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 culling (the premature removal of animals from a herd or flock), survival to weaning, and pregnancy (*eg* the probability of an animal becoming pregnant within a specified time period). The format for calculating these is the same as it is for morbidity and mortality.

Because both morbidity and mortality are strongly associated with animal (or herd) attributes, and different diseases have different impacts, we usually calculate these measures for specific host attributes (*eg* age, sex, and breed) 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 period of time over which the study is conducted. 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 in calendar time or by the event at which the data are collected (*eg* at slaughter or at birth).

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 retained placenta in dairy cows, the risk period is short—a day or 2 at most; whereas, for diseases such as lameness or foot problems, 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 COUNT, PROPORTION, ODDS AND RATE

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

Count This is a simple enumeration of the number of cases of disease or number of animals 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 200 cows are tested for enzootic bovine leukosis (EBL) and 40 of them are positive, the

proportion positive is 40/200=0.2 (or 20%). Prevalence (Section 4.7) and risk (Sections 4.3, 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 time 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 stillborn animals 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 EBL (based on the data given above) is 40/160=0.25 (or 1:4).

Rate A rate is a ratio in which the denominator is the number of animal-time units at risk. For example, if there are 30 cases of kennel cough in a 100-dog kennel over a 3-month period, the incidence rate is 30/(100*3)=0.1 cases per dog-month. Note the 300 dog-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 animal-time units. Similarly, we often say that animals 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 an animal becoming ill. Although incidence deals with 'new cases' of disease, it does not necessarily imply just the 'first case' within an animal. For some diseases (*eg* clinical mastitis in dairy cows), multiple cases are possible within an animal, either by involving different quarters of the udder or recurring in the same quarter after a period of absence from that quarter.

For reasons perhaps related to their unique susceptibility, or due to the effect of the first disease occurrence in the animal, animals that develop one case of a disease are often at a much 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 animal 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:

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

Incidence times are the times at which incident cases occur. They are usually measured as the elapsed time since a reference event (*eg* days after calving for diseases of dairy cows) 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 (*eg* country X has had 12 cases of bovine spongiform encephalopathy (BSE)). It might also be used for some common diseases (*eg* case counts of *Salmonella* in humans) but without data on the number of samples/animals 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 BSE 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 animal 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 cow having a case of clinical mastitis in the next year is very different (*ie* much higher) than the risk of having a case in the next week. In addition, only the first occurrence of a disease in the time period of interest is relevant because, once an animal 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 that a 7-year-old boxer will develop some form of detectable neoplasia over the next year is 14%. 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 animal-time during a given time period. It has units of 1/animal-time, and is positive without an upper bound. If a cattery housing 50 cats has 72 cases of upper respiratory disease over a period of a year, the incidence rate is 72/50, which is 1.44/cat-year (or 0.12/cat-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 (below) focuses on cases of disease. Risk can either be expressed at the animal level (*eg* the probability of an 8-year-old dog developing lymphosarcoma within the next year) or at the population level (eg the proportion of 8-year old dogs that develop lymphosarcoma within a year). Rothman and Greenland (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 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 animals 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* a herd of dairy cows followed for the next year) or in terms of some life event (*eg* all cows in a dairy herd followed for the first 2 months of lactation—regardless of when the lactation starts—to determine the risk of ketosis). Only disease-free animals in the population at the start of the study period are considered to be at risk and are monitored for the outcome of interest. Animals which 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 animals are leaving and entering the population throughout the study period. For example, if you wanted to determine the frequency of lymphosarcoma over a one-year period in a population of dogs served by a single veterinary clinic (assuming that all cases are diagnosed at the veterinary clinic), the population at risk would be an open population of dogs that were served by that clinic. 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, dairy and swine herds are inherently open in the sense that new animals enter the at-risk group (this use of open is not the same as saying that a farmer does or does not purchase new 'outside' animals). However, if we observe a set of animals, *eg* post-partum, for a full, defined risk period, then the population becomes closed.

4.5 CALCULATING INCIDENCE RATES

Incidence rates are calculated as:

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

An **animal-time unit** is one animal for a defined period of time (*eg* a cow-month, a dog-day (not to be confused with the 'dog days' in August)).

As noted above, incidence rates can be calculated using only the first occurrence of disease for any given animal (and from then on they are not considered to be at risk), or using all occurrences of disease. For example, a neoplastic disease would likely occur only once in an animal's lifetime but some infectious diseases such as mastitis can occur more than once in a dairy cow. However, even for diseases that might occur multiple times, we might only be interested in an animal's first case of mastitis as risk factors for a first case might be different from 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 animals achieve it if they live long enough).

As with calculating the number of animals at risk for R, there are several methods for calculating animal-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 when animals are at risk for multiple disease episodes, as opposed to only one disease episode per animal. The important thing to remember is that, if you are only interested in the first case of disease, then, after the animal contracts the disease of interest, it is **no longer** at risk and it no longer contributes to the pool of animal-time units at risk, even if it remains in the herd or study.

