4

MEASURES OF DISEASE FREQUENCY

OBJECTIVES

After reading this chapter, you should be able to:

1. Explain the different ways of measuring disease frequency, and differentiate among counts, proportions, odds, risks, and rates.

2. Describe the difference between incidence and prevalence, and when each should be used.

3. Describe the difference between risk and rate as applied to measures of incidence.

4. Elaborate upon the concepts of 'cause-specific measures', proportional morbidity/mortality rates, and case fatality rates.

5. Apply all of the above concepts, and select the appropriate measures of disease frequency to be used in specific circumstances.

6. Compute the appropriate measures when provided with the necessary data, and calculate exact and/or approximate confidence intervals.

4.1 Introduction

Measurement of disease (or event) frequency is the basis for many epidemiological activities. These include routine surveillance, observational research, and outbreak investigations, among others. In observational studies, measuring the frequency of a disease and an exposure, and subsequently linking (or associating) the exposure and the disease, are the first steps to inferring causation. The hypothesis we test is described qualitatively, but the process involves quantification and begins with measurement of events and exposures.

Morbidity and mortality are the 2 main categories of events for which frequency measures are calculated. However, there are other events of interest, such as parturition (giving birth), vaccination, hospital admission, *etc*. The format for calculating these is the same as it is for morbidity and mortality.

Because both morbidity and mortality are strongly associated with an individual's attributes, and because different diseases have different impacts, we often calculate these measures for specific host attributes (*eg* age, sex, and race) and for specific diseases (*ie* outcomes of interest).

4.1.1 Some factors affecting the choice of frequency measure

Study period When selecting a measure of disease frequency for use in a study, it is important to consider both the study period and the risk period. The study period is the time interval during which the study subjects are observed for the outcome (usually a disease) of interest. It is usually measured in terms of calendar time, but sometimes the study period is a point in time. In either instance, the study period could be specified by calendar time, by the event at which the outcome data are collected (*eg* at birth) or both (*eg* congenital defects during 2008–2010).

Risk period The risk period is the time during which the individual could develop the disease of interest. Thus an important question is: how long is the risk period? For example, for diseases such as post-partum eclampsia, the risk period is generally short—usually less than 2 days—whereas for diseases such as migraine headaches, the risk period is 'lifelong'.

Both the risk and study period relate to whether the population is deemed to be closed or open (see Section 4.4.1). However, disregarding this, diseases with a short risk period (relative to the study period) are good candidates for risk measures. Diseases with long risk periods are likely candidates for rate-based measures. These 2 approaches to measuring the incidence of disease are discussed in Section 4.3.

4.2 Counts, proportions, odds, and rates

Before specific measures of disease frequency can be discussed, it is necessary to review the mathematical forms that these measures can take: counts, proportions, odds, and rates.

Count This is a simple enumeration of the number of cases of disease or number of individuals affected with a condition in a given population. Because the size of the population is not taken into consideration, counts of events are of **very** limited use for epidemiologic research.

Proportion This is a ratio in which the numerator is a subset of the denominator. For example, if 188 samples are tested for norovirus (as in the data used in Chapter 5) and 82 of them are positive, the proportion positive is 82/188=0.436 (or 43.6%). Prevalence (Section 4.7) and risk

(Sections 4.3 and 4.4) are both proportions. In the former, both the numerator and denominator are measured at a point in time. In the latter, the numerator relates to the number of new cases over a period of time so, although proportions have no units, the period must be specified for the proportion to make sense.

Odds This is a ratio in which the numerator is not a subset of the denominator. For example, if there are 3 stillbirths and 120 live births, the odds of stillbirth is 3:120=0.025:1 or 25 stillbirths to 1,000 live births. The odds of norovirus (based on the data given above) is 82/106=0.77 (or approximately 3:4).

Rate A rate is a ratio in which the denominator is the number of person-time units at risk. For example, if there are 30 cases of diarrhea in a 100-patient nursing home over a 3-month period, the incidence rate is 30/(100*3)=0.1 cases per person-month. Note the 300 person-months in the denominator.

Note The term 'rate' is often used in a general sense to refer to all types of measures of disease frequency. Strictly speaking, though, it should only be used to refer to measures based on the concept of person-time units. Similarly, we often say that people with a high 'chance' of having or getting the disease have a high 'risk', although the underlying measure of frequency might not be a risk.

