

## 9 Analyzing Simple Epidemiologic Data

This chapter provides the statistical tools to analyze simple epidemiologic data, such as crude data from a study with no confounding. Because our emphasis is on estimation rather than statistical significance testing, we concentrate on formulas for obtaining confidence intervals for basic epidemiologic measures, although we also include formulas to derive *P* values.

The equations presented in this chapter give only approximate results and are valid only for data with sufficiently large numbers. More accurate estimates can be obtained by using what is called *exact* methods. It is difficult to determine a precise threshold of data above which we can say that the approximate results are good enough and below which we would say that exact calculations are needed. Fortunately, even for studies with modest numbers, the interpretation of results rarely changes when exact rather than approximate results are used to estimate confidence intervals. For those who inappropriately place emphasis on whether a confidence interval contains the null value (thereby converting the confidence interval into a statistical test), it may appear to matter if the limit changes its value slightly with a different equation and the limit is near the null value—a situation equivalent to being on the borderline of statistical significance. As explained in the previous chapter, however, placing emphasis on the exact location of a confidence interval, equivalent to placing emphasis on statistical significance, is an inappropriate and potentially misleading way to interpret data. With proper interpretation, which ignores the precise location of a confidence limit and instead considers the general width and location of an interval, the difference between results from approximate and exact formulas becomes much less important.

### CONFIDENCE INTERVALS FOR MEASURES OF DISEASE FREQUENCY

#### Risk Data and Prevalence Data

Suppose we observe that 20 people of 100 become ill with influenza during the winter season. We would estimate the risk, *R*, of influenza to be 20/100, or 0.2.

To obtain a confidence interval, we need to apply a statistical model. For risk data or prevalence data, the model usually applied is the binomial model. To use the model to obtain a confidence interval, it helps to have some simple notation. We can use *a* to represent cases and *N* to represent people at risk. Using this notation, our estimate of risk is the number of cases divided by the total number of people at risk:  $R = a/N$ . We can obtain a confidence interval from the following equation:

$$R_L, R_U = R \pm Z \cdot SE(R) \quad [9-1]$$

In Equation 9-1, the minus sign is used to obtain the lower confidence limit and the plus sign is used to obtain the upper confidence limit. *Z* is a fixed value, taken from the standard normal distribution, that determines the confidence level. If *Z* is set at 1.645, the result is a 90% confidence interval; if it is set at 1.96, the result is a 95% confidence interval. *SE*(*R*) is the *standard error* of *R*. The standard error is a measure of the statistical variability of the estimate. Under the binomial model, the standard error of *R* would be

$$SE(R) = \sqrt{\frac{a(N-a)}{N^3}}$$

#### Example: Confidence Limits for a Risk or Prevalence

Using the following equation with the example of 20 cases of influenza among 100 people, we can calculate a 90% confidence interval for the risk as follows. The lower bound would be

$$R_L = R - Z \cdot SE(R) = 0.20 - 1.645 \cdot \sqrt{\frac{20 \cdot 80}{100^3}} = 0.13$$

The upper bound could be obtained by substituting a plus sign for the minus sign in the calculation. Making this substitution gives a value 0.27 for the upper bound. With 20 influenza cases in a population of 100 at risk, the 90% confidence interval for the risk estimate of 0.2 is 0.13 to 0.27.

#### Incidence Rate Data

For incidence rate data, we use *a* to represent cases and *PT* to represent person-time. Although the notation is similar to that for risk data, these data differ conceptually and statistically from the binomial model used to describe risk data. For binomial data, the number of cases cannot exceed the total number of people at risk. In contrast, for rate data, the denominator does not relate to a specific number of people but rather to a time total. We do not know from the value of the person-time denominator, *PT*, how many people might have contributed time.

For statistical purposes, we invoke a model for incidence rate data that allows the number of cases to vary without any upper limit. It is the Poisson model.



## 9 Analyzing Simple Epidemiologic Data

This chapter provides the statistical tools to analyze simple epidemiologic data, such as crude data from a study with no confounding. Because our emphasis is on estimation rather than statistical significance testing, we concentrate on formulas for obtaining confidence intervals for basic epidemiologic measures, although we also include formulas to derive *P* values.