Exact calculation An exact calculation requires that the exact amount of animal-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 animal 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 animals).

If multiple cases of disease per animal 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 animals 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 animals were observed for exactly one month (30 days). The history for each individual was as follows:

1 animal not sick at all		1.00	animal-month at risk
1 animal sick on day 10		0.33	animal-months at risk
1 animal sick on day 20		0.67	animal-months at risk
1 animal sold on day 15		0.50	animal-months at risk
Total 'population at risk'	=	2.50	animal-months at risk
Total new cases of disease	=	2	
$I = 2/2.5$	=	0.80	cases/animal-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 t = 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 pigs over 6 weeks in a feeder barn 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 96 pig-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 that have 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 animals in the population at risk at the same point in time.

For example, if you bleed 75 horses from a large riding stable and test for equine infectious anemia (swamp fever) and 3 test results are positive, P is:

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

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 subclinical mastitis in a dairy herd is 0.3/cow-year (*ie* 30 new infections/100 cows per year) and the mean duration of an infection is 3 months (0.25

year), then we would expect P to be:

$$P = \frac{0.3 * 0.25}{0.3 * 0.25 + 1} = 0.07 = 7\%$$

so, on any given day throughout the year, we would expect 7% of cows to have subclinical mastitis. However, given that both the incidence rate and duration of infections probably vary substantially throughout lactation, more complex formulae that take this variability into account are probably required (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 animals become infected with a certain pathogen. For example, by bleeding a group of cats at regular intervals and testing for feline leukaemia virus, the rate at which cats 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 animals 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 animals that die from (or with) a specific disease during a defined time 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 cow regurgitates and contracts aspiration pneumonia and then dies, did it die:

- due to 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 decision 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 are really measures of risk.

Example 4.2 Calculation of risk and rate

You are interested in determining the frequency of new intramammary infections (IMI) with *Staph. aureus* in dairy cattle so you identify 5 cows in a dairy herd, follow them for one full lactation (10 months) and culture milk samples at months 0 (calving), 2, 4, 6, 8 and 10 (dry-off). The results are presented in the table below. A cow is only considered to have a new intramammary infection if it was negative on the preceding sample.

	Sampling times						Total months at risk	
Cow	0	2	4	6	8	10	First case only	All cases
A	0	X	0	0	X	X	2	6
B	0	0	0	–	–	–	4	4
C	X	0	0	X	X	X	0	4
D	0	0	0	0	0	0	10	10
E	0	0	X	0	X	X	4	6

where:

X = positive culture

X = positive culture that represents a new IMI

0 = negative culture

– = cow removed from herd

par = population at risk

a) risk of infection during first 2 months of lactation
par = 4 cows
new IMI = 1 cow
2-month R = 1/4 = 0.25

b) risk of first infection during lactation
par = 4 - 1/2 (1 withdrawal) = 3.5 cows
new IMI = 2 cows
lactation R = 2/3.5 = 0.57

c) rate of IMI—considering first cases only
par = 20 cow-months
new IMI = 2 first cases
I = 2/20 = 0.1 cases/cow-month
= 1 case/cow-lactation

d) rate of IMI—considering all new IMI
par = 30 cow-months
(*eg* cow A at risk for months 0 to 2 and 4 to 8)
new IMI = 5 cases
I = 5/30 = 0.17 cases/cow-month
= 1.7 cases/cow-lactation

e) lactation risk estimated from lactation rate (first cases only)
I = 1 case/cow-lactation
$R = 1\text{-}e^{-1} = 0.63$

f) prevalence at dry-off
par = 4 cows
existing IMI = 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 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* herd, 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 animal-to-animal spread and that due to common exposure (*eg* BSE in cattle). 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 animals with a specific disease that die from it (within a specified time period). It is actually a 'risk' measure (*ie* a proportion) instead of 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:

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

where Z_α is the (1-α/2) percentile of the standard normal distribution. In small samples, the *t*-distribution should be used in place of the *Z*-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 these cases, exact CIs based on probabilities derived from the binomial distribution (for proportions) or the Poisson distribution (for rates) will be more appropriate.

Example 4.3 shows the calculation of approximate and exact CIs for a prevalence proportion and exact CIs for some estimated incidence rates.

Example 4.3 Confidence intervals for proportion and rate
data = dairy_dis (herd 1)

Prevalence data for several infectious diseases were obtained from a sample of dairy herds. See Example 2.1 or Chapter 31 (dairy_dis) for a more complete description of these data.

Approximate and exact CIs for the prevalence proportion of leukosis and Johne's disease in herd 1 (27 cows) in this dataset were computed.

Disease type		Number of positives	P	SE	95% CI	
Leukosis	approximate	22	0.815	0.075	0.658	0.971
	exact				0.619	0.937
Johne's	approximate	3	0.111	0.060	-0.016	0.238
	exact				0.024	0.292

This shows that approximate CIs might go beyond the theoretically possible boundaries of 0 and 1.