4.3 INCIDENCE

Incidence relates to the number of new events (*eg* new cases of a disease) in a defined population within a specific period (Vandenbroucke, 1985). Because incidence deals with new cases of disease, studies based on incident cases of disease are used to identify factors associated with a person becoming ill. Although incidence deals with 'new cases' of disease, it does not necessarily imply just the 'first case' within an individual. For some diseases (*eg* migraine headaches), multiple cases are possible within an individual over time. Similarly, multiple cases of glaucoma are possible in one individual, because either eye could be affected.

For reasons perhaps related to their unique susceptibility, or due to the effect of the first disease occurrence, people who develop one case of a disease are often at a higher risk of developing a subsequent case. Thus, it might be preferable to count only the first case in terms of a disease frequency measure, but to enumerate separately the number of occurrences per individual in the study period. Regardless of whether you are considering only first cases of a disease or all cases, it is imperative that you have a clear case definition (*ie* what criteria need to be met for a 'case' to be considered as such). For the estimate to be reliable, you also need a surveillance programme capable of identifying all such cases.

There are 4 ways of expressing incidence:

- incident times
- incidence count
- incidence risk (*R*)
- incidence rate (*I*).

Incident times are the times at which incident cases occur. They are usually measured as the elapsed time since a reference event (*eg* days after parturition to the onset of post-partum eclampsia), but the reference time may be common for a whole population (*eg* days after

exposure to an environmental toxin). Incident times form the basis of survival analyses which are discussed at length in Chapter 19, and will not be considered further in this chapter.

Incidence count is the simple count of the number of cases of disease observed in a population. It is often used to describe the frequency of a disease in a population in which the disease did not previously exist or was very rare (*eg* country X has had 12 cases of new variant Creutzfeldt-Jakob Disease (nv-CJD)). It might also be used for some common diseases (*eg* case counts of *Salmonella*), but without data on the number of samples/people examined, there are limits to the inferences we can make from count data. Incidence counts are rarely used in epidemiologic research unless they are combined with information about the population at risk (*eg* Poisson regression, Chapter 18). Incidence counts are sometimes expressed as **absolute rates**, in which the number of cases of disease is related to the time period of observation. For example, if the 12 cases of nv-CJD were observed over 4 years, the absolute rate would be 3 cases per year.

Incidence risk An incidence risk (R) is the probability that an individual will contract or develop a disease in a defined time period. Risk, as a measure of frequency, should be restricted to closed populations (Section 4.4.1) where the individual is observed for the full risk period. Because risk is a probability, it is dimensionless (that is, it has no units) and ranges from 0 to 1. Although risk is dimensionless, the time period to which the risk applies must be specified. For example, the risk of a 50-year old male having a heart attack in the next decade is very different (*ie* much higher) than his risk of having one in the next week. In addition, only the first occurrence of a disease in the period of interest is relevant because, once an individual has had one case, it contributes to the numerator of the proportion and what happens to it after that is irrelevant. Risk is used in studies in which making individual predictions is the objective. For example, a study might determine that the probability of recurrence of breast cancer (following mastectomy), over the next year is 14%. This would be referred to as the 1-year risk. Incidence risk is sometimes referred to as **cumulative incidence**. In the context of survival analysis (Chapter 19), survival ((S)—staying free of the event) is defined as: S=1-R.

Incidence rate An incidence rate (I) is the number of new cases of disease in a population per unit of person-time during a given time period. It has units of 1/person-time, and is positive without an upper bound. If a nursing home housing 50 residents has 72 cases of upper respiratory disease over a period of a year, the incidence rate is 72/50, which is 1.44/person-year (or 0.12/person-month). Incidence rates are used in studies designed to determine what factors are related to diseases and what the effects of those diseases are. Incidence rates are sometimes referred to as **incidence density**. A related concept is the hazard rate, which expresses the theoretical limit of I as the time period approaches zero. Hazard rates are used in survival analysis.

4.4 Calculating risk

Risk focuses on individuals, whereas incidence rate (Section 4.5) focuses on cases of disease. Risk can be expressed either at the individual level (*eg* the probability of recurrence of breast cancer within the next year) or at the population level (*eg* the proportion of women who have a recurrence of breast cancer within a year following mastectomy). Rothman *et al* (2008) distinguish between these 2 measures and refer to the latter as an **incidence proportion**. We will use the term risk for both measures, but we recognise that it can only be estimated from a population.

