

©1987 S. Wayne Martin, Alan H. Meek, Preben Willeberg

Veterinary Epidemiology: Principles and Methods

Part 1: Basic Principles

Chapter 3: Measurement of Disease Frequency and Production

Originally published 1987 by Iowa State University Press / Ames

Rights for this work have been reverted to the authors by the original publisher. The authors have chosen to license this work as follows:




License Information:

1. The collection is covered by the following Creative Commons License:



Attribution-NonCommercial-NoDerivs 4.0 International license

You are free to copy, distribute, and display this work under the following conditions:

	Attribution: You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work.) Specifically, you must state that the work was originally published in <i>Veterinary Epidemiology: Principles and Methods (1987)</i> , authored by S. Wayne Martin, Alan Meek, and Preben Willeberg.
	Noncommercial. You may not use this work for commercial purposes.
	No Derivative Works. You may not alter, transform, or build upon this work.

For any reuse or distribution, you must make clear to others the license terms of this work.

Any of these conditions can be waived if you get permission from the copyright holder.

Nothing in this license impairs or restricts the author's moral rights.

The above is a summary of the full license, which is available at the following URL:

<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>

2. The authors allow non-commercial distribution of translated and reformatted versions with attribution without additional permission.

Full text of this book is made available by Virginia Tech Libraries at: <http://hdl.handle.net/10919/72274>

Measurement of Disease Frequency and Production

3.1 Disease Frequency: General Considerations

Counts of individuals that are infected, diseased, or dead may be used to estimate workload, costs, or the size of facilities required to provide adequate health care for a specific animal population. However, epidemiologists usually wish to estimate the probability of events such as becoming infected, diseased, or dying in populations containing different numbers of individuals. Hence they express these counts as a fraction of the number of animals biologically capable of experiencing the event. The latter group of animals is called the population at risk. Fractions having the general form $a/(a + b)$ (where a is the number of animals with the event of interest, and b is the number of animals at risk of but not experiencing that event) are called either rates or proportions (Elandt-Johnson 1975). In practical terms rates are fractions, but they usually are multiplied by 100 or 1000, etc., so the result is a number greater than 1.

Morbidity and mortality are the two main categories of events for which rates are calculated. However, there are other events of interest to veterinarians and their clients, including culling (the premature removal of animals from a herd or flock), survival to weaning, and pregnancy rate (the probability of becoming pregnant within a specified period). The format for calculating these rates is the same as for morbidity and mortality; hence only the latter will be described in detail in this chapter.

3.1.1 Rates: Specifying the Denominator and Time Components

All rates have an external time component which refers to a period or a point in calendar time (called the study period). This should be specified

when reporting results because the rate may change with time, from season to season, or year to year. In addition, a rate is based on an internal time component (ITC), a time period having a duration of less than or equal to the study period. An investigation of the rate of calf mortality might last for a period of three years, but the calculation of the rate could be based on a daily, monthly, yearly, or 3-year basis.

A basic rule in forming a rate is that each animal can only experience the event of interest once during a time period; they cease to be at risk after the event of interest occurs, and for the duration of the internal time period on which the rate is based. Although mastitis can occur more than once during a lactation, if one is calculating the rate of mastitis during a lactation (ITC), only the first occurrence is counted. The easiest way to handle multiple occurrences is to shorten the ITC sufficiently to make the constraints reasonable. That is, several rates of mastitis, one for each 30-day interval postpartum, could be calculated.

In general, there are two different types of rates. The first, called a true rate (in technically precise terms, an incidence density rate), describes the average speed at which the event of interest (i.e., infection, disease occurrence, culling, death) occurs per unit of animal time at risk (Green 1982; Kleinbaum et al. 1982). In human medicine the most common time unit used for the period of risk is a year; however, shorter periods such as days or months are appropriate and often are used in veterinary medicine. The concept of animal time may require elaboration; for example, one animal year of risk may result from one animal being at risk of the event of interest for one year, or 12 animals being at risk for one month (1/12 of a year), or 365 animals being at risk for one day (1/365 of a year). Many other combinations are possible, but the general rule is to multiply the number of animals by their average period at risk to obtain the animal time of risk.

If the data are available, an exact denominator for a true rate is formed by adding each individual time period at risk for all animals in the study. Often, calculating an exact denominator is not practical or necessary. An approximate denominator may be formed by adding the number of animals at risk at the beginning of the time period to the number at risk at the end, dividing the sum by 2 to obtain the average number at risk (NAR), and multiplying the number at risk by the appropriate ITC.

Thus the general formula for a true rate is:

$$\frac{\text{no. animals acquiring event of interest}}{\text{average NAR} \times \text{ITC}}$$

EXAMPLE CALCULATIONS To illustrate this method of calculating a rate, assume that 3 animals were observed in a study period lasting 1 year.

During the year, 2 develop a disease, 1 at day 120 (0.33 years) and 1 at day 240 (0.67 years). The true rate of disease per animal year using the exact denominator is:

$$2/(1 + 0.33 + 0.67) = 2/2 = 1 \text{ per animal year}$$

The true rate using the approximate denominator is:

$$2/[(3 + 1)/2] \times 1 = 2/2 = 1 \text{ per animal year}$$

The two rates agree because the animals experiencing the event of interest were at risk for an average of exactly 1/2 year. Note that 2 animal years of risk were experienced by these 3 animals during the 1-year study period. Also, the time period on which the rates were based (the ITC) is 1 year, the same as the period of study (the external time component). The ITC of 1 year is represented by $\times 1$ in the above calculations. If the rate was desired on an animal week basis, the ITC factor $\times 52$ would be used.

True rates are used when the animal population being studied is very dynamic (with additions and/or withdrawals) during the period representing the ITC. As mentioned, the approximate denominator is used when the exact period of risk of individual animals is unavailable or impractical to obtain. True rates have a minimum value of zero and a maximum value of infinity; true rates apply only to populations and have no interpretation at the individual level. Had both animals developed disease on day 30 (0.08 of a year), the total animal years of risk would have been 1.16 and thus the rate would be 1.72 per animal year, or 172% (172 per 100 animal years). This cannot be sensibly interpreted at the individual animal level.

If a true rate has been calculated based on one internal time period, say x months, and it is desired to determine the rate on the basis of some other time period, say y months, then assuming a constant rate, the rate in the latter period is: true rate in $y = \text{true rate in } x(y/x)$.

EXAMPLE CALCULATIONS In the initial example, the true rate per animal month would be $1 \times 1/12 = 0.08$ per animal month.

The second type of rate, called a risk rate (in technically precise terms, a cumulative incidence rate), provides a direct estimate of the probability as defined in statistics of an animal experiencing the event of interest during the internal time period. (In this text, risk will be used as a synonym for probability and the specific measure of risk will be referred to as a risk rate. The words "at risk" may be used in their usual sense, namely, to denote animals susceptible to that disease.) This method requires that each animal initially at risk be observed for the full duration of the stated time period or until the event of interest occurs. Also, there can be no additions to the

number initially at risk. (These constraints are the major reasons that true rates often are used to describe the rapidity with which disease occurrence is changing in natural populations.) If there are withdrawals (losses from the study), for reasons other than the event of interest, the effective denominator is determined by subtracting one half of the number withdrawn from the initial number at risk. (The reason for subtracting one half rather than some other number is more pertinent in biometrics courses.) Risk rates have a minimum value of 0 and a maximum value of 1; risk rates may be interpreted at either the population or individual level.

The general format for a risk rate is:

$$\frac{\text{no. animals acquiring event of interest}}{\text{initial NAR} - \frac{1}{2} \text{ withdrawals}}$$

The risk (probability) form of rate is used whenever possible for analytic purposes (comparing rates statistically), since comparing true rates poses both practical and theoretical problems in terms of testing for statistical significance.