The equations presented in this chapter give only approximate results and are valid only for data with sufficiently large numbers. More accurate estimates can be obtained by using what is called *exact* methods. It is difficult to determine a precise threshold of data above which we can say that the approximate results are good enough and below which we would say that exact calculations are needed. Fortunately, even for studies with modest numbers, the interpretation of results rarely changes when exact rather than approximate results are used to estimate confidence intervals. For those who inappropriately place emphasis on whether a confidence interval contains the null value (thereby converting the confidence interval into a statistical test), it may appear to matter if the limit changes its value slightly with a different equation and the limit is near the null value—a situation equivalent to being on the borderline of statistical significance. As explained in the previous chapter, however, placing emphasis on the exact location of a confidence interval, equivalent to placing emphasis on statistical significance, is an inappropriate and potentially misleading way to interpret data. With proper interpretation, which ignores the precise location of a confidence limit and instead considers the general width and location of an interval, the difference between results from approximate and exact formulas becomes much less important.

### CONFIDENCE INTERVALS FOR MEASURES OF DISEASE FREQUENCY

#### Risk Data and Prevalence Data

Suppose we observe that 20 people of 100 become ill with influenza during the winter season. We would estimate the risk, *R*, of influenza to be 20/100, or 0.2.

To obtain a confidence interval, we need to apply a statistical model. For risk data or prevalence data, the model usually applied is the binomial model. To use the model to obtain a confidence interval, it helps to have some simple notation. We can use *a* to represent cases and *N* to represent people at risk. Using this notation, our estimate of risk is the number of cases divided by the total number of people at risk:  $R = a/N$ . We can obtain a confidence interval from the following equation:

$$R_L, R_U = R \pm Z \cdot SE(R) \quad [9-1]$$

In Equation 9-1, the minus sign is used to obtain the lower confidence limit and the plus sign is used to obtain the upper confidence limit. *Z* is a fixed value, taken from the standard normal distribution, that determines the confidence level. If *Z* is set at 1.645, the result is a 90% confidence interval; if it is set at 1.96, the result is a 95% confidence interval. *SE*(*R*) is the *standard error* of *R*. The standard error is a measure of the statistical variability of the estimate. Under the binomial model, the standard error of *R* would be

$$SE(R) = \sqrt{\frac{a(N-a)}{N^3}}$$

#### Example: Confidence Limits for a Risk or Prevalence

Using the following equation with the example of 20 cases of influenza among 100 people, we can calculate a 90% confidence interval for the risk as follows. The lower bound would be

$$R_L = R - Z \cdot SE(R) = 0.20 - 1.645 \cdot \sqrt{\frac{20 \cdot 80}{100^3}} = 0.13$$

The upper bound could be obtained by substituting a plus sign for the minus sign in the calculation. Making this substitution gives a value 0.27 for the upper bound. With 20 influenza cases in a population of 100 at risk, the 90% confidence interval for the risk estimate of 0.2 is 0.13 to 0.27.

#### Incidence Rate Data

For incidence rate data, we use *a* to represent cases and *PT* to represent person-time. Although the notation is similar to that for risk data, these data differ conceptually and statistically from the binomial model used to describe risk data. For binomial data, the number of cases cannot exceed the total number of people at risk. In contrast, for rate data, the denominator does not relate to a specific number of people but rather to a time total. We do not know from the value of the person-time denominator, *PT*, how many people might have contributed time.

For statistical purposes, we invoke a model for incidence rate data that allows the number of cases to vary without any upper limit. It is the Poisson model.



We take  $a/PT$  as the estimate of the disease rate, and we calculate a confidence interval for the rate using Equation 9-1 with the following standard error:

$$SE(R) = \sqrt{\frac{a}{PT^2}}$$

### DO RATES ALWAYS DESCRIBE POPULATION SAMPLES?

Some theoreticians propose that if a rate or risk is measured in an entire population, there is no point to calculating a confidence interval, because a confidence interval is intended to convey only the imprecision that comes from taking a sample from a population. According to this reasoning, if the entire population is measured instead of a sample, there is no sampling error to worry about and therefore no confidence interval to compute. There is another side to this argument, however. Others hold that even if the rate or risk is measured in an entire population, that population represents only a sample of people from a hypothetical superpopulation. In other words, the study population, even if enumerated completely without any sampling, represents merely a biologic sample of a larger set of people; therefore, a confidence interval is justified.