Risk of disease is estimated as:

$$R = \frac{\text{number of newly affected individuals in a defined time period}}{\text{the population at risk}} \qquad \textit{Eq 4.1}$$

4.4.1 Population at risk

While counting the new cases of disease presents some challenges, estimating the population at risk can be even more difficult. The population at risk might be considered 'closed' or 'open'. Regardless of whether the population is closed or open, only people free of the disease at the start of the study period are considered to be at risk.

Closed population A closed population is one in which there are no additions to the population for the duration of the study, and few to no losses. The duration of the study might be defined in terms of calendar time (*eg* residents of a nursing home followed for the next year) or in terms of some life event (*eg* women followed for the first week after giving birth—regardless of when the birth occurred—to determine the risk of post-partum eclampsia). Only disease-free individuals in the population at the start of the study period are considered to be at risk, and are monitored for the outcome of interest. People who are lost to follow-up during the study period are called **withdrawals**, and the simplest way of dealing with them is to subtract half of the number of withdrawals from the population at risk when computing R (this assumes that, on average, the withdrawals leave halfway through the study period). This correction for withdrawals is derived from (or related to) actuarial life-table methods. Unless there are no withdrawals, the risk estimate is biased. Nonetheless, provided the number of withdrawals is small relative to the population size being studied, the bias is small.

Open population An open population is one in which individuals are leaving and entering the population throughout the study period. For example, if you wanted to determine the frequency of recurrence of breast cancer following mastectomy over a one-year period in a population of women served by a single cancer centre, the population at risk would be an open population of people who had mastectomies at that centre. An open population is considered to be **stable** (also referred to as **stationary** or **steady state**) if the rate of additions and withdrawals and the distribution of host attributes are relatively constant over time.

It is not possible to compute risk directly from an open population, but it can be estimated from I (Section 4.6). Risk can also be estimated in open populations using methods for the analysis of 'survival' data (Chapter 19).

Sometimes we can define a follow-up period after a specified exposure/event in a manner that converts an open population to a closed population. For example, a cancer centre is inherently open in the sense that new patients are continually entering the at-risk group. However, if we observe a set of women, post-mastectomy, for a full, defined risk period, then the population becomes closed.

4.5 Calculating incidence rates

Incidence rates are calculated as:

$$I = \frac{\text{number of new cases of disease in a defined time period}}{\text{number of person-time units at risk during the time period}} \qquad \textit{Eq 4.2}$$

A **person-time unit** is one person observed for a defined period (*eg* a person-month, a person-day).

As noted above, incidence rates can be calculated using only the first occurrence of disease for any given individual (from then on they are not considered to be at risk), or using all occurrences of disease. For example, while death can occur only once in an individual's lifetime, some diseases (*eg* migraine headaches) can occur many times. However, even for diseases that occur multiple times, we might be primarily interested in a person's first case, as risk factors for a first case might be different than risk factors for recurrences.

Note The inverse of I ($1/I$) is an estimate of the average time to the occurrence of the disease if the population is closed, or open and stable, providing the outcome is inevitable (all individuals have it if they live long enough).

As with calculating the number of people at risk for R, there are several methods for calculating person-time units at risk for I. The exact method is always preferred, but often the information is not available for you to use the exact method and an approximation must be substituted.

Exact or approximate methods can be adapted for situations in which people are at risk for multiple disease episodes, as opposed to only one disease episode per person. The important thing to remember is that, if you are only interested in the first case of disease, then after the individual contracts the disease of interest, they are **no longer** at risk and no longer contributes to the pool of person-time units at risk, even if they remain in the study.

Exact calculation An exact calculation requires that the exact amount of person-time contributed by each member of the study population be known. Example 4.1 presents a simple exact calculation.

Approximate calculation If only one case of disease per individual is considered, then I is calculated as:

$$I = \frac{\text{cases}}{(\text{start} - 1/2\ \text{sick} - 1/2\ \text{wth} + 1/2\ \text{add}) * \text{time}} \qquad \textit{Eq 4.3}$$

where: cases = # of new cases
start = # at risk at start of study period
sick = # developing disease
wth = # withdrawn from the population
add = # added to the population
time = length of study period (same for all individuals).

If multiple cases of disease per individual are possible, then I is calculated as:

$$I = \frac{\text{cases}}{(\text{start} - 1/2\ \text{wth} + 1/2\ \text{add}) * \text{time}} \qquad \textit{Eq 4.4}$$

Note Both the exact and approximate calculations take into account the fact that individuals withdrawn from a population no longer contribute to the time at risk (Bendixen, 1987). Also, for relatively rare diseases, the second formula might be used even if the investigator is only interested in 'first cases', because the adjustment to the average population at risk by removing those cases will be very small.