EXAMPLE CALCULATIONS To illustrate the method for determining a risk rate (the probability of an animal developing disease during a time period of one year) using data from the previous example, is: $2/3 = 0.67$.

The risk form of rate may be multiplied by 100 or 1000 to express it on a per 100 or 1000 animals basis. For example 67% means 67 events per 100 animals initially at risk.

If the risk form of rate has been calculated based on one internal time period (e.g., x months) and it is desired to express the risk rate for a different length of time (e.g., y months), then assuming a constant rate, the risk in the latter period is: risk rate in $y = 1 - (1 - \text{risk rate in } x)^{y/x}$.

EXAMPLE CALCULATIONS If the risk rate of disease in one year is 0.67, the risk rate in two years is: $1 - (1 - 0.67)^2 = 0.89$.

If a true rate is available and the risk of an animal experiencing the event of interest (in the same time period) is required, the formula to convert a true rate to a risk rate is: risk rate = $1 - e^{-\text{true rate}}$, where e is the base of natural logarithm. This approximation is extremely good when the true rate is below 0.05 per unit of animal time.

When rates are low (<15%), the technical differences between true rates and the risk form of rates may be ignored primarily because the difference in magnitude between them is of little practical importance. For example, in Table 3.1 the true rate of foot problems is 0.24 per cow year. Using the above formula, the risk rate per year is 0.21, for practical purposes, nearly the same magnitude. On the other hand, there is merit in

Table 3.1. Example calculations: true rates and risk rates

A herd of dairy cows provides the following data for the year 1983: On January 1, there were 60 cows in the herd, 6 of which had foot problems; 42 of the 60 cows calved during the year.

Ten new cows entered the herd during the year, all at the time of calving. Eight of the original cows were culled; 4 of these 8 had calved and subsequently developed left displaced abomasum (LDA) and foot problems (FP); the other 4 cows had no diseases and had not calved.

A total of 8 cows developed left displaced abomasum, 6 of these also developed foot problems. Six other cows acquired foot problems; 32 other cows experienced one or more other diseases.

Two cows died; 1 of these had left displaced abomasum, the other no disease.

What are the morbidity, mortality, culling (crude), and the proportional morbidity rates?

In order to proceed make the following assumptions: The period of risk for left displaced abomasum is short and only cows that calve are at risk; hence, use the initial population at risk—adjusting it for any losses—as the denominator. The period of risk for foot problems is long and cows are affected for their lifetime; hence, use the average population at risk for the denominator.

Morbidity risk rate (LDA) = $8 / [(42 + 10) - 0.5 \times 1 \text{ died}] = 8 / 51.5 = 0.16$ per year

Mortality risk rate (LDA) = $1 / [(42 + 10) - 0.5 \times (1 \text{ died} + 4 \text{ culls})]$
 $= 1 / 49.5 = 0.02$ per year

Case fatality rate (LDA) = $1 / 8 = 0.125$. Only deaths shortly after the disease occurrence are of interest, so the 4 culls are not counted as withdrawals.

Proportional morbidity rate (LDA) = $8 / (32 \text{ others} + 6\text{FP} + 8\text{LDA}) = 8 / 46 = 0.17$

Morbidity true rate (FP) = $12 / [(60 - 6) + (54 - 2 \text{ deaths} - 4 \text{ culls} - 12 \text{ cases} + 10 \text{ additions}) / 2] \times 1$
 $= 12 / [(54 + 46) / 2] \times 1 = 12 / 50 = 0.24$ per cow year

Crude mortality true rate = $2 / [(60 + 60) / 2] \times 1 = 2 / 60 = 0.03$ per cow year

Proportional morbidity rate (FP) = $12 / 46 = 0.26$

Culling true rate = $8 / [(60 + 60) / 2] \times 1 = 8 / 60 = 0.13$ per cow year

noting the differences to avoid confusion when the rates are $> 15\%$.

A practical method of calculating risk rates in dynamic populations circumventing the use of exponentials is:

$$\frac{\text{no. animals acquiring event of interest}}{\text{average NAR}}$$

This formula is very much like the true rate formula given earlier, but in calculating the average NAR the animals developing the event of interest are not subtracted from the NAR at the end of the stated time period. For example, in Table 3.1 the risk rate of foot problems may be calculated using the average of 54 and 58 ($46 + 12$ cases) as the denominator; namely, $12 / 56 = 0.21$.

3.2 Morbidity Rates

Morbidity rates describe the level of clinical disease in an animal population and may be crude, cause-specific, attribute-specific (i.e., host characteristic) or a combination of the latter two. Crude rates specify neither disease nor host attributes (e.g., the morbidity rate in feedlot cattle during July was 5%). Such rates may be made more meaningful by specifying the disease (e.g., the morbidity rate due to pneumonia in feedlot cattle during July was 4%) or attributes of the host (e.g., the morbidity rate in feedlot calves less than 8 months of age during July was 9%) or both. The extent to which one should make a rate specific depends on the circumstances involved. Morbidity rates also differ depending on whether new cases (incidence) or only existing cases (prevalence) are of interest. Although it is possible to include the number of new and existing cases in the same rate (called period prevalence), it is usually advisable to keep them separate.

Incidence rates describe the probability, or rapidity, of a new case developing during the stated internal time interval. The general formula for a crude true incidence rate is:

$$\frac{\text{no. animals developing disease during time period}}{\text{average population at risk during time period} \times \text{ITC}}$$

For example, in a study of calf morbidity the formula for the true morbidity rate per animal month would be:

$$\frac{\text{no. calves developing disease during a month}}{\text{no. calf-months at risk during that month}}$$

In most instances, the denominator would be calculated by counting the number of live disease-free calves on the first day of the month, adding this to the number of live disease-free calves on the last day of the month and dividing the sum by 2 (the implied time component being $\times 1$ month). Calves that developed disease during the month would not be at risk at the end of the month and hence should not be included even if they are alive and disease-free at that time. If detailed calf records were available, the exact denominator could be determined, but often such accuracy is not required.

To directly calculate the probability of disease occurrence in a group of animals (e.g., pigs born in July, cattle entering a feedlot in October, dogs whelping in May), one should use the risk form of incidence rate. For example, the formula for the risk rate of disease in calves born in July would be:

$$\frac{\text{no. calves born in July developing disease}}{\text{no. calves born alive in July}}$$

Note that the disease does not have to occur in July. Usually one specifies a reasonable period of risk for the disease in question, say 28 days for most neonatal diseases.

Host characteristics (attributes) often have a dramatic effect on the probability of disease events (see Chapter 4). Therefore, most rates are restricted to selected ages or breeds of the species in question; the restrictions apply to both the numerator and the denominator of the rate. An example of an attribute specific rate is a neonatal rate, indicating disease or death within 28 days of birth.

The risk form of rate is frequently used when the event(s) of interest is closely related, temporally, to occurrences such as farrowing (birth), entry to a feedlot, or the start of a racing season; the period of follow up begins at the time of the latter events. In these instances, the biologic period of risk usually is short relative to the average duration of observation (study period) of individual animals. For example, since the majority of cases of displaced abomasum (DA) occur within a few weeks of calving, the risk rate formula would be:

$$\frac{\text{no. cows developing DA of those calving in June}}{\text{no. cows calving in June}}$$

In calculating risk rates, the animals in the numerator must belong to the group defined in the denominator. Of course, if individuals cannot be identified readily, or if new animals are added to the at risk group, the true rate formula:

$$\frac{\text{no. cows developing DA in June}}{\text{no. cow months of risk in June}}$$

may be used. Both formulas require that the at risk period for DA be defined. One can convert from the true rate to the risk rate using the formula previously shown. Note that some cows developing DA may not have calved in June and may have contributed little to the denominator. Further, some of those calving in June might develop DA in July, but would not be counted in the numerator although they contributed to the denominator. However, in general and particularly in large, stable populations, these discrepancies cancel each other and the rate remains valid. (See Table 3.1 for illustrative calculations.)

