiologic measures of effect, such as risk and rate ratios and, in cohort studies, risk and rate differences as well. The overall strategy in a data analysis is to obtain a good point estimate of the epidemiologic measure that we seek and an appropriate confidence interval.

Confidence intervals are usually calculated on the presumption that the estimate comes from the statistical distribution called a *normal distribution*, the usual bell-shaped curve. Estimates based on the normal distribution are always reasonable with enough data. When data are sparse, it may be necessary to use more specialized formulas for small numbers (usually called *exact methods*), but such situations are the exception rather than the rule. A given normal distribution is described with regard to its mean, or center point, and its spread, or *standard deviation* (*standard error* is an equivalent term for the standard deviation in the applications discussed here).

In estimating the confidence interval for an incidence rate difference (RD), we can use basic formulas to calculate an estimate of the standard deviation (SD) of the RD from the data, and from the SD , we can calculate a confidence interval around the point estimate. To calculate a 95% confidence interval, we would add and subtract $1.96 \times SD$ to the point estimate to get the 95% confidence limits. The value 1.96 is a constant multiplier for the SD that determines an interval encompassing 95% of the normal distribution. For a different level of confidence for the interval, we would use a different multiplier. For example, the multiplier 1.645 corresponds to 90% confidence.

We have the following formula for a 90% confidence interval for the rate difference:

$$RD_L, RD_U = RD \pm 1.645 \times SD \quad [8-1]$$

In Equation 8-1, RD_L refers to the lower confidence limit, obtained using the minus sign, and RD_U refers to the upper confidence limit, obtained by using the

STATISTICAL SIGNIFICANCE TESTING VERSUS ESTIMATION

Statistical significance testing is so ingrained that it is almost ubiquitous. Even those who acknowledge the impropriety of basing a conclusion on the results of a statistical significance test often fall into the bad habit of equating a lack of significance with a lack of effect and the presence of significance with "proof" of an effect. Significance testing evaluates only one theory that is an alternative to causation to explain the data, the theory that chance accounts for the findings. Nonchance alternative theories, such as confounding, selection bias, and bias from measurement error, are all more important to consider. For example, if an investigator finds a non-significant result and consequently does not explore it further, he or she may be ignoring an important and even strong association that has been underestimated because of confounding or nondifferential misclassification. To evaluate these issues, it is crucial to take a quantitative view of the data and their interpretation. That is, it is essential to think in terms of estimation rather than testing.

Significance testing is qualitative, not quantitative. When P values are calculated, they are often reported using inequalities, such as $P < .05$, rather than equalities, such as $P = .023$. Nothing is gained by converting the continuous P -value measure into a dichotomy, but even the numeric P value is far inferior to an estimate of effect, such as that obtained from a confidence interval. Estimation using confidence intervals allows the investigator to quantify separately the strength of a relation and the precision of an estimate and to reach a more reasonable interpretation. The key issue in interpreting a confidence interval is not to take the limits of the interval too literally. Instead of sharp demarcation boundaries, they should be considered gray zones. Ideally, a confidence interval should be viewed as a tool to conjure up an image of the full P -value function, a smooth curve with no boundary on the estimate. In most instances, there is no need for any test of statistical significance to be calculated, reported, or relied on, and we are much better off without them.

plus sign. We use the point estimate of RD as the center point for the interval. If 95% confidence limits had been desired, the value 1.96 could be substituted for 1.645. The values for the RD point estimate and for the SD would be calculated from the data, as is demonstrated in subsequent chapters.

Equation 8-1 is the same general form that would be used to calculate risk differences or rate differences. We modify this equation slightly, however, when we estimate risk or rate ratios. The reason is that for small or moderate amounts of data, the distribution of ratio measures is asymmetrically skewed toward large values. To counteract the skewness, it is customary to set the confidence limits on the log scale (ie, after a logarithmic transformation). For the RR , we can use the following equation to determine a 90% confidence interval.

$$\ln(RR_L), \ln(RR_U) = \ln(RR) \pm 1.645 \times SD(\ln(RR))$$

The term $\ln()$ refers to the natural logarithm transformation. A natural logarithm is a logarithm using the base $e \approx 2.7183$. Because this equation gives confidence limits on the log scale, the limits need to be converted back to the RR scale after they are calculated, by reversing the transformation, which involves taking antilogarithms. The whole process can be summarized by Equation 8-2 (for a 90% confidence interval):

$$RR_L, RR_U = e^{(\ln(RR) \pm 1.645 \times SD(\ln(RR)))} \quad [8-2]$$

In the next chapter, we apply these equations to the analysis of simple epidemiologic data.

QUESTIONS

1. Why should confidence intervals and P values have a different interpretation in a case-control study or a cohort study than a randomized experiment? What is the effect of the difference on the interpretation?
2. Which has more interpretive value, a confidence interval, a P value, or a statement about statistical significance? Explain.
3. In what way is a P value inherently confounded?
4. What are the two main messages that should come with a statistical estimate? How are these two messages conveyed by a P -value function?
5. Suppose that a study showed that former professional football players experienced a rate ratio for coronary heart disease of 3.0 compared with science teachers of the same age and sex, with a 90% confidence interval of 1.0 to 9.0. Sketch the P -value function. What is your interpretation of this finding, presuming that there is no confounding or other obvious bias that distorts the results?
6. One argument sometimes offered in favor of statistical significance testing is that it is often necessary to come to a yes-or-no decision about the effect of a given exposure or therapy. Significance testing has the apparent benefit of providing a dichotomous interpretation that could be used to make a yes-or-no decision. Comment on the validity of the argument that a decision is sometimes needed based on a research study. What would be the pros and cons of using statistical significance to judge whether an exposure or a therapy has an effect?
7. Are confidence intervals always symmetric around the point estimate? Why or why not?
8. What is the problem with using a confidence interval to determine whether or not the null value lies within the interval?

9. Consider two study designs, A and B, that are identical apart from the study size. Study A is planned to be much larger than study B. If both studies are conducted, which of the following statements is correct? (1) The 90% confidence interval for the rate ratio from study A has a greater probability of including the true rate ratio value than the 90% confidence interval from study B. (2) The 90% confidence interval for the rate ratio from study A has a smaller probability of including the true rate ratio value than the 90% confidence interval from study B. (3) The 90% confidence intervals for the rates ratio from study A and study B have equal probabilities of including the true rate ratio value. Before answering, be sure to take into account the fact that no study is without some bias.

REFERENCES

1. Rothman KJ, Fyler DC, Goldblatt A, et al. Exogenous hormones and other drug exposures of children with congenital heart disease. *Am J Epidemiol.* 1979;109:433-439.
2. Eisenberger MA, Blumenstein BA, Crawford ED, et al. Bilateral orchiectomy with or without flutamide for metastatic prostate cancer. *N Engl J Med.* 1998;339:1036-1042.
3. Prostate Cancer Trialists' Collaborative Group. Maximum androgen blockade in advanced prostate cancer: an overview of 22 randomised trials with 3283 deaths in 5710 patients. *Lancet.* 1995;346:265-269.
4. Rothman KJ, Johnson ES, Sugano DS. Is flutamide effective in patients with bilateral orchiectomy? *Lancet.* 1999;353:1184.
5. Shelton RC, Keller MB, Gelenberg A, et al. Effectiveness of St John's wort in major depression: a randomized controlled trial. *JAMA.* 2001;285:1978-1986.