Example 4.1 Exact incidence rate calculation

Assume 4 previously healthy people were observed for exactly 1 month (30 days). The history for each individual was as follows:

1 person not sick at all		1.00	person-month at risk
1 person sick on day 10		0.33	person-months at risk
1 person sick on day 20		0.67	person-months at risk
1 person moved away on day 15 (and lost to follow-up)		0.50	person-months at risk
Total 'population at risk'	=	2.50	person-months at risk
Total new cases of disease	=	2	
$I = 2/2.5$	=	0.80	cases/person-month

In general, if the risk period is much shorter than the study period, using risk as a measure of disease is appropriate. If the risk period is longer than the study period, then I is a more appropriate measure of disease incidence and the question of whether only one case, or all cases of disease will be counted must be considered.

4.6 Relationship between risk and rate

Another approach to estimating risk is to use the functional relationship between R and I. If complete data are available for a closed population then:

$$R = A/N \quad \text{and} \quad I = A/(N\,\Delta t)$$

so

$$R = I\,\Delta t$$

where A = number of cases, N = population at risk and Δt = length of study period.

If the population can only be considered closed for short subintervals of the study period, and incident risks or rates in those subintervals are known and small, we can make use of the fact that for small values of x (*eg* $x<0.1$)

$$x \approx 1 - e^{-x} \qquad \textit{Eq 4.5}$$

Thus, if $I\Delta t<0.1$ is a subinterval, then $R \approx I\Delta t$ for that subinterval. The risk for the full study period (consisting of k subintervals) is then

$$R = 1 - \exp\left(-\sum I_k \Delta t_k\right) \qquad \textit{Eq 4.6}$$

Calculations based on deaths among 100 people in a nursing home experiencing a norovirus outbreak over 6 weeks are shown in Table 4.1.

Table 4.1 Estimation of R from average I

Week k	Population at risk N_k	Cases A_k	Weekly I I_k
1	100	1	0.0100
2	99	2	0.0202
3	97	1	0.0103
4	96	3	0.0313
5	93	1	0.0108
6	92	0	0.0000
Total		8	0.0826

The estimate of the 6-week risk is

$$R = 1-\exp\left(-\sum I_k \Delta t_k\right) = 1-e^{(-0.0826)} = 0.079$$

However, if only an average rate I (8 deaths in 100-1/2*8=96 person-weeks, I=0.0833) is available for a population, then assuming that I is constant over the time period:

$$R = 1-e^{-I\Delta t} = 1-e^{-0.0833} = 0.080 \qquad \textit{Eq 4.7}$$

4.7 Prevalence

Prevalence relates to cases of disease existing at a specific point in time rather than new cases occurring over a period of time. Hence, the prevalence count is the number of individuals in a population having an attribute or disease at a particular time.

The prevalence proportion (P) (also referred to simply as prevalence) is calculated as:

$$P=\frac{\text{cases}}{\text{par}} \qquad \textit{Eq 4.8}$$

where cases = # of cases of disease in a population at a point in time

par = # of individuals in the population at risk at the same point in time.

For example, if you test 75 athletes for performance-enhancing drug-use, and 3 test positive, then:

$$P=\frac{3}{75}=0.04=4\%$$

Relationship between prevalence and incidence In a stable population in which I of a disease remains constant (which it rarely does for contagious diseases), P (at any point in time) and I and disease duration (D) are related as follows:

$$P=\frac{I*D}{I*D+1} \qquad \textit{Eq 4.9}$$

For example, if the incidence rate of influenza in an urban population is 0.3/person-year (*ie* 30 cases/100 people per year), and the mean duration of an infection is 3 weeks (0.058 years), then

we would expect P to be:

$$P = \frac{0.3 * 0.058}{0.3 * 0.058 + 1} = 0.0171 = 1.7\%$$

so, on any given day throughout the year, we would expect 1.7% of individuals to have the flu. Of course, this completely ignores the strong seasonal pattern to influenza; more complex formulae that take this variability into account are required to determine the expected prevalence at any point in time (see Alho (1992) for details).