For many infectious diseases, animals previously exposed or vaccinated may not be biologically at risk. Thus the rates can be made more accurate if adjustments are made to the denominator for the number of immune animals in the population, and this information should be used if the circumstances allow. Frequently, however, the number of truly immune

animals (as distinct from animals with high-serum titers) is unknown; thus if animals are apparently at risk of the event or disease of interest, they should be counted in the denominator.

In contrast to incidence (a dynamic measure of disease occurrence), the prevalence proportion (also called the point prevalence rate) is a static measure of disease frequency. It is the fraction of the population that is diseased at a point in time. The general formula for a crude prevalence proportion is:

$$\frac{\text{no. animals with disease at a point in time}}{\text{no. animals at risk at that point in time}}$$

Note that for a diseased animal to exist, the animal must first develop the disease (a function of incidence); then the disease must persist and the animal must survive (both a function of duration). Thus, in diseases of short duration or with a high case fatality rate, the incidence rate will likely be greater than the prevalence proportion. Chronic diseases tend to produce prevalence proportions that are greater than the incidence rates. In keeping with common usage, prevalence proportion will be referred to hereafter as prevalence. An approximation that explicitly links incidence rate (IR), prevalence (P) and duration of disease (D) is: $P = IR \times D$. All three quantities must be stated in the same time period (e.g., days).

The terms incidence and prevalence often are used incorrectly, particularly in the reporting of the results of mass serologic or microbiologic testing. By definition, incidence rates require two tests—one at the start of the period of observation to ensure that the animals did not have the disease, and the second to investigate whether the disease developed during the observation period. Rates based on one test or examination are by definition measuring prevalence (existing cases). Quite often, rates derived from clinical diagnostic data are treated as incidence rates, as if they were measuring the relative frequency of new cases. However, these rates most often are based on time of diagnosis, not on time of occurrence of the disease. For diseases that may remain subclinical for months or years before becoming clinically apparent, ignoring this difference could lead to inferential errors. For example, animals born with congenital abnormalities are often thought of as new cases and therefore as incidence cases. However, in order to exist at birth, the abnormality must develop in utero and the fetus must persist (not be resorbed or aborted at an early stage of development). Variation in the severity of the abnormality, with respect to survivability of the fetus, could drastically alter the number of animals with abnormalities observed at or after birth, with no change in the number of new abnormalities. Thus, congenital abnormalities measure prevalence not incidence.

As demonstrated above, it is quite important to differentiate incidence

rates from prevalence proportions. First, their magnitude may differ greatly, particularly with chronic diseases. Second, factors associated with acquiring new disease may differ from those associated with having a disease, and only the former are of value for disease prevention. Finally, knowledge of the time period when the disease was acquired assists in demarcating the time period during which causal factors may have operated and, hence, assists in the identification of these factors.

A subtype of an incidence rate is an attack rate. The latter is used when the period of risk is limited, as in simultaneous exposure of a group of animals to noxious gases or contaminated water or food. The general formula for an attack rate (AR) is similar to that for the risk form of rate, namely:

$$\frac{\text{total no. animals that develop disease during specified time period following exposure}}{\text{total no. animals exposed}}$$

Because the biologic period of risk is limited, an attack rate represents the total incidence rate; no new cases would arise from that exposure even if the period of observation were lengthened.

A further modification of morbidity rates, primarily used to study the spread of infectious diseases in defined subgroups (e.g., households) of the population, is the secondary attack rate (SAR), which is calculated as:

$$\frac{\text{total no. animals exposed to first case (proband) that develop disease within range of incubation period}}{\text{total no. animals exposed to proband}}$$

Secondary attack rates are usually applied to natural groupings of animals such as pens or farms. They may also be used to evaluate the communicability of diseases of unknown etiology in an attempt to see if infectious agents might be involved. For infectious diseases, the higher the SAR the more contagious the agent. However, some noninfectious diseases can occur in a manner that may result in a high secondary attack rate. This may occur if there is a variable latent period following a common exposure of individuals within the group, and hence the disease may appear to spread from animal to animal.

3.3 Mortality Rates

Mortality rates describe the quantitative impact of death in an animal population. Two frequently used measures of mortality are the crude and

cause-specific mortality rates. The formula for the crude mortality (true) rate is:

$$\frac{\text{total deaths in time period}}{\text{average population at risk in time period} \times \text{ITC}}$$

and the formula for the cause-specific mortality (true) rate is:

$$\frac{\text{total deaths from disease } X \text{ in time period}}{\text{average population at risk in time period} \times \text{ITC}}$$

The probability (i.e., risk) of dying in a specified time period may be determined by restricting the denominator to those alive at the start of the time period and adjusting this number for any withdrawals, as was described for risk rates. All animals must be observed for the full time period, or until death or withdrawal occurs.

The risk of death in animals with a specific disease may be described using the case fatality rate. The formula for a case fatality rate is:

$$\frac{\text{total deaths from disease } X \text{ within specified time after diagnosis}}{\text{total no. animals acquiring disease } X}$$

Case fatality rates are of greater value in acute than in chronic diseases and are used to describe the virulence of the agent and/or the severity of the disease. (See Table 3.1 for example calculations.)

An approximation that links case fatality rates (CFR), cause-specific mortality rates (CSMTR), and incidence rates (IR) is $\text{CFR} = \text{CSMTR}/\text{IR}$. Thus under certain assumptions, if any two of these rates are known, the third may be calculated.

3.4 Proportional Rates

Sometimes, (e.g., when summarizing disease occurrence on one farm or in one clinic) an investigator divides the number of animals with a given disease by the total number of diseased animals. In other instances, the number of animals dying from a given disease is divided by the total number of deaths. These are called proportional morbidity or proportional mortality rates respectively. Although they have the form of a rate and often are mistakenly referred to as incidence or prevalence rates, the denominator is only a portion of the actual population at risk. Proportional

rates may be affected by independent changes in the numerator, the denominator, or both. Hence proportional rates are potentially misleading, and their use is discouraged in favor of the morbidity or mortality rates described previously.

3.5 Variability of Rates

Risk rates and prevalence proportions are averages subject to variability from sampling error. In calculating this sampling error, the number of animals used to calculate the rate is regarded as if it was a random sample from a larger population. If repeated samples of the same number of individuals n were selected, the calculated rate \bar{p} would vary from sample to sample. The extent of this variability is described by the standard error of the mean and is estimated from the sample to be:

$$SE(\bar{p}) = [\bar{p}(1 - \bar{p})/n]^{1/2}$$

A 95% confidence interval may be constructed using the upper and lower limits of the interval defined by $\bar{p} \pm 1.96 \times SE(\bar{p})$ (see Table 2.1). The interpretation to be placed on the confidence interval is that if many samples were selected and a confidence interval constructed for each, 95% would contain the true population rate. This approximation is quite good provided both $n\bar{p}$ and $n(1-\bar{p})$ are > 5 .

EXAMPLE CALCULATIONS Suppose that in a pen of 100 pigs, 30 develop pneumonia and 5 of these die during the first month on feed. If all pigs were free of pneumonia at the start of the feeding period, the true rate of pneumonia per month is $30/[(100 + 70)/2] = 0.35$ or 35% (i.e., 35 per 100 pig months). The probability of a pig developing pneumonia during the 1-month period (risk rate) is: $30/100 = 0.3$ or 30%.

If the above risk (0.3) remains constant during a 3-month feeding period, the probability of a pig developing pneumonia at least once during the 3-month period is:

$$\begin{aligned} \text{risk rate (3)} &= 1 - [1 - \text{risk rate}(1)]^{3/1} \\ &= 1 - (1 - 0.3)^3 = 1 - 0.7^3 = 0.66 \end{aligned}$$

This means that 66% of the pigs (or $100 \times 0.66 = 66$ pigs) would be expected to develop pneumonia in the 3-month period.