The validity of each argument may depend on the context. If one is measuring voter preference, it is the actual population in which one is interested, and the first argument is reasonable. For biologic phenomena, however, what happens in an actual population may be of less interest than the biologic norm that describes the superpopulation. Therefore, for biologic phenomena the second argument is more compelling.

### Example: Confidence Limits for an Incidence Rate

Consider as an example a cancer incidence rate estimated from a registry that reports 8 cases of astrocytoma among 85,000 person-years at risk. The rate is 8/85,000 person-years, or 9.4 cases/100,000 person-years. A lower 90% confidence limit for the rate would be estimated as

$$R_L = R - Z \cdot SE(R) = \frac{8}{85,000 \text{ person-years}} - 1.645 \cdot \sqrt{\frac{8}{(85,000 \text{ person-years})^2}} \\ = 3.9/100,000 \text{ person-years}$$

Using the plus sign instead of the minus sign in the equation gives 14.9/100,000 person-years for the upper bound.

### CONFIDENCE INTERVALS FOR MEASURES OF EFFECT

Studies that measure the effect of an exposure involve the comparison of two or more groups. Cohort studies may be conducted using a fixed follow-up period for

each person. These studies allow direct calculation of risks, which may then be compared. Alternatively, cohort studies may allow for different follow-up times for each person, giving rise to data from which incidence rates may be estimated and compared. Case-control studies also come in more than one variety, depending on how the controls are sampled. Usually, the analysis of case-control studies is based on a single underlying statistical model that describes the statistical behavior of the odds ratio. Prevalence data, obtained from surveys or cross-sectional studies, usually may be treated as risk data for statistical analysis because, like risk data, they are expressed as proportions. Similarly, case-fatality rates, which are more aptly described as data on risk of death among those with a given disease, may usually be treated as risk data.

### Cohort Studies with Risk Data or Prevalence Data

Consider a cohort study of a dichotomous exposure, classified into exposed and unexposed. If the study followed all subjects for a fixed period of time and there were no important competing risks and no confounding, we could display the essential data as follows:

	Exposed	Unexposed
Cases	$a$	$b$
People at risk	$N_1$	$N_0$

From this table, it is easy to estimate the risk difference, RD, and the risk ratio, RR:

$$RD = \frac{a}{N_1} - \frac{b}{N_0}$$

and

$$RR = \frac{a/N_1}{b/N_0}$$

To apply Equation 8-1 and 8-2 to get confidence intervals for the risk difference and the risk ratio, we need formulas for the standard error of the RD and the  $\ln(RR)$ :

$$SE(RD) = \sqrt{\frac{a(N_1 - a)}{N_1^3} + \frac{b(N_0 - b)}{N_0^3}} \quad [9-2]$$

and

$$SE(\ln(RR)) = \sqrt{\frac{1}{a} - \frac{1}{N_1} + \frac{1}{b} - \frac{1}{N_0}} \quad [9-3]$$



### Example: Confidence Limits for Risk Difference and Risk Ratio

As an example of risk data, consider Table 9-1, which describes recurrence risks among women with breast cancer treated either with tamoxifen or with a combination of tamoxifen and radiotherapy. From the data in Table 9-1, we can calculate a risk of recurrence of  $321/686 = 0.47$  among women treated with tamoxifen and radiotherapy and a risk of  $411/689 = 0.60$  among women treated with tamoxifen alone. The risk difference is  $0.47 - 0.60 = -0.13$ , with the minus sign indicating that the treatment group receiving both tamoxifen and radiotherapy had the lower risk. To obtain a 90% confidence interval for this estimate of risk difference, we use Equation 8-1 and 9-2 as follows:

$$\begin{aligned} RD_L &= -0.13 - 1.645 \cdot \sqrt{\frac{321 \cdot 365}{686^3} + \frac{411 \cdot 278}{689^3}} \\ &= -0.13 - 1.645 \cdot 0.027 = -0.17 \\ RD_U &= -0.13 + 1.645 \cdot \sqrt{\frac{321 \cdot 365}{686^3} + \frac{411 \cdot 278}{689^3}} \\ &= -0.13 + 1.645 \cdot 0.027 = -0.08 \end{aligned}$$