A series of prevalence studies is often used to determine I of diseases which are not easily detected on the basis of clinical signs. This is particularly relevant for determining the rate at which individuals become infected with a certain pathogen. For example, by collecting blood samples from a population of people at regular intervals over their life and testing for toxoplasmosis, the rate at which people are becoming infected can be estimated.

Note P is less useful than I for research into risk factors for diseases because factors that contribute to either the occurrence of disease or its duration will both affect prevalence.

Example 4.2 shows the calculation of various measures of P, R, and I.

4.8 Mortality statistics

These statistics are calculated in exactly the same way as P, R, and I. The disease event of interest in these statistics is by definition, death. The term **mortality rate**, strictly speaking, refers to the incidence rate of mortality. However, it is often misused to describe the risk of mortality. You should be alert to this and interpret the literature accordingly. Overall, the mortality rate describes the number of individuals that die from all causes in a defined time period, and is analogous to I except that the outcome of interest is death.

The **cause-specific mortality rate**, as one would expect, describes the number of individuals that die from (or with) a specific disease during a defined period. This is also calculated the same as I.

Mortality statistics can describe the number of deaths due to a disease or the number of deaths with a disease, but it is often difficult to determine the specific cause of death. For example, if a recumbent patient regurgitates and contracts aspiration pneumonia and then dies, did they die:

- from the initial condition causing recumbency?
- due to pneumonia?
- with pneumonia?

Usually the ‘cause’ will be the factor which is deemed to be the proximate cause (*ie* the straw that broke the back). As indicated above, that might be a difficult determination to make.

4.9 Other measures of disease frequency

Virtually all disease frequency measures can be defined in terms of P, R, and I provided the outcome of interest, the population at risk, and the study period are adequately defined. However, a few specific terms that appear frequently in the literature warrant some attention. Most of these are referred to as rates, but they are really measures of risk.

Example 4.2 Calculation of risk and rate

You are interested in determining the frequency of new malarial infections in young children in a tropical country. You examine blood smears annually from 1–6 years of age for malarial parasites. The results are presented in the table below. A child is only considered to have a new infection if it was negative on the preceding sample, except for samples taken at year 1 (all positives are considered to be new infections at that point).

	Sampling times						Total years at risk	
Child	1	2	3	4	5	6	First case only	All cases
A	0	X	0	0	X	X	2	4
B	0	0	0	–	–	–	3	3
C	X	0	0	X	X	X	1	3
D	0	0	0	0	0	0	6	6
E	0	0	X	0	X	X	3	4

where:

X = positive culture

X = positive culture that represents a new infection

0 = negative culture

– = lost to follow-up

par = population at risk

a) risk of new infection in first 2 years of life
par=5 children
new infections=2
2-year R=2/5=0.4

b) risk of first infection in first 6 years of life
par=5 - 1/2 (1 withdrawal)=4.5 children
first infections=3
6-year R=3/4.5=0.67

c) rate of new infections (considering first cases only)
par=15 person-years
new infections=3
I=3/15=0.2 cases/person-year

d) rate of new infections (considering all new infections)
par=20 person-years
(*eg* person A at risk for years 1, 2, 4, 5)
new infections=6
I=6/20=0.3 cases/person-year

e) 6-year risk estimated from annual rate (first cases only)
I=0.2 cases/person-year
$R=1-e^{-0.2}=0.18$

f) prevalence at 6 years of age
par=4 children
existing infections=3
P=3/4=0.75

Note We are using the sampling time as the time of occurrence (or withdrawal). Some might prefer to use the midpoint between samplings; we have not done this aiming to keep the calculations simple.

4.9.1 Attack rates

Attack rates are used to describe the frequency of disease in outbreak situations. They are computed as the number of cases divided by the size of the population exposed. Consequently, they are really a measure of risk. Attack rates (risk) are used in situations such as outbreaks where the risk period is limited and all cases arising from the exposure are likely to occur within that risk period.

4.9.2 Secondary attack rates

Secondary attack rates are used to describe the 'infectiousness' (or ease of spread) of living agents. The assumption is that there is spread of an agent within the aggregate (*eg* family) and that not all cases are a result of a common-source exposure. When the latent period is long, it is often difficult to distinguish between individual-to-individual spread and that due to common exposure. Secondary attack rates are the number of cases minus the initial case(s) divided by the population at risk.

4.9.3 Case fatality rates

The case fatality rate describes the proportion of individuals with a specific disease that die from it (within a specified time period). It is actually a 'risk' measure (*ie* a proportion) rather than a 'rate', and is often used to describe the impact of epidemic-type diseases or the severity of acute diseases for affected individuals.