The true rate of mortality is $5/[(100 + 95)/2] = 0.051$ per month, whereas the probability of a pig dying during the first month (risk rate) is $5/100 = 0.050$. (Note that as the true rate decreases, it approximates the

risk rate very closely.) If the probability of mortality remained constant for the 3-month feeding period, the probability of a pig dying in the 3-month period is $1 - (1 - 0.05)^3 = 1 - 0.95^3 = 0.14$. This means that 14% of the pigs would be expected to die during the 3-month feeding period.

The probability of a pig dying if it develops pneumonia is found by using the case fatality rate. In this example, the case fatality rate for pneumonia is $5/30 = 0.17$ or 17%. (Note that since the only disease present is pneumonia, the above morbidity and mortality rates are cause-specific.)

If the 100 pigs were viewed as a sample of the feeder-pig population on this farm, one could construct confidence intervals for the average morbidity and mortality risk rates.

For the average morbidity risk rate, the standard error of \bar{p} (0.30) is $SE(\bar{p}) = (0.30 \times 0.70/100)^{1/2} = 0.046$ and hence the 95% confidence limits are 0.21–0.39 (21%–39%).

For the average mortality risk rate, the standard error of \bar{p} (0.05) is $SE(\bar{p}) = (0.05 \times 0.95/100)^{1/2} = 0.022$ and hence the 95% confidence limits are 0.007–0.093 (0.7%–9.3%).

If the 100 animals were obtained by formally sampling a defined population (a herd) with individual pigs being the sampling unit, and if the number studied was greater than 10% of the population, more precise estimates of the standard error may be obtained by adjustment using the finite population correction factor (see Table 2.1). Hence, if there were only 500 pigs in the population, $n/N = 0.2$ and the correction factor for the standard error is $(1 - 0.2)^{1/2} = 0.89$. Thus the best estimate of the standard error of the morbidity rate is $0.046 \times 0.89 = 0.04$, and the best estimate of the standard error of the mortality rate is $0.022 \times 0.89 = 0.02$. The resulting confidence intervals will be slightly narrower; a reflection that 20% of the population was sampled. The reader will now be aware that it is quite difficult to establish standard errors for true rates, hence no discussion of this topic will be presented. If standard errors are desired and the true rate is low (<10%), one may use the same approach as demonstrated above for risk rates.

3.6 Measuring Production: Basic Statistics

As previously mentioned, the level of production is often used in veterinary medicine as a proxy or surrogate measure for health. As such, production is frequently the outcome of concern (dependent variable) in many veterinary epidemiologic studies. Production, whether it be kilograms of milk per lactation, number of pigs per litter, number of litters per year, weight gain per day, or eggs per bird per year, is considered to be a quantitative variable. The sample distribution of a quantitative variable is best

described by the mean (\bar{y}) or average, the standard deviation (s), or variance (s^2), and the median. The sampling variability of the mean is described by the standard error of the mean [$SE(\bar{y})$] (see Table 2.1).

The mean is a measure of central tendency and a formula for calculating it is $\bar{y} = \Sigma y_i/n$ where y_i is the i th observation, n is the number of observations, and Σ means take the sum of the y_i . The median is another measure of central tendency and is the middle value when the n values are placed in order of magnitude. If n is even, the median is the average of the middle two values of y_i . The median is useful to describe central tendency when the distribution of a variable is not Gaussian (i.e., not bell-shaped or normal), since the median is affected less by extreme values than is the mean. If a distribution has a right skew (long tail to the right) the mean will be greater than the median and vice versa if the distribution has a left skew. Another way of treating skewed data is to transform them (e.g., by taking logarithms of the values) and then taking the mean of the logarithmic values. A common example of this approach is in the description of somatic cell counts in milk.

The standard deviation s is the square root of the variance or mean square s^2 and describes the variability of individual values of y around their mean. Two formulas for calculating s^2 are:

$$s^2 = \Sigma (y_i - \bar{y})^2 / (n - 1) \quad \text{or} \quad [\Sigma (y_i^2) - (\Sigma y_i)^2 / n] / (n - 1)$$

$$s = (s^2)^{1/2}$$

A number of relatively inexpensive calculators are programmed to calculate \bar{y} and s^2 ; nonetheless, the above formulas are instructive about the meaning of these statistics.

The n animals on which y and s are based may be viewed as a sample of size n from a much larger population. Repeated samples of the same size would provide other estimates of the average in the population. (One does not actually draw repeated samples but uses the central limit theorem to describe the variability of the sample mean.) The variability among these means is described by the standard error of the mean and this may be calculated as $SE(\bar{y}) = (s^2/n)^{1/2} = s/(n)^{1/2}$.

The standard error may be used to construct a confidence interval for the mean. The upper and lower limits of a 95% confidence interval are calculated using $\bar{y} \pm 1.96 \times SE(\bar{y})$.

When measuring rates of events (e.g., disease) at any aggregate level (e.g., farm level), the rates may be treated as quantitative variables for purposes of description and analysis.

3.6.1 Choosing the Production Parameter

In many studies only a few production parameters are available. However, even if the number of choices is limited, the investigator should try to select parameters that not only measure production, but that may be used as economic indicators, and hence are of value for decision making (Williamson 1980). For example, selecting the number of services per conception as a parameter of reproductive efficiency in dairy health management would probably be unwise; first, many factors including time of first postpartum breeding affect it, and second, this parameter is not a good indicator of important economic aspects of reproduction. The open interval (i.e., the period between parturition and conception) or the percentage pregnant by 100 days postpartum would be more appropriate parameters. A hierarchy of parameters should be used to monitor and/or investigate production decreases in health management programs (see 12.2). Both the mean and the standard deviation are important to note in such instances.

Choosing a suitable measure for milk production in dairy cows will serve as an example of some other considerations that must be taken into account in selecting a parameter. Absolute measures of milk production include the total kilograms of milk produced in a lactation (kg tot) and the kilograms of milk produced in a 305 day period (kg 305). The value of the kg 305 over the kg tot is that differences due to variation in days-in-milk are removed. However, other factors such as the age of cow and the season of calving can also have a major effect on the kg 305 produced. To circumvent these problems, the effects of age and calving season can be removed using an index known as the breed class average for milk production (BCM). In a simple sense, the production of a typical cow is assigned a value of 100 and all other cows are assigned a breed class average score based on their kg 305 adjusted for their age and their season of calving. In general, each BCM unit in a two-year-old cow represents about 45 kg of milk. The BCM allows one to compare the milk production between two groups of cows in the same herd or between two groups of herds without having to worry about the age structure or seasonal distribution of calvings within the groups.

If an investigator wished to compare the milk production of cows with a particular disease to that of cows without that disease, and production data from more than one herd were to be used, the comparison could be biased by differences in the level of production among herds, unless equal numbers of cows with and without the disease were selected from each herd. Another way to obviate this problem is to express each cow's level of production as the deviation (in BCM units) from the average production in her source herd. This parameter is known as the deviation-from-herd-average and is frequently used to remove the herd effect when making cow-level comparisons across many herds.

Although the above example is based on the dairy industry, similar indices for other parameters in other industries are available or can be derived.

