7

Analyzing Simple Epidemiologic Data

In this chapter, we provide the statistical tools to analyze simple epidemiologic data. By simple data, we refer to the most elementary data that could be obtained from an epidemiologic study, such as crude data from a study with no confounding. Because our emphasis is on estimation as opposed to 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 formulas presented here give only approximate results and are valid only for data with sufficiently large numbers. More accurate estimates can be obtained 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 can 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. True, for those who 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 formula 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, that is, 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 out 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, the model usually applied is the binomial model. To use the model to obtain a confidence interval, it helps to have some simple notation. Let us use *a* to represent cases and *N* to represent people at risk. Using this notation, our estimate of risk would be the number of cases divided by the total number of people at risk: $R = a/N$. We can calculate a confidence interval for the lower and upper confidence limits of *R* using the following formula.

$$R_L, R_U = R \pm Z \cdot \mathrm{SE}(R) \tag{7–1}$$

In this expression, 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 estimate. Under the binomial model, the standard error of *R* would be as follows.

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

Example: Confidence Limits for a Risk or Prevalence

Using this formula with the example of 20 cases of influenza in 100 people, we can calculate the lower bound of a 90% confidence interval for the risk as follows.

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

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

Wilson's confidence limits for a binomial

The application of the preceding formula to risk data to obtain confidence limits is straightforward, but the approach is useful only as a large-number approximation. With scanty data, and especially for risks that are considerably less than (or greater than) 50%, the confidence limits are apt to be inaccurate. Suppose we have 20 people, among whom there is one case, for a risk estimate of 1/20 or 0.05. The text formula on page 131 would give a 90% confidence interval from −0.03 to 0.13. The lower limit is a negative risk, which does not even make sense. The lower limit should theoretically never go below 0, and the upper limit should never go above 1. These numbers are too small to use the binomial standard error formula. Because the risk estimate for this example is 0.05, there is less room for the risk to vary in the low direction than in the high direction, and an exact calculation of the confidence interval would produce limits that were not symmetrically placed around 0.05. The exact procedure, a more complicated calculation that goes beyond the scope of this book, gives a 90% confidence interval of 0.005–0.19. There is an approximate approach that comes close to exact limits for a binomial distribution, however. The formula was proposed in 1927 by Wilson:[1]

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

As before, 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. This formula, much easier than an exact calculation, gives a 90% confidence interval for risk of 0.01–0.20, close to the exact confidence limits even for these small numbers. With this formula, there is little reason to obtain an exact binomial confidence interval for any data.

Incidence Rate Data

For incidence rate data, let us use a to represent cases and PT to represent person-time. Although the notation is similar to that for risk data, these data differ both 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 our estimate of the disease rate and calculate a confidence interval for the rate using formula 7–1 with the following standard error.

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

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 in 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 one measures the entire population 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 super-population; therefore, 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 per 100,000 person-years. A lower 90% confidence limit for the rate would be estimated as follows.

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

$$= 3.9 \text{ / } 100{,}000 \text{ person-years}$$

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

Byar's confidence limits

With very small numbers, the approximate formula given in the text for confidence limits for incidence rate data will be inaccurate. Once again, the ideal would be to calculate exact confidence limits, but as for risk data, there is also a convenient approximate formula that is nearly as good as exact methods and much easier to use. This formula, adapted from D. Byar (unpublished), is:

$$R_L, R_U = \frac{a'\left(1 - \frac{1}{9a'} \pm \frac{Z}{3}\sqrt{\frac{1}{a'}}\right)^3}{PT}$$

where a' equals $a + 0.5$, PT is the rate denominator, and Z is the multiplier from the standard normal distribution that corresponds to the level of confidence. As before, the minus sign is used to calculate the lower limit and the plus sign, the upper limit.

Suppose that the observed rate were 3 cases in 2500 person-years, for a rate of 12 cases per 10,000 person-years. The large-number formula for confidence limits in the text gives a symmetric 90% confidence interval of 0.6–23 cases per 10,000 person-years, whereas the above formula gives a 90% confidence interval of 4.3–28 cases per 10,000 person-years, which is much more accurate. (The exact confidence limits are 4.0 and 29.) Like Wilson's binomial formula and the exact confidence interval, the Byar confidence interval is asymmetrical.

Confidence Intervals for Measures of Effect

Studies that measure the effect of an exposure involve comparison of two or more groups. Cohort studies may be conducted using a fixed follow-up period for each person, obtaining effect estimates from a comparison of risk data; or they may allow for varying follow-up times for each person, obtaining effect estimates from a comparison of incidence rate data. Case-control studies come in several varieties, depending on how the controls are sampled; but for the most part, 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, may usually be treated as risk data for statistical analysis because, like risk data, they represent 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 considered risk data.

Cohort Studies with Risk Data (or Prevalence Data)

Consider a cohort study of a dichotomous exposure with the categories 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 array, one can easily estimate the risk difference, *RD*, and the risk ratio, *RR*.