4.9.4 Proportional morbidity/mortality rates

These rates are used when the appropriate denominator is unknown, and they are calculated by dividing the number of cases (or deaths) due to a specific disease by the number of cases (or deaths) from all diseases diagnosed. Proportional morbidity/mortality rates are often used for diagnostic laboratory data, and are subject to variation in the numerator or the denominator. Hence, they are less preferable than measures of risk.

4.10 Standard errors and confidence intervals

When estimating a rate or proportion (*eg* risk, prevalence), you usually also want an estimate of its standard errors (SE) as a measure of the precision of the estimate. The SE for a proportion is:

$$\mathrm{SE}(p) = \sqrt{p(1-p)/N} \qquad \textit{Eq 4.10}$$

where p is the estimate of the proportion and N is the sample size. The SE for an incidence rate is:

$$\mathrm{SE}(p) = \sqrt{A/t^2} \qquad \textit{Eq 4.11}$$

Where A is the number of cases and t is the time at risk.

Approximate CIs can be computed based on the estimate (θ) and the SE of the parameter of interest. The lower and upper limits of the CI are then:

$$\theta - Z_{\alpha} * \text{SE} \quad , \quad \theta + Z_{\alpha} * \text{SE} \qquad \textit{Eq 4.12}$$

where Z_{α} is the (1-α/2) percentile of the standard normal distribution.

However, in small samples, or in situations where the frequency of disease is very low (or very high), the approximate CIs might be misleading (and lower limits might be negative). In such cases, exact CIs based on probabilities derived from the binomial distribution (for proportions) or the Poisson distribution (for rates) will be more appropriate. See Agresti and Coull (1998), Newcombe (1998), and Vollset (1993) for additional discussion on confidence intervals. Example 4.3 shows the calculation of approximate and exact CIs for a prevalence proportion, and of exact CIs for some estimated incidence rates.

Example 4.3 Confidence intervals for proportion and rate
data = gi_surv

Incidence data on acute gastrointestinal (GI) disease were obtained from 3 surveys in Canadian populations. Data on the occurrence of any acute GI disease and diarrhea, specifically over the preceding month, were obtained by telephone surveys. See Chapter 31 (gi_surv) for a more complete description of these data.

Approximate and exact CIs for the risk of any GI disease and diarrhea in children from 0–4 years of age in this dataset were computed.

Disease	CI type	Number of positives	Risk	SE	95% CI	
any GI disease	approximate	46	0.134	0.018	0.098	0.170
	exact				0.100	0.174
diarrhea	approximate	19	0.055	0.012	0.031	0.079
	exact				0.034	0.085

Exact CIs were slightly wider. In extreme cases, approximate CIs might go beyond the theoretically possible boundaries of 0 and 1.

Incidence rates were computed by assuming that:

- each month was 30 days long
- each case of GI disease removed 7 days from a person's at-risk time
- multiple cases of GI disease were possible during the 1-month period

The exact CIs for the incidence of any GI disease (for children aged 0–4 years) was then determined based on the Poisson distribution. Of the 46 children with GI disease, 12 had 1 episode and 34 had 2 (total of 80 episodes).

Disease	Number of cases	Child-months at risk	I	SE	Exact 95% CI	
any GI disease	80	9675	0.0083	0.0009	0.0066	0.0103

The incidence rate of GI disease was 0.0083, or 8.3 cases per 1,000 child-months.

4.11 Standardisation of risks and rates

4.11.1 Accounting for differences in populations

Often our intent is to describe the occurrence of disease in a manner that allows valid inferences to be made about factors which affect the frequency of specific diseases. Frequently, host factors are confounders and bias the comparison of risks (rates), whether they be from different geographical areas or have a different exposure history. This confounding can be prevented by standardising the risks or rates. See Chapter 13 for a more complete discussion of confounding.

'**Technical' aspects**
A population might be divided into strata (denoted by the subscript *j*), based on one or more host characteristics (*eg* age, sex, geographical location). The overall frequency of disease in the population is a function of the size of the strata determined by the host factor(s) (denoted here as H_j) and the rates (I_j) or risks of disease (R_j) in each of the strata. The H_j for risks is N_j/N (the proportion of the study group or population in that stratum), and for rates the H_j is T_j/T (the proportion of individual-time in that stratum). Specifically, the crude risk (R) in a population is:

$$R = \sum H_j R_j \qquad \textit{Eq 4.13}$$

where $H_j = N_j/N$

And the crude rate (I) is:

$$I = \sum H_j I_j \qquad \textit{Eq 4.14}$$

where $H_j = T_j/T$.