3.7 Detecting Subclinical Disease with Screening Tests

The previous sections have been concerned with measuring the frequency and impact of visible events such as clinical disease or death in animal populations. Screening is the application of a test to apparently healthy animals in order to detect infection or subclinical disease. In domestic animals, probably the major economic loss is due to the effects of hidden or subclinical disease. For example, subclinical mastitis is a mild inapparent condition, yet because of its high prevalence, it has a much greater impact on the productivity of dairy herds than the sporadic yet dramatic clinical forms of the disease. In addition, knowledge of the frequency and distribution of infectious and noninfectious agents of disease and of immune responses to these agents can greatly assist our understanding of disease processes and the importance of various agents in manifestationally classified syndromes such as pneumonia or gastroenteritis. Certainly, as mentioned in Chapter 2, the frequency, distribution, and importance of subclinical disease may be very different from that of clinical cases. From an epidemiologic perspective, it may be argued that greater success at preventing disease occurrence can be realized if investigations are concentrated on how infections occur and persist in the absence of disease, rather than using only diseased animals as models of study.

Because the disease process is clinically inapparent, special tests (e.g., the California mastitis test) are required to detect subclinical disease. Also, in addition to what one might consider conventional laboratory tests, epidemiologists include any device or process designed to detect or elicit a sign, substance, tissue change, or response as a test. Thus, examples of tests include common serologic and microbiologic tests for detecting agents or the animal's response to an agent; clinical-pathologic tests designed to measure the number of particular cell types, the levels of tissue enzymes or minerals; as well as questions in personal or mail surveys. Using one or more of our senses during the diagnostic process for the detection of signs or tissue changes (including pregnancy diagnosis and meat inspection findings) could also be included as tests.

Tests are usually considered to be either pathognomonic or surrogate. Pathognomonic tests are those for which the detection of a sign, substance, response, or tissue change is an absolute predictor of the presence of the disease or disease agent. Surrogate tests detect secondary changes, which it is hoped will predict the presence or absence of disease or the disease agent. For example, a positive culture of *Brucella abortus* from a cow's milk

sample is pathognomonic for brucella infection. Testing the milk for antibodies to *Brucella abortus*, however, is a surrogate test; since it is not measuring the presence of *Brucella abortus* per se, but rather the body's reaction to brucella organisms or cross-reacting antigens. Surrogate tests may produce false-positive results, whereas pathognomonic tests do not. Both types of tests can have false-negative results. Such false results and the question of assessing tests and interpreting the results lead to the subject of sensitivity and specificity (Robertson 1963; Martin 1977; Dodd 1978; Seiler 1979; Martin 1984).

3.7.1 Sensitivity and Specificity

Suppose it is possible to correctly classify animals into two categories—those having disease X and those not having disease X ,—using a set of available tests. A new test has been developed, and its ability to differentiate between diseased and nondiseased animals needs to be evaluated. (Disease here is used in its broadest sense and includes subclinical disease and/or infection.)

The initial step in the evaluation is to select a sample of animals known to have disease X and a sample known not to have disease X . Although infrequently used in practice, formal random samples of each of these populations will help to ensure that animals to be tested are representative of diseased and nondiseased animals respectively, as this is crucial for accurate evaluation of the new test (Ransohoff and Feinstein 1978). It is also important that the new test is biologically independent of the methods initially used to define the true health status of the animals. After appropriate animals are selected, they are tested and classified as being positive or negative on the basis of the new test results. The resultant cross classification of n animals according to their true health status and the results of the screening test may be displayed as follows:

Test result	Actual health status (Disease X)	
	Present ($D+$)	Absent ($D-$)
Positive ($T+$)	a	b
Negative ($T-$)	c	d
	$a+c$	$b+d$

The sensitivity of the test is its ability to detect diseased animals and is defined as the proportion of the diseased animals that test positive, i.e., $a/(a+c)$. The specificity of the test is its ability to detect nondiseased animals and is defined as the proportion of nondiseased animals that test negative, i.e., $d/(b+d)$. (Nondiseased indicates animals that do not have the event of interest; it does not mean 100% healthy.) In combination these

two statistics describe how well a test can discriminate between nondiseased and diseased individuals. Note that the epidemiologic usage of "sensitivity" differs from immunologic or pharmacologic usage. In the latter disciplines, a sensitive test is one that detects a small amount of antibody, toxin, enzyme, etc. An immunologically sensitive test may not be epidemiologically sensitive, so one should be careful not to confuse the different meanings. Sensitivity and specificity are calculated in the same manner as risk rates because they are probability statements. To summarize:

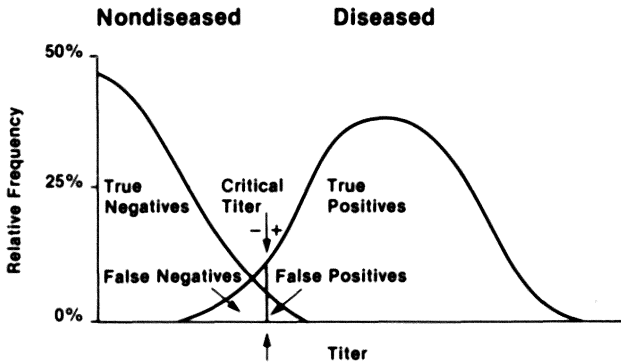
$$\begin{aligned} \text{sensitivity} &= a/(a+c) = p(T+/D+) \\ \text{specificity} &= d/(b+d) = p(T-/D-) \end{aligned}$$

In a random sample of the overall population, the true prevalence proportion of disease in the population $P(D+)$ would be estimated by $p(D+)$, i.e., $(a + c)/n$. However, in practice this parameter is almost always unknown; only the test results ($T+$ and $T-$) are available, and hence the estimate of $P(D+)$ is the apparent prevalence proportion $p(T+)$, namely, $(a + b)/n$. Obviously, the true and apparent prevalence proportions are equal only if $b = c$. In general, b tends to be numerically greater than c and thus the apparent prevalence is usually somewhat higher than the actual prevalence, sometimes by a surprising amount.

To summarize, in a random sample of the population,

$$\begin{aligned} \text{apparent prevalence} &= (a + b)/n = p(T+) \\ \text{true prevalence} &= (a + c)/n = p(D+) \end{aligned}$$

Note that for most surrogate tests there is an inverse relationship between sensitivity and specificity. That is, if the critical value of the test is altered so that the sensitivity is increased, the specificity will be automatically decreased. This is because the substances being measured may be present in nondiseased as well as diseased animals, although at different levels and with different frequencies, and often their distributions overlap. For example, Figure 3.1 displays the distribution of antibody titers to agent X in a sample of healthy nondiseased (do not have agent X) and a sample of diseased (have agent X) animals. Note that most nondiseased animals do not have a titer to the agent, some have low titers and a very few have high titers. On the other hand, in diseased animals the distribution is somewhat bell-shaped (i.e., a normal or Gaussian distribution). Very few diseased animals have low titers; most have moderate titers, and some have very high titers to the agent. Although the diseased animals have higher titers on average, the two distributions of titers overlap, and this produces an inverse relationship between the sensitivity and the specificity of tests measur-



3.1. Distribution of titers to agent X in sample of nondiseased and diseased animals.

ing this antibody response. The resultant sensitivity and specificity will depend on the critical titer selected.

In practice, a critical titer is selected so that animals having titers above that point are considered positive, and those having titers equal to or below that point are considered negative. In terms of the previous 2 x 2 table, diseased animals with titers above the critical titer are the true positives, their number being represented by a ; the nondiseased animals with titers below the critical titer are the true negatives, their number being represented by d ; the nondiseased animals with titers above the critical titer are false positives, their number being represented by b , and the diseased animals with titers equal to or less than the critical titer are false negatives, their number being represented by c .

If the critical titer is adjusted to increase the sensitivity (i.e., lowered or moved to the left in Fig. 3.1), the number of false-positive animals will increase, hence this decreases the specificity. If the critical titer is altered by moving it to the right to increase the specificity, the sensitivity of the test will decrease, thus there will be a larger number of false negatives. An example of the effect of changing the critical titer when testing for visceral larva migrans using an ELISA test is shown in Table 3.2 (Glickman et al. 1978).