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

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

To apply formulas 6–1 and 6–2 to obtain confidence intervals for the risk difference and the risk ratio, we need formulas for the standard error of *RD* and the ln(*RR*):

$$SE(RD) = \sqrt{\frac{a(N_1 - a)}{N_1^3} + \frac{b(N_0 - b)}{N_0^3}} \qquad (7\text{–}2)$$

$$SE[\ln(RR)] = \sqrt{\frac{1}{a} - \frac{1}{N_1} + \frac{1}{b} - \frac{1}{N_0}} \qquad (7\text{–}3)$$

Example: Confidence Limits for Risk Difference and Risk Ratio

As an example of risk data, consider Table 7–1, which describes recurrence risks among women with breast cancer treated with either tamoxifen or a combination of tamoxifen and radiotherapy.

Table 7–1. 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 Overgaard et al.[2]

From the data in Table 7–1, 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 formulas 6–1 and 7–2 as follows.

$$RD_L = -0.13 - 1.645 \cdot \sqrt{\frac{321 \cdot 365}{686^3} + \frac{411 \cdot 278}{689^3}}$$

$$= -0.13 - 1.645 \cdot 0.027 = -0.17$$

$$RD_U = -0.13 + 1.645 \cdot \sqrt{\frac{321 \cdot 365}{686^3} + \frac{411 \cdot 278}{689^3}}$$

$$= -0.13 + 1.645 \cdot 0.027 = -0.08$$

This calculation gives 90% confidence limits around -0.13 of -0.17 and -0.08. In other words, the 90% confidence interval for the risk difference ranges from a benefit of 17% smaller risk to a benefit of 8% smaller risk 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.

$$RR_L = e^{\ln(0.78) - 1.645 \cdot \sqrt{\frac{1}{321} - \frac{1}{686} + \frac{1}{411} - \frac{1}{689}}}$$

$$= e^{-0.24 - 1.645 \cdot 0.051} = e^{-0.327} = 0.72$$

Substituting a plus sign for the minus sign before the z-multiplier of 1.645 gives 0.85 for the upper limit. Thus, the 90% confidence interval for the risk ratio estimate of 0.78 is 0.72–0.85. In other words, the 90% confidence interval for this benefit of combined treatment ranges from a 28% lower risk to a 15% lower risk. (It is common when describing a reduced risk to convert the risk ratio to a relative decrease in risk by subtracting it from unity; thus, a lower limit for the risk ratio equal to 0.72 indicates a 28% lower risk because $1 - 0.72 = 0.28$, or 28%.) 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
Person-time at risk	PT_1	PT_0

The incidence rate among the exposed is a/PT_1 and that among the 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 formulas for the standard error of the rate difference and the logarithm of the incidence rate ratio.

$$SE(ID) = \sqrt{\frac{a}{PT_1^2} + \frac{b}{PT_0^2}} \quad (7\text{–}4)$$

$$SE[\ln(IR)] = \sqrt{\frac{1}{a} + \frac{1}{b}} \quad (7\text{–}5)$$

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

The data in Table 7–2 are taken from a study by Feychting et al.,[3] comparing 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 circulating level of melatonin protects against cancer; melatonin production is greater among the blind because, among those who see, visual detection of light suppresses melatonin production by the pineal gland.)

Table 7–2. Incidence rate of cancer among a blind population and a population that is severely visually impaired but not blind*

	Blind	Severely Visually Impaired but Not Blind
Cancer cases	136	1709
Person-years	22,050	127,650

*Data from Feychting et al.[3]

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 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 the "exposed" group. To obtain a 90% confidence interval for this estimate of rate difference, we use formula 6–1 in combination with formula 7–4, as follows.

$$ID_L = \frac{-7.2}{1000 \text{ person-years}} - 1.645 \cdot \sqrt{\frac{136}{22{,}050^2} + \frac{1709}{127{,}650^2}}$$

$$= \frac{-7.2}{1000 \text{ person-years}} - 1.645 \cdot \frac{0.62}{1000 \text{ person-years}} = \frac{-8.2}{1000 \text{ person-years}}$$

$$ID_U = \frac{-7.2}{1000 \text{ person-years}} + 1.645 \cdot \sqrt{\frac{136}{22{,}050^2} + \frac{1709}{127{,}650^2}}$$

$$= \frac{-7.2}{1000 \text{ person-years}} + 1.645 \cdot \frac{0.62}{1000 \text{ person-years}} = \frac{-6.2}{1000 \text{ person-years}}$$

This calculation gives 90% confidence limits around the rate difference of −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 7–2 is (136/22,050)/(1709/127,650) = 0.46, indicating a rate among the blind that is less than half that among the comparison group. The lower limit of the 90% confidence interval for this rate ratio is calculated as follows.

$$IR_L = e^{\ln(0.46) - 1.645 \cdot \sqrt{\frac{1}{136} + \frac{1}{1709}}}$$

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

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 0.40–0.53.