Note For simplicity, we will primarily refer to rates for the rest of this discussion, but the methods apply equally to risks.

Differences in disease rates (I) between populations of individuals might be due to different distributions of host characteristics (H_j), or to actual differences in the stratum-specific rates (I_j). We can remove the effect of differences in host characteristics by 'standardising' the risks or rates. We can carry out this standardisation by using a set of standard rates (I_j) from a referent population (called **indirect standardisation**), or by using a set of H_j from a standard population (called **direct standardisation**).

4.11.2 Indirect standardisation of rates

One method to control the potential confounding effect of host characteristics when comparing rates from different populations is to compute standardised morbidity/mortality ratios (*SMR*). These are based on a set of stratum-specific rates from a reference or standard population (Is_j) together with the observed proportion of individual-time in each of the strata in the study group. The process is called indirect standardisation. It is useful if the actual stratum-specific rates are not available for the study population, or if the estimates of those rates are based on small sample sizes.

The standard rates from the reference population will allow us to calculate the adjusted or expected rate (I_e) as:

$$I_e = \sum H_j Is_j \qquad \textit{Eq 4.15}$$

The expected number of cases in the study population (denoted as if the reference population rates apply) is:

$$E = T * I_e \qquad \textit{Eq 4.16}$$

where T is the total time at risk.

If A is the observed number of cases in the area, the ratio A/E is the standardised morbidity rate ratio (similarly $I/I_e=SMR$). To obtain the indirect standardised rate (I_{ind}), we use the overall rate in the standard population (Is) multiplied by the SMR.

$$I_{ind} = Is * SMR \qquad \textit{Eq 4.17}$$

The SE of the log of the standardised rate ratio [lnSMR] is:

$$\mathrm{SE}[\ln(SMR)] = 1/\sqrt{A} \qquad \textit{Eq 4.18}$$

and the confidence limits for the SMR can be calculated using:

$$e^{[\ln(SMR)] \pm Z_\alpha * \mathrm{SE}} \qquad \textit{Eq 4.19}$$

4.11.3 Indirect standardisation of risks

We can use the same strategy for risks as described above for rates. The only difference is that H_j is based on the proportion of individuals in each stratum, rather than on the proportion of person-time. The expected number of cases, if the reference population risks apply to the study group's distribution of individuals, is $E=N*Rs$, where Rs is the overall risk in the standard population. The ratio of observed to expected cases, A/E, is the standardised morbidity risk ratio. Again, the indirect standardised risk for the area is $Rs*SMR$. The variability of an SMR based on risks is somewhat more complex than one based on rates and, because most standardisation is done on rates, the formulae for variance will not be given here. Example 4.4 demonstrates the indirect standardisation of risks.

4.11.4 Direct standardisation of rates

We can also address the problem through direct standardisation. Here, we use a standard distribution of the population time-at-risk in each level (stratum) of the confounder (or combination of confounders) for the factor(s) of interest (*ie* the Ts_j). The direct standardised rate (I_{dir}) is:

$$I_{dir} = \sum Ts_j I_j \qquad \textit{Eq 4.20}$$

where Ts_j is the proportion of the total subject time-at-risk allotted to the j^{th} stratum of subjects.

A major drawback to the direct method is that there is no adjustment for the variance of the stratum-specific rates; they all have equal weight even if they are based on a very few individuals.

Example 4.4 Indirect standardisation of risks
data = gi_surv

Assume you would like to compare the GI survey data from the 3 geographical regions of Canada. However, the proportion of people in each of 4 age categories that you have recorded differs across the regions and you know that this factor influences the risk of GI disease. You obtain a set of standard age-specific, monthly incidence risks (Rs_j) based on data from the whole country:

- 0–4 years 20% affected
- 5–9 years 15% affected
- 10–69 years 12% affected
- 70+ years 10% affected
- overall risk 15%

The data for Study A are:

Type	Number of person-months (T_j)	Person-months distribution (H_j)	Observed number of cases	Observed risk (R_j)	Standard risk (Rs_j)	Expected number of cases
0–4 years	137	0.034	28	0.204	0.2	27.4
5–9 years	147	0.037	20	0.136	0.15	22.1
10–69 years	3271	0.817	469	0.143	0.12	392.5
70+ years	451	0.113	27	0.060	0.10	45.1
Total	4006		544			487.1
Overall risk*				0.136	0.122	