In general, sensitivity and specificity describe the discriminatory power of a test based on a single biologic sample taken at a point in time. They do not describe how well the test would function if applied very late in the disease process as compared to early in the disease process; nor do they

Table 3.2. Sensitivity, specificity, and predictive value of the enzyme-linked immunosorbent assay (ELISA) for the diagnosis of visceral larva migrans

Cut-off log titer of a positive ELISA test	Sensitivity (%)	Specificity (%)	Predictive value	
			Positive (%)	Negative (%)
1	91.3	76.9	70.0	93.8
2	91.3	79.5	72.4	93.9
3	82.6	82.1	73.1	89.2
4	82.6	84.6	76.0	88.9
5	78.3	92.3	85.7	87.8
6	65.2	94.9	88.2	82.2
7	56.5	97.4	92.9	79.2
8	43.5	97.4	90.9	74.5
9	30.4	100.0	100.0	70.9
10	30.4	100.0	100.0	70.9
11	21.7	100.0	100.0	68.4
> 12	17.4	100.0	100.0	67.2

Source: Glickman et al. 1978, with permission.

describe how well one could classify the health status of animals based on results from using the test sequentially on the same animals. The same principles apply, however, to the situation where acute and chronic (convalescent) titers are measured, and an animal is declared infected or diseased if there is say a two-fold or four-fold titer rise. Here the question of interest is the ability (i.e., sensitivity and specificity) of a specified increase in titer to discriminate between diseased and nondiseased animals.

3.7.2 Indirect Estimates of Sensitivity and Specificity

Sometimes the test to be evaluated is biologically similar to those available to diagnose the disease, yet estimates of sensitivity and specificity are desired. This is frequently the case with diseases of viral etiology where the virus is difficult to culture, and secondary binding tests are used to detect the presence of antibody to viral antigens. In this instance, the results of the new test can be compared with the results of a bank of standard tests. For this purpose, animals positive to all tests in the bank are assumed to be diseased, and animals negative to all tests in the bank are considered disease-free. Animals with intermediate types of response are excluded from further analyses. The sensitivity and specificity calculations proceed in the usual manner, but the results of the comparisons should be prefixed with "relative" to indicate that the determinations are based on biologically related tests. Usually the results obtained by this method represent maximum values of sensitivity and specificity. The reader should note that comparing the results of one test to the results of a biologically related surrogate test does not allow the establishment of sensitivity or specificity. This procedure can establish which test gives more positive results and the extent of agree-

ment between the tests, but not their ability to differentiate diseased from nondiseased animals.

In other situations, it may prove very difficult to assemble a sufficiently large representative group of nondiseased animals in order to determine the specificity of a test. However, if test results on a relatively large number of representative animals ($n > 1000$) are available, and if it is reasonable to assume that the prevalence of disease is less than 1% and that the test has high sensitivity, an approximation may be used. The approximation is based on the assumptions that all test-positive individuals are false-positives and that disease is rare. Thus specificity can be estimated by $1 - (\text{number of test positives})/n = 1 - (a + b)/n$.

EXAMPLE CALCULATIONS If 17 of 2000 representative animals have positive tests, then assuming all are false-positive reactions the minimum specificity would be $1 - 17/2000 = 0.9915$ or 99.15%.

Under some circumstances, it may be possible to conduct a detailed follow-up on the test-positive animals and classify them into diseased and nondiseased (false-positive) groups. In this case, assuming a reasonable sensitivity, specificity may be more accurately estimated by:

$$1 - \frac{\text{number of false positives}}{n - \text{number of diseased among test positives}} = 1 - b/(n - a)$$

Finally, if estimates of sensitivity are available, the above estimate may be improved by using a^* instead of a , where $a^* = a/\text{sensitivity}$ (a^* estimates $a + c$).

EXAMPLE CALCULATIONS If 12 of the above 17 reactors were found to be diseased, the minimum specificity would be $1 - 5/(2000 - 12) = 1 - 0.0025 = 0.9975$ or 99.75%. In addition, if the test was known to be 80% sensitive, an improved estimate of specificity would be $1 - 5/(2000 - 15)$, which to four decimals in this case is also 0.9975 or 99.75%.

3.7.3 Predictive Value of Screening Test Results

The predictive value of a positive test is defined as the proportion of diseased animals among those that test positive; that is, the quantity $p(D+/T+)$ which is calculated using $a/(a + b)$. (Unless otherwise stated, this discussion will be restricted to the predictive value of a positive test result.) Caution is required here because this quantity sounds and looks like $p(T+/D+)$ (i.e., sensitivity), but it is quite different. Predictive value is important because it reflects the way test results are used in the field. Here the question is, Given that an animal has a positive test, what is the likelihood that the animal has the disease or infection under study? This ques-

tion arises because the true state of health is unknown, hence, the practitioner must argue backward from test results to the likelihood of disease, not from disease status to the likelihood of a specific test result.

The predictive value of a test has been used as a method of test selection. However, the predictive value of any given test is affected by both the sensitivity and the specificity of the test, as well as by the true prevalence of disease in the population. Since the latter usually is unknown, it makes the selection of the "best" test difficult, because the direction of the inequality of predictive values of two tests can be reversed depending on the prevalence of disease. One cannot assume that the test with the highest predictive value is necessarily the most sensitive or specific.

The data in Table 3.3 demonstrate the effect of prevalence of disease on the predictive value of the test result. Note that when the prevalence of disease is 3%, the predictive value of the test is 79.5%. (This is found by dividing 234, the number of test positives, into 186, the number of true positives.) When the prevalence of disease is 0.1% (i.e., one animal per thousand is diseased) the predictive value is 10.7%, and when the prevalence of disease is 0.01% (i.e., one animal per ten thousand) the predictive value of a positive test is 1.2%. Note that the assumed level of sensitivity and specificity, 62% and 99.5% respectively, have not changed except for rounding to obtain whole numbers (animals). The example in Table 3.2,

Table 3.3. Relationship between true prevalence of disease and the predictive value of a positive test result

Sensitivity = $p(T+/D+) = 62\%$ Specificity = $p(T-/D-) = 99.5\%$			
Example 1: $p(D+) = 3\%$			
	D+	D-	Total
T+	186	48	234
T-	114	9652	9766
	300	9700	10,000
Predictive value = $p(D+/T+) = (186/234) \times 100 = 79.5\%$			
Example 2: $p(D+) = 0.1\%$			
	D+	D-	Total
T+	6	50	56
T-	4	9940	9944
	10	9990	10,000
Predictive value = $p(D+/T+) = 10.7\%$			
Example 3: $p(D+) = 0.01\%$			
	D+	D-	Total
T+	6	500	506
T-	4	99,490	99,494
	10	99,990	100,000
Predictive value = $p(D+/T+) = 1.18\%$			

which is based on testing for visceral larva migrans, illustrates the relationship between predictive value, sensitivity, and specificity; the prevalence of disease being constant. The predictive value of this test is quite good (being at least 70%). This is only true because of the high prevalence proportion of visceral larva migrans of 37%.

The predictive value of a positive test result in a variety of circumstances can be estimated using the formula:

$$p(D+/T+) = \frac{p(D+) \times p(T+/D+)}{p(D+) \times p(T+/D+) + p(D-) \times p(T+/D-)}$$

Although valuable from a theoretical viewpoint, since it explicitly describes the factors influencing predictive value, the true prevalence of disease is rarely known, and hence this formula is not often used in practice. Its major value is to demonstrate what the predictive value would be if the test was used at a specified prevalence proportion.