This calculation gives 90% confidence limits around  $-0.13$  of  $-0.17$  and  $-0.08$ . The 90% confidence interval for the risk difference ranges from a risk that is 17% lower in absolute terms to a risk that is 8% lower in absolute terms for women receiving the combined tamoxifen and radiotherapy treatment.

We can also compute the risk ratio and its confidence interval from the same data. The risk ratio is  $(321/686)/(411/689) = 0.78$ , indicating that the group receiving combined treatment faces a risk of recurrence that is 22% lower ( $1 - 0.78$ ) relative to the risk of recurrence among women receiving tamoxifen alone. The 90% lower confidence bound for the risk ratio is calculated as follows:

$$\begin{aligned} RR_L &= e^{\ln(0.78) - 1.645 \cdot \sqrt{\frac{1}{321} - \frac{1}{686} + \frac{1}{411} - \frac{1}{689}}} \\ &= e^{-0.24 - 1.645 \cdot 0.051} = e^{-0.327} = 0.72 \end{aligned}$$

Substituting a plus sign for the minus sign before the Z multiplier of 1.645 gives 0.85 for the upper limit. The 90% confidence interval for the risk ratio estimate of

Table 9-1 RISK OF RECURRENCE OF BREAST CANCER IN A RANDOMIZED TRIAL OF WOMEN TREATED WITH TAMOXIFEN AND RADIOTHERAPY OR TAMOXIFEN ALONE

	Tamoxifen and Radiotherapy	Tamoxifen Only
Women with recurrence	321	411
Total women treated	686	689

Data from Feychting et al.<sup>1</sup>

0.78 is 0.72 to 0.85, which is equivalent to saying that the benefit of combined treatment ranges from a 28% lower risk to a 15% lower risk, measured in relative terms. (It is common when describing a reduced risk to convert the risk ratio to a relative decrease in risk by subtracting the risk ratio from unity; a lower limit for the risk ratio equal to 0.72 indicates a 28% lower risk because  $1 - 0.72 = 0.28$ , or 28%.) Keep in mind that these percentages indicate a risk measured in relation to the risk among those receiving tamoxifen alone: the 28% lower limit refers to a risk that is 28% lower than the risk among those receiving tamoxifen alone.

### CONFIDENCE INTERVALS VERSUS CONFIDENCE LIMITS

A confidence interval is a range of values about a point estimate that indicates the degree of statistical precision that describes the estimate. The level of confidence is set arbitrarily, but for any given level of confidence, the width of the interval expresses the precision of the measurement. A wider interval implies less precision, and a narrower interval implies more precision. The upper and lower boundaries of the interval are the confidence limits.

### Cohort Studies with Incidence Rate Data

For cohort studies that measure incidence rates, we use the following notation:

	Exposed	Unexposed
Cases	$a$	$b$
People-time at risk	$PT_1$	$PT_0$

The incidence rate among exposed is  $a/PT_1$ , and that among unexposed is  $b/PT_0$ . To obtain confidence intervals for the incidence rate difference (ID),  $a/PT_1 - b/PT_0$ , and the incidence rate ratio (IR),  $(a/PT_1)/(b/PT_0)$ , we use the following equations for the standard error of the rate difference and the logarithm of the incidence rate ratio:

$$SE(ID) = \sqrt{\frac{a}{PT_1^2} + \frac{b}{PT_0^2}} \quad [9-4]$$

$$SE(\ln(IR)) = \sqrt{\frac{1}{a} + \frac{1}{b}} \quad [9-5]$$

### Example: Confidence Limits for Incidence Rate Difference and Incidence Rate Ratio

The data in Table 9-2 are taken from a study by Feychting et al.<sup>1</sup> that compared cancer occurrence among the blind with occurrence among those who were not blind but had severe visual impairment. The study hypothesis was that a high