Incidence rates were computed by assuming that:

- the age of each cow (in years) was her current lactation number plus 2.
- all infections arose immediately before the cow was tested (*ie* her period of risk was equal to her age). (This is a very untenable assumption for these 2 diseases and has been done only for the sake of this example.)

Exact CIs for the incidence of these 2 disease rates were then determined based on the Poisson distribution.

Disease	Number of positives	Cow-years at risk	I	SE	Exact 95% CI	
Leukosis	22	158	0.139	0.030	0.087	0.211
Johne's	3	158	0.019	0.011	0.004	0.056

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 host factor distribution (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 animal-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, for the rest of this discussion, we will primarily refer to rates, but the methods are equally applicable to risks.

Differences in disease rates (I) between populations of animals 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 animal-time in each of the strata in the study group. The process is called indirect standardisation. It is very 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:

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

Example 4.4 demonstrates the indirect standardisation of rates.

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 animals in each stratum instead of the proportion of animal-time. The expected number of cases, if the reference population risks apply to the study group's distribution of animals, 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.

4.11.4 Direct standardisation of rates

A second way of addressing the problem is 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 animals. Example 4.5 presents the calculation of direct standardised rates.

Example 4.4 Indirect standardisation of rates

Assume that you have data on the herd rate of tuberculosis (*ie* incidence rate of herds found to be positive) from 2 geographical regions which you would like to compare. However, the proportion of dairy and beef herds differ in the 2 regions and you know that this factor influences the rate of herd infections. You obtain a set of standard incidence rates based on data from the whole country and they are:

- the rate in beef herds is 0.025 cases/herd-year,
- the rate in dairy herds is 0.085 cases/herd-year, and
- the overall rate is 0.06 cases/herd-year.

In Region A, you have data from 1,000 herds over one year and in Region B, data on 2,000 herds for one year. The data are:

Type	Number of cases	Number of herd-years (T_j)	Observed rate (I_j)	Herd-years distribution (H_j)	Standard rate (Is_j)
Region A					
Beef	17	550	0.031	0.55	0.025
Dairy	41	450	0.091	0.45	0.085
Total	58	1000			
Overall rate*			0.058		0.052
SMR = 0.058/0.052 = 1.12					
Indirect standardised rate (I_{ind}) = 0.06 * 1.12 = 0.067					
Region B					
Beef	10	500	0.020	0.25	0.025
Dairy	120	1500	0.080	0.75	0.085
Total	130	2000			
Overall rate*			0.065		0.07
SMR = 0.065/0.07 = 0.93					
Indirect standardised rate = 0.06 * 0.93 = 0.056					

* Overall rate is the sum of the stratum-specific rates times the H_j distribution (eg overall observed rate in Region A=(0.031*0.55)+(0.091*0.45)=0.058 (except for slight rounding errors).

Although the stratum-specific rates in Region A are higher than in Region B, the crude overall rate would suggest (incorrectly) a lower rate in Region A (0.058 vs 0.065) whereas the standardised rates show (correctly) a higher rate in Region A (0.067 vs 0.056).

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 animals (Hs_j) in each category in the reference population is used instead of the proportion of animal-time (Ts_j) in each category. Example 4.5 demonstrates the direct standardisation of rates.

Example 4.5 Direct standardisation of rates

Using the same data presented in Example 4.4, and a suitable reference population which had a cattle type time-at-risk distribution (Ts_j) of:

- beef 40%
- dairy 60%.

Direct standardised rates can be computed as:

Cattle type	Observed rate (I_j)	Reference population distribution (Ts_j)	Product (I_j* Ts_j)
Region A			
Beef	0.031	0.4	0.012
Dairy	0.091	0.6	0.055
Direct standardised rate (I_{dir})			0.067
Region B			
Beef	0.02	0.4	0.008
Dairy	0.08	0.6	0.048
Direct standardised rate			0.056

Standardisation has once again revealed that the rate of tuberculosis is actually higher in Region A.

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.

One example stems from work in Ireland on tuberculosis. There, one measure of progress of the control programme is to monitor the annual risk (actually, prevalence) of lesions in supposedly tuberculosis-free cattle at slaughter. A number of factors affects the lesion risk. Two of the more important factors are slaughter plant (not all plants do an equally good job at finding lesions) and class of animal slaughtered (cows tend to have higher lesion prevalence than heifers, steers or bulls). Season also has an effect. One might think that, on an annual basis, season would cancel out but, if the slaughter distribution shifted seasonally, this would impact the lesion risk. Thus, with approximately 18 major slaughter plants, 4 classes of animal and 4 seasons, we would have 288 strata for each year. For each stratum, one needs the number slaughtered and the number of tuberculous lesions found (from which the stratum-specific risks can be computed). Then the number of cattle in that stratum is expressed as a proportion of the total slaughtered (*eg* using national data from a 10-year period as the standard population). We then have the H_j and an R_j for each stratum which are combined to compute a direct standardised annual risk. In this manner, the annual lesion risks could be compared without concern about the effects of season, animal class, or slaughter plant biasing them.

REFERENCES

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

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

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

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