Since the prevalence proportion of disease is usually below 0.2, the lack of specificity in most screening tests is responsible for the apparent prevalence of disease often being somewhat higher than the true prevalence of disease. This may be verified by comparing the apparent and true prevalence of disease for the data presented in Table 3.3. In general, the apparent prevalence is frequently not a good estimate of the true prevalence because of the false-negative and false-positive animals. However, if the sensitivity and specificity are known, the true prevalence may be estimated by:

$$p(D+) = \frac{p(T+) - p(T+/D-)}{1 - [p(T+/D-) + p(T-/D+)]}$$

Note that $p(T+/D-) = 1 - \text{specificity}$, and $p(T-/D+) = 1 - \text{sensitivity}$. For example, using the data in Table 3.3 example 2:

$$p(D+) = \frac{0.0056 - 0.005}{1 - (0.005 + 0.38)} = \frac{0.0006}{1 - 0.385} = 0.001 = 0.1\%$$

3.7.4 Methods for Improving Predictive Value

One method of improving the predictive value of a screening test is to screen only high risk populations; that is, populations likely to have a high rate of infection or disease. Observational studies (e.g., cross-sectional, cohort, and case-control) are used to identify subgroups with an elevated risk of infection or disease, and the screening program can then be concentrated on those individuals with a high risk, hence ensuring a relatively good predictive value.

A second method of improving the predictive value is to use more than one screening test. This may be done in several ways. The first example assumes that a relatively sensitive, inexpensive screening test is available for use on all animals in the population, and a more sensitive but expensive test is available for use on a limited number of individuals. Table 3.4 contains the expected results given that the initial test (with a sensitivity of 95% and specificity of 99%) is used on all individuals in the population, and the second test (with a sensitivity of 98% and specificity of 99%) is subsequently applied to the animals positive to the first test.

Table 3.4. Results expected after application of one test to all animals and a second test to all reactors from the primary test

Results of initial test:			Initial test:
	<i>D+</i>	<i>D-</i>	Total
<i>T+</i>	95	99	194
<i>T-</i>	5	9801	9806
	100	9900	10,000
Results of second test:			Second test:
	<i>D+</i>	<i>D-</i>	Total
<i>T+</i>	93	1	94
<i>T-</i>	2	98	100
	95	99	194

Note: Overall sensitivity = $(100 - 7)/100 = 93.0\%$

Overall specificity = $(9801 + 98)/9900 = 99.9\%$

Overall predictive value = $93/94 = 98.9\%$

The overall results of using these two tests is a combined sensitivity of 93% and a specificity of 99.9%. Notice that 5 diseased animals were missed on the first test and in order to reduce the number of false positives from the first test, an additional 2 infected animals were declared negative on the second test. However, the use of the second test reduced the number of false positives from 99 to 1. This demonstrates the general results to be expected utilizing tests in this manner. The actual results probably would not be this good, because if the two tests were biologically similar the results would be correlated; that is, they would tend to give similar results on samples from the same animal.

Another method of using multiple tests is to apply two or more tests simultaneously to all individuals. When tests are used in this manner, the resultant sensitivity and specificity are dependent on the way the results are interpreted. One method of interpretation used when a high sensitivity is required is known as parallel interpretation. Using parallel interpretation, an animal is considered positive if it reacts positively to one or the other or

both tests. This increases the sensitivity but tends to decrease the specificity of the combined tests. This makes intuitive sense since it gives a diseased animal the greatest opportunity to react positively. The second method of interpretation used whenever a high specificity is required is known as series interpretation. In series interpretation, an animal must be positive to both of the tests to be considered positive. As mentioned, this will increase specificity but decrease sensitivity because the likelihood of a diseased animal reacting positive to both tests is less than the likelihood of it reacting positive to both, or positive to the first and negative to the second or vice versa.

The outcome from using series and parallel interpretation with two tests is shown in Table 3.5. The sensitivity of the first test is 50% and its specificity 98.7%. The sensitivity of the second test is 60% and its specificity 98.6%. When the tests are interpreted in parallel, 150 of the 200 diseased animals are considered positive for a resultant sensitivity of 75%. A total of 7620 of the nondiseased animals are considered negative and thus the specificity is 97.7%. When the results are interpreted in series, only 70 of the 200 diseased animals are considered positive for a resultant sensitivity of 35%. However, 7770 animals that are not diseased are considered to be negative for a specificity of 99.6%.

Obviously the above example could be expanded to include more than two tests and, again, the results would be similar to that indicated here—parallel interpretation increases sensitivity and series interpretation increases specificity. In general, the greater number of tests involved, the greater the increase in sensitivity or specificity depending upon the method of interpretation. To identify the optimal classification (i.e., minimizing the overall misclassification rates) requires the use of more elaborate techniques such as discriminant analysis; however, these are beyond the scope of this book.

Table 3.5. Sensitivity and specificity of combined screening tests, with test results interpreted in series and in parallel

Test 1	Test 2	Diseased	Not diseased
+	-	30	70
-	+	50	80
+	+	70	30
-	-	50	7620
		200	7800
		Sensitivity	Specificity
Both tests in "parallel"		150/200 = 75%	7620/7800 = 97.7%
Both tests in "series"		70/200 = 35%	7770/7800 = 99.6%

3.7.5 Accuracy and Precision

Unlike sensitivity and specificity, which relate to the discriminatory powers of a test to differentiate healthy and diseased individuals, accuracy and precision relate more to quality control within the laboratory. Obviously if a test is inaccurate and lacks precision, the results will certainly influence the sensitivity and specificity of the test. However, for ease of discussion, accuracy and precision will be treated independently of sensitivity and specificity.

An accurate test gives a true measure of the substance, lesion, or structure of concern (i.e., the number of white blood cells, the level of blood sugar, the level of lead in blood, the size of follicles on ovaries). On the other hand, precision is the ability of the test to give a consistent measure upon repeated testing of the same sample. Each test will have its own inherent level of accuracy and precision.

Within limitations, accuracy is less important than precision in terms of screening tests. For example, if the extent to which a test tends to overestimate or underestimate the true level of the substance being measured is known, a correction for this may be made. When tests are not precise, more than one measurement should be made, and the average of the set of measurements used instead of just one test result.

Both precision and accuracy of a test are influenced by the variability of the test itself, the variability of the person who performs the test, and the differences between laboratories. This text is not concerned with how precision and accuracy of a test are evaluated. Nonetheless, a simple way of assessing the precision of a test performed by one person is to submit repeat samples in a blind manner and calculate the variability (variance) among results. (A blind technique is also essential when comparing test results for agreement and/or sensitivity and specificity. That is, the person performing test B should not have knowledge of the results of test A; otherwise, serious bias can occur.) Often, when using complicated tests requiring standardization on a daily basis, such a procedure will indicate that within-day precision is acceptable but between-day precision is poor. Hence paired sera (acute and convalescent) from the same animal should be tested on the same day.

The results of a study of intra- and inter-individual variation (precision) in the interpretation of canine chest radiographs are shown in Table 3.6 (Reif et al. 1970). The extent of agreement between the two radiologists was 74% and, on average, the radiologists agreed with their previous findings 82% of the time. Note the average sensitivity and specificity of chest radiographs for detecting pulmonary disease, assuming histologic diagnosis to be correct. Given the low specificity of only 87%, radiography would not be an appropriate method of screening canine populations for respira-

Table 3.6. Some findings on the sensitivity, specificity, and precision of radiographic techniques used to determine pulmonary disease in dogs

	Histological diagnosis	
	<i>D</i> +	<i>D</i> -
Radiographic <i>T</i> +	100	8
Interpretation <i>T</i> -	38	54
	138	62
	Sensitivity = $100/138 = 72.4\%$	
	Specificity = $54/62 = 87.1\%$	

In rereading 130 of the above radiographs, the two researchers disagreed with themselves 24 times and with each other 34 times, giving the following:

Intraindividual precision = 81.5% (18.5% error)

Interindividual precision = 73.9% (26.1% error)

Source: Reif et al. 1970.

tory disease if the true prevalence of disease was low. If used in this situation, the predictive value of positive radiographs would be extremely low.