Table 9-2 INCIDENCE RATE OF CANCER AMONG A BLIND POPULATION AND A POPULATION THAT IS VISUALLY SEVERELY IMPAIRED BUT NOT BLIND

	Totally Blind	Visually Severely Impaired but Not Blind
Cancer cases	136	1,709
Person-years	22,050	127,650

Data from Petitti et al.<sup>2</sup>

circulating level of melatonin protects against cancer. Melatonin production is greater among the blind because visual detection of light suppresses melatonin production by the pineal gland.

From these data, we can calculate a cancer rate of 136/22,050 person-years = 6.2/1000 person-years among the blind, compared with 1709/127,650 person-years = 13.4/1000 person-years among those who were visually impaired but not blind. The incidence rate difference (ID) is (6.2 - 13.4)/1000 person-years = -7.2/1000 person-years. The minus sign indicates that the rate is lower among the group with total blindness, which is here considered to be the exposed group. To get a 90% confidence interval for this estimate of rate difference, we use Equations 8-1 in combination with Equation 9-4, as follows.

$$ID_L = \frac{-7.2}{1,000 \text{ pyrs}} - 1.645 \cdot \sqrt{\frac{136}{22,050^2} + \frac{1,709}{127,650^2}}$$

$$= \frac{-7.2}{1,000 \text{ pyrs}} - 1.645 \cdot \frac{0.62}{1,000 \text{ pyrs}} = \frac{-8.2}{1,000 \text{ pyrs}}$$

$$ID_U = \frac{-7.2}{1,000 \text{ pyrs}} + 1.645 \cdot \sqrt{\frac{136}{22,050^2} + \frac{1,709}{127,650^2}}$$

$$= \frac{-7.2}{1,000 \text{ pyrs}} + 1.645 \cdot \frac{0.62}{1,000 \text{ pyrs}} = \frac{-6.2}{1,000 \text{ pyrs}}$$

This calculation gives 90% confidence limits around the rate difference, -7.2/1000 person-years, of -8.2/1000 person-years and -6.2/1000 person-years.

The incidence rate ratio for the data in Table 9-2 is (136/22,050)/(1709/127,650) = 0.46, indicating a rate among the blind that is less than one half that among the comparison group. The lower limit of the 90% confidence interval for this rate ratio is calculated as follows:

$$IR_L = e^{\ln(0.46) - 1.645 \cdot \sqrt{\frac{1}{136} + \frac{1}{1,709}}}$$

$$= e^{-0.775 - 1.645 \cdot 0.089} = e^{-0.922} = 0.40$$

A corresponding calculation for the upper limit gives  $IR_U = 0.53$ , for a 90% confidence interval around the incidence rate ratio of 0.46 of 0.40 to 0.53.

## Case-Control Studies

This and later chapters deal with methods for the analysis of a density case-control study or a cumulative case-control study. The analysis of case-cohort studies and case-crossover studies is slightly different and is left for more advanced texts. For the data display from a case-control study, we use the following notation:

	Exposed	Unexposed
Cases	a	b
Controls	c	d

The primary estimate of effect that we can derive from these data is the incidence rate ratio or risk ratio, depending on how controls were sampled. In either case, the effect measure is estimated from the odds ratio (OR),  $ad/bc$ . We obtain an approximate confidence interval for the odds ratio using the following equation for the standard error of the logarithm of the odds ratio:

$$SE(\ln(OR)) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \quad [9-6]$$

### Example: Confidence Limits for the Odds Ratio

Consider as an example the data in Table 9-3 on amphetamine use and stroke in young women, from the study by Petitti et al.<sup>2</sup> For these case-control data, we can calculate an odds ratio (OR) of  $(10)(1,016)/[(5)(337)] = 6.0$ . An approximate 90% confidence interval for this odds ratio can be calculated from the standard error Equation 9-6 in combination with Equation 8-1:

$$OR_L = e^{\ln(6.0) - 1.645 \cdot \sqrt{\frac{1}{10} + \frac{1}{337} + \frac{1}{5} + \frac{1}{1,016}}}$$

$$= e^{1.797 - 1.645 \cdot 0.551} = e^{1.797 - 0.907} = e^{0.890} = 2.4$$

Using a plus sign instead of the minus sign in front of the Z multiplier of 1.645, we get  $OR_U = 14.9$ . The point estimate of 6.0 for the odds ratio is the geometric mean between the lower limit and the upper limit of the confidence interval. This relation applies whenever we set confidence intervals on the log scale, which we do for all approximate intervals for ratio measures. The limits are symmetrically placed about the point estimate on the log scale, but the upper bound appears

Table 9-3 FREQUENCY OF RECENT AMPHETAMINE USE AMONG STROKE CASES AND CONTROLS AMONG WOMEN BETWEEN 15 AND 44 YEARS OLD

	Amphetamine Users	No Amphetamine Use
Stroke cases	10	337
Controls	5	1,016

Adapted from Petitti et al.<sup>2</sup>



farther from the point estimate on the untransformed ratio scale. This asymmetry on the untransformed scale for a ratio measure is especially apparent in this example because the OR estimate is large.

## CALCULATION OF P VALUES

Although the investigator is better off relying on estimation rather than tests of statistical significance for inference, for completeness, we give the basic formulas from which traditional *P* values can be derived that test the null hypothesis that exposure is not related to disease.

### Risk Data

For risk data, we use the following expansion of the notation used earlier in the chapter:

	Exposed	Unexposed	Total
Cases	<i>a</i>	<i>b</i>	<i>M</i> <sub>1</sub>
Noncases	<i>c</i>	<i>d</i>	<i>M</i> <sub>0</sub>
People at risk	<i>N</i> <sub>1</sub>	<i>N</i> <sub>0</sub>	<i>T</i>

The *P* value testing the null hypothesis that exposure is not related to disease can be obtained from the following equation for  $\chi$ :

$$\chi = \frac{a - \frac{N_1 M_1}{T}}{\sqrt{\frac{N_1 N_0 M_1 M_0}{T^2 (T-1)}}} \quad [9-7]$$

For the data in Table 9-1, Equation 9-7 gives  $\chi$  as follows:

$$\chi = \frac{321 - \frac{686 \cdot 732}{1375}}{\sqrt{\frac{686 \cdot 689 \cdot 732 \cdot 643}{1375^2 \cdot 1374}}} = \frac{321 - 365.20}{\sqrt{85.64}} = -4.78$$

The *P* value that corresponds to this  $\chi$  statistic must be obtained from tables of the standard normal distribution (see Appendix). For a  $\chi$  of -4.78 (minus sign indicates only that the exposed group had a lower risk than the unexposed group), the *P* value is very small (roughly 0.0000009). The Appendix tabulates values of  $\chi$  only from -3.99 to +3.99.

### Incidence Rate Data

For incidence rate data, we use the following notation, which is an expanded version of the table we used earlier:

	Exposed	Unexposed	Total
Cases	<i>a</i>	<i>b</i>	<i>M</i>
Person-time	<i>PT</i> <sub>1</sub>	<i>PT</i> <sub>0</sub>	<i>T</i>

for which we can use the following equation to calculate  $\chi$ :

$$\chi = \frac{a - \frac{PT_1 M}{T}}{\sqrt{M \frac{PT_1}{T} \frac{PT_0}{T}}} \quad [9-8]$$

Applying this equation to the data of Table 9-2 gives the following result for  $\chi$ :

$$\chi = \frac{136 - \frac{22,050 \cdot 1845}{149,700}}{\sqrt{1845 \cdot \frac{22,050}{149,700} \cdot \frac{127,650}{149,700}}} = \frac{136 - 271.76}{\sqrt{231.73}} = -8.92$$

This  $\chi$  is so large in absolute value that the *P* value cannot be readily calculated. The *P* value corresponding to a  $\chi$  of -8.92 is much smaller than  $10^{-20}$ , implying that the data are not readily consistent with a chance explanation.