SMR = .136/.122 = 544/487.07 = 1.12

Indirect standardised risk (I_{ind}) = .15 * 1.12 = .168

* Overall risk is the sum of the standard stratum-specific risk times the H_j distribution (eg overall observed risk in Study A=(.204*.034)+(.136*.037)+(.143*.817)+(.06*.113)=0.136

Across all studies, the crude (observed) and adjusted risks were:

Study	Cases observed	Crude risk	Adjusted risk	95% CI	
1	544	0.136	0.168	0.154	0.182
2	417	0.125	0.154	0.139	0.169
3	179	0.088	0.111	0.095	0.128

... and the SMRs were:

Study	Cases observed	Cases expected	SMR	95% CI	
1	544	487.1	1.12	1.02	1.21
2	417	406.8	1.03	0.93	1.13
3	179	242.9	0.74	0.63	0.85

The crude and adjusted risks were different in each of the 3 studies. The *SMR*s show that the observed risk in Study 1 was higher, and in Study 3 was lower than would have been expected given the age structure of the sample.

To express the variability of the direct standardised rate, the SE is:

$$\mathrm{SE}(I_{\mathrm{dir}}) = \sqrt{\sum (Ts_j^2 * I_j * S_j / N_j)} \qquad \textit{Eq 4.21}$$

where Ts_j is the proportion of the total subject time-at-risk allotted to the j^{th} stratum of subjects.

where S_j=1-I_j.

The confidence interval can be calculated using:

$$I_{\mathrm{dir}} \pm Z_{\alpha} * \mathrm{SE}(I_{\mathrm{dir}}) \qquad \textit{Eq 4.22}$$

The **direct standardisation of risks** proceeds in an analogous manner to that of rates. The actual proportion of individuals (Hs_j) in each category in the reference population is used instead of the proportion of person-time (Ts_j) in each category. Example 4.5 demonstrates the direct standardisation of risks.

Example 4.5 Direct standardisation of risks
data = gi_surv

Using the same data presented in Example 4.4, and using a reference population based upon the data from all 3 studies pooled, the reference population distribution (Ts_j) was:

- 0–4 year 3.2% of people
- 5–9 years 3.7% of people
- 10–69 years 80.5% of people
- 70+ years 12.6% of people

Direct standardised risks for Study A were:

Age group	Observed risk (R_j)	Reference population distribution (Ts_j)	Product (R_j* Ts_j)
0–4 years	0.2044	0.032	0.0066
5–9 years	0.1361	0.037	0.0050
10–69 years	0.1434	0.805	0.1155
70+ years	0.0599	0.126	0.0075
Direct standardised risk (I_{dir})			0.1346

The crude and adjusted risk for the 3 studies were:

Study	N	Crude risk	Adjusted risk	95% CI	
1	4006	0.136	0.135	0.124	0.145
2	3339	0.125	0.125	0.113	0.136
3	2037	0.088	0.090	0.077	0.102

The differences between the crude and adjusted risks are smaller than with indirect standardisation, primarily because data from the 3 studies were used for the reference population.

4.11.5 Application of standardisation

There are a number of areas where rate standardisation is really useful. It allows us to compare a set of rates without being concerned about whether or not they are confounded—provided we can measure the confounders. Rate standardisation works best when the confounders are categorical in nature. Age standardisation of cancer statistics is one common application of these techniques, *eg* the rates published by the Canadian Cancer Society (cancer.ca).

References

Agresti A, Coull BA. Approximate is better than "exact" for interval estimation of binomial proportions. The American Statistician. 1998;52:119-26.

Alho JM. On prevalence, incidence, and duration in general stable populations. Biometrics. 1992;48 (2):587-92.

Bendixen PH. Notes about incidence calculations in observational studies. Prev Vet Med. 1987;5:151-6.

Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. Stat Med. 1998;17 (8):857-72.

Rothman KJ, Greenland S, Lash TL. Modern Epidemiology, 3rd Ed. Philadelphia: Lippincott Williams & Wilkins; 2008.

Vandenbroucke JP. On the rediscovery of a distinction. Am J Epidemiol. 1985;121 (5):627-8.

Vollset SE. Confidence intervals for a binomial proportion. Stat Med. 1993;12 (9):809-24.