3.8 Measuring Agreement

In many circumstances it is very difficult and costly to establish the true state of nature with regard to disease status. For example, the latter may require post mortem examinations, or as in the case of many viral diseases, culturing for the agent is both tedious and insensitive. Hence, in practice, veterinarians often have to utilize imperfect tests for which there are no quantitative estimates of sensitivity and specificity. In so doing, the tacit assumption is that the predictive values will be acceptable enough for practical purposes.

Under these circumstances, when a new test for disease is developed, its results are often compared to those from the current, standard, yet imperfect, test. A fictitious example of such a comparison is shown in Table 3.7. The standard test gives an apparent prevalence of 8%, the new test 10%, and both tests are positive in 4.2% of the animals. Note that these data do not directly indicate whether a positive test indicates disease (or infection) or a negative test indicates health (no infection). Thus, other than ascertaining if one test gives more positive responses than the other, all one can do is assess the extent of agreement between the test results.

An obvious measure of agreement is to calculate the observed percentage of agreement between the tests; in this example it is 90.4%. On the surface this seems quite good. However, in making this inference the implicit comparison level is no (i.e., 0%) agreement. This is incorrect, however, as there should be some agreement by chance alone. This is analagous

Table 3.7. Agreement between two tests

		Standard test		Total	Apparent prevalence
		+	-		
New	+	42	58	100	0.1
Test	-	38	862	900	...
		80	920	1000	
Apparent prevalence,		0.08			
Observed proportion agreement		$(42 + 862)/1000 = 0.904$			
Chance proportion agreement (both +)		$0.1 \times 0.08 = 0.008$			
Chance proportion agreement (both -)		$0.9 \times 0.92 = 0.828$			
Chance proportion agreement		$0.008 + 0.828 = 0.836$			
Observed minus chance agreement		$0.904 - 0.836 = 0.068$			
Maximum possible agreement beyond chance level		$1 - 0.836 = 0.164$			
Kappa		$0.068/0.164 = 0.41$			

to tossing two coins and noting the percentage of tosses in which both coins land “heads” (representing positive) or both land “tails” (representing test negative). In coin tossing, the probability of obtaining a head is 0.5 for both coins; hence, one expects agreement 50% of the time (25% of the time for heads and 25% of the time for tails). In test comparisons the probability of being test positive is given by the apparent prevalence for each test. Hence, the probability of both tests being positive is given by the product of the two apparent prevalences. Similarly, the probability of both tests being negative is given by the product of 1 minus the apparent prevalence of each test. The sum of these two probabilities gives the level of agreement expected by chance alone, 83.6% in this example. The chance level of agreement is the explicit level of comparison for assessing agreement, the observed level being 6.8% higher than the chance level in this example. To evaluate the relative magnitude of this difference, it is divided by the maximum possible agreement beyond chance, which in this example is 16.4%. The quotient (often called kappa) is 0.41. No agreement beyond chance gives a kappa of 0, and a kappa of 1 indicates perfect agreement.

A qualitative assessment of kappa suggests that if it is high, the tests are measuring what they purport to measure. If kappa is low, much uncertainty exists and in the absence of sensitivity and specificity data it is difficult to say which test provides the more valid answers. In the comparison of tests, a kappa of at least 0.4–0.5 indicates a moderate level of agreement.

In recent years, kappa has also been applied to the assessment of agreement between clinical diagnoses and to measure the “repeatability” of a clinician’s assessments on two separate occasions. Obviously, a blind technique should be used to prevent bias in these assessments. The study referred to in Table 3.6 contains sufficient data to assess between-clinician

and within-clinician agreement in the interpretation of radiographs. The levels of precision cited reflect only observed levels of agreement, not the extent of agreement beyond chance. A fictitious example based on agreement between the diagnoses of front limb lameness in horses by two clinicians is shown in Table 3.8. In this example, the observed level of agreement was 84%, the expected level by chance was 54.8%, and kappa was 0.65. Although there is little data in veterinary medicine on this subject, a kappa of 0.5–0.6 would appear to be the level anticipated from experienced clinicians when attempting to diagnose conditions of moderate difficulty. Within-clinician agreement of diagnoses made on the same subjects on different occasions will likely be somewhat higher, resulting in kappa values of 0.6–0.8.

Elucidating reasons for disagreement may allow the improvement of the test's (or clinician's) ability to correctly detect the true state of nature. General reasons for disagreement in the results of serologic tests are the absence of certain antibody classes in animals during the very early or terminal stages of disease and the presence of microorganisms antigenetically similar to those of the agent the test is designed to detect. Disagreement in clinicians' diagnoses may reflect the lack of a standardized diagnostic workup procedure, a different knowledge base, being misled by a biased history, or the inappropriate selection (or interpretation) of ancillary tests.

In any event, the application of sensitivity and specificity concepts as well as measures of agreement beyond chance to the evaluation of tests and clinician abilities should result in more refined tests and improved diagnostic ability.

Table 3.8. Agreement between two clinicians diagnosing reasons for front limb lameness in horses

		Clinician 2		Total	Apparent prevalence
		ND	OD		
Clinician 1	ND	26	4	30	0.3
	OD	12	58	70	...
		38	62	100	
Apparent prevalence,		0.38			
Observed proportion agreement				$(26 + 58)/100 = 0.84$	
Chance proportion agreement ND				$0.3 \times 0.38 = 0.114$	
Chance proportion agreement OD				$0.7 \times 0.62 = 0.434$	
Chance proportion agreement				$0.114 + 0.434 = 0.548$	
Observed minus chance agreement				$0.84 - 0.548 = 0.292$	
Maximum possible agreement beyond chance level				$1 - 0.548 = 0.452$	
Kappa				$0.292/0.452 = 0.65$	

Note: ND = Navicular disease; OD = Other disease

References

- Dodd, K. 1978. Estimation of the sensitivity, specificity and predictive value of the intradermal tuberculin test. *Irish Vet. J.* 32:87-89.
- Elandt-Johnson, R. C. 1975. Definition of rates: Some remarks on their use and misuse. *Am. J. Epidemiol.* 102:267-71.
- Glickman, L., P. Shantz, R. Dombroske, and R. Cypess. 1978. Evaluation of sero-diagnostic tests for visceral larva migrans. *Am. J. Trop. Med. Hyg.* 27:492-98.
- Green, A. 1982. The epidemiologic approach to studies of association between HLA and disease. I. The basic measures, concepts and estimation procedures. *Tissue Antigens* 19:245-58.
- Kleinbaum, D. G., L. L. Kupper, and H. Morgenstern. 1982. *Epidemiological Research: Principles and Quantitative Methods*. Belmont, Calif.: Wadsworth.
- Martin, S. W. 1977. The evaluation of tests. *Can. J. Comp. Med.* 44:19-25.
- _____. 1984. Estimating disease prevalence and the interpretation of screening test results. *Prev. Vet. Med.* 2:463-72.
- Ransohoff, D. F., and A. R. Feinstein. 1978. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *New Eng. J. Med.* 299:926-30.
- Reif, J. S., W. H. Rhodes, and D. Cohen. 1970. Canine pulmonary disease and the urban environment. I. The validity of radiographic examination for estimating the prevalence of pulmonary disease. *Arch. Environ. Health* 20:676-83.
- Robertson, T. G. 1963. Diagnosis of bovine tuberculosis. I. The evaluation of tuberculin tests. *N.Z. Vet. J.* 11:6-10.
- Seiler, R. J. 1979. The non-diseased reactor: Considerations on the interpretation of screening test results. *Vet. Rec.* 105:226-28.
- Williamson, N. 1980. Reproductive performance and recording systems. Proc. 13th Am. Assoc. Bov. Pract., November 19-22, Toronto, Canada.