### Case-Control Data

For case-control data, we can apply Equation 9-7 to the data in Table 9-3.

$$\chi = \frac{10 - \frac{15 \cdot 347}{1368}}{\sqrt{\frac{15 \cdot 1353 \cdot 347 \cdot 1021}{1368^2 \cdot 1367}}} = \frac{10 - 3.80}{\sqrt{2.81}} = 3.70$$

From the appendix table, we see that this result corresponds to a *P* value of 0.00022.

## QUESTIONS

1. With person-time data, the numerators of rates are considered Poisson random variables, and the denominators are treated as if they were constants, not subject to variability. Nevertheless, the person-time must be measured and is therefore subject to measurement error. Why are the denominators treated as constants if they are subject to measurement error? What would be the effect on the confidence interval of taking this measurement error into account instead of ignoring it?

2. The approximate formulas for confidence intervals described in this chapter do not work well with small numbers. Suppose 20 people are followed, and 1 develops a disease of interest, giving a risk estimate of  $1/20 = 0.05$ .



The binomial model would give a 90% confidence interval for the risk from -0.03 to 0.13. The lower limit implies a negative risk, which does not make sense. The lower limit should never go below zero, and the upper limit should never go above 1. These risk estimates, based on only one case, are too small for these approximate formulas. Instead, exact formulas based on the binomial distribution can be used. Would you expect a confidence interval for risk calculated from an exact formula to be symmetric around the point estimate (0.05), as the approximate confidence interval is?

3. There is another approximation for obtaining the confidence interval for a binomial proportion that comes closer to the exact method. It is an expression that was proposed in 1927 by Wilson<sup>3</sup>:

$$\frac{N}{N+Z^2} \left[ \frac{a}{N} + \frac{Z^2}{2N} \pm Z \sqrt{\frac{a(N-a)}{N^3} + \frac{Z^2}{4N^2}} \right]$$

In this formula,  $a$  is the number of cases (numerator),  $N$  is the number at risk (denominator), and  $Z$  is the multiplier from the standard normal distribution that corresponds to the confidence level. The  $\pm$  sign gives the lower bound when the minus sign is used and the upper bound when the plus sign is used. Even with only 1 case among 20 people, this formula gives results very close to the exact confidence interval, and its accuracy only improves with larger numbers. What is the 90% confidence interval for the risk estimate of 1/20 using Wilson's equation? If Wilson's equation is so accurate, why do you suppose that it has not been adopted more widely as the usual approach to getting confidence limits for a binomial variable?

4. Why are the estimation equations to obtain confidence intervals the same for prevalence data and for risk data (see Equations 9-2 and 9-3)?

5. Why do the estimation equations for confidence intervals differ for risk data and case-control data (see Equations 9-3 and 9-6), whereas the formula for obtaining a  $\chi$  statistic to test the null hypothesis is the same for risk data and case-control data (see Equation 9-7)?

6. Does it lend a false sense of precision to present a 90% confidence interval instead of a 95% confidence interval?

7. Calculate a 90% confidence interval and a 95% confidence interval for the odds ratio from the following crude case-control data relating to the effect of exposure to magnetic fields on risk of acute leukemia in children<sup>4</sup>:

	Median Nighttime Exposure		
	$\geq 2 \mu\text{T}$	$< 2 \mu\text{T}$	Total
Cases	9	167	176
Controls	5	409	414
Total	14	576	590

## REFERENCES

1. Feychting M, Osterlund B, Ahlbom A. Reduced cancer incidence among the blind. *Epidemiology*. 1998;9:490-494.
2. Petitti DB, Sidney S, Quesenberry C, Bernstein A. Stroke and cocaine or amphetamine use. *Epidemiology*. 1999;9:596-600.
3. Wilson EB. Probable inference. The law of succession and statistical inference. *J Am Stat Assoc*. 1927;22:209-212.
4. Michaelis J, Schütz, Meinert R, et al. Combined risk estimates for two German population-based case-control studies on residential magnetic fields and childhood acute leukemia. *Epidemiology*. 1997;9:92-94.