Case Control Studies

Here and in later chapters we deal with methods for the analysis of a density case-control study, the most common form of 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 will 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, which in case-control studies is estimated from the odds ratio (OR), ad/bc. We obtain an approximate confidence interval for the odds ratio using the following formula for the standard error of the logarithm of the odds ratio:

$$\text{SE}[\ln(OR)] = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \qquad (7\text{–}6)$$

Example: Confidence Limits for the Odds Ratio

Consider as an example the data in Table 7–3 on amphetamine use and stroke in young women, from the study by Petitti et al.[3]

For these case-control data, we can calculate an OR of (10)(1016)/[(5)(337)] = 6.0. An approximate 90% confidence interval for this odds ratio can be calculated from the standard error expression 7–6 above in combination with formula 6–1.

$$OR_L = e^{\ln(6.0) - 1.645 \cdot \sqrt{\frac{1}{10} + \frac{1}{337} + \frac{1}{5} + \frac{1}{1016}}}$$

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

Table 7–3. Frequency of recent amphetamine use among stroke cases and controls among women age 15–44*

	Amphetamine Use	No Amphetamine Use
Stroke cases	10	337
Controls	5	1016

*Adapted from Petitti et al.[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 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 reader is better off relying on estimation rather than tests of statistical significance for inference, for completeness we give here 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 will use the following expansion of the notation used earlier in the chapter.

	Exposed	Unexposed	Total
Cases	a	b	M_1
Noncases	c	d	M_0
People at risk	N_1	N_0	T

The p value testing the null hypothesis that exposure is not related to disease can be obtained from the following formula for χ.

$$\chi = \frac{a - \dfrac{N_1 M_1}{T}}{\sqrt{\dfrac{N_1 N_0 M_1 M_0}{T^2 (T - 1)}}} \qquad (7\text{–}7)$$

For the data in Table 7–1, formula 7–7 gives χ as follows.

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

The p value that corresponds to the χ statistic must be obtained from tables of the standard normal distribution (see Appendix). For a χ of

-4.78 (the 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 χ 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	a	b	M
Person-time	PT_1	PT_0	T

We can use the following formula to calculate χ.

$$\chi = \frac{a - \dfrac{PT_1}{T} M}{\sqrt{M \dfrac{PT_1}{T} \dfrac{PT_0}{T}}} \qquad (7\text{–}8)$$

Applying this formula to the data of Table 7–2 gives the following result for χ.

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

This χ is so large in absolute value that the p value cannot be readily calculated. The p value corresponding to a χ 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 formula 7–7 to the data in Table 7–3.

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

This result corresponds to a p value of 0.00022.

Questions

1. With person-time data, the numerators are considered Poisson random variables and the denominators are treated as if they were constants, not subject to variability. In fact, however, 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 usual approximate formula for confidence limits for risk or prevalence data, based on the binomial distribution, will not work if there are zero cases in the numerator. The Wilson formula, however, is still useful in such situations. It gives zero for the lower limit, which is appropriate, and it gives a meaningful upper limit. Suppose that you were interested in the case fatality rate among patients undergoing bypass cardiac surgery in a new cardiac surgery unit. Among the first 30 patients to undergo surgery, none died within 30 days. Using the Wilson formula, calculate a 90% confidence interval for the risk of dying within 30 days after surgery.
3. Why do you suppose that the estimation formulas to obtain confidence intervals are the same for prevalence data and risk data (formulas 7–2 and 7–3)?
4. Why do you suppose that the estimation formulas for confidence intervals differ for risk data and case-control data (formulas 7–3 and 7–6) but the formula for obtaining a χ statistic to test the null hypothesis is the same for risk data and case-control data (formula 7–7)?
5. Does it lend a false sense of precision if one presents a 90% confidence interval instead of a 95% confidence interval?
6. 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.[5]

	Median night-time exposure		
	≥2 μT	<2 μT	Total
Cases	9	167	176
Controls	5	409	414
Total	14	576	590

References

1. Wilson EB: Probable inference. The law of succession and statistical inference. *J Amer Stat Assn* 1927; 22:209–212.
2. Overgaard, M, Jensen, M-B, Overgaard, J, et al: Postoperative radiotherapy in high-risk postmenopausal breast-cancer patients given adjuvant tamoxifen: Danish Breast Cancer Cooperative Group DBCG 82c randomized trial. *Lancet* 1999;353:1641–1648.
3. Feychting, M, Osterlund, B, Ahlbom, A. Reduced cancer incidence among the blind. *Epidemiology* 1998;9:490–494.
4. Petitti, DB, Sidney, S, Quesenberry, C, et al: Stroke and cocaine or amphetamine use. *Epidemiology* 1998;9:596–600.
5. Michaelis, J, Schuz, J, Meinert, R, et al: Combined risk estimates for two German population-based case-control studies on residential magnetic fields and childhood acute leukemia. *Epidemiology* 1998;9:92–94.