

Veterinary Epidemiology: Principles and Methods

Part 1: Basic Principles

Chapter 2: Sampling Methods

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>

Sampling Methods

Good sample design is an essential component of surveys and analytic studies. Hence, this chapter contains methods for obtaining data from a representative subset (sample) of a population and makes inferences about the characteristics of the population. Other aspects of data collection (e.g., questionnaire design) are discussed in 6.1.

Sometimes data from a census are available to describe events in a population; no sampling is required and hence no information is lost, as can occur when selecting only a subset of the population. More frequently, data are available from only a subset of the population, and that subset may or may not have been selected by formal sampling methods. For example, data from outbreak investigations or routinely collected data from hospitals or client records (e.g., case reports) may be viewed as arising from a sample of the population, although no formal sampling is used. As will become apparent, there are fewer problems in extrapolating from data obtained by formal planned sampling than from data whose collection was unplanned.

There are two reasons why an epidemiologist would take a planned sample of a population. One is to describe the characteristics (i.e., frequency and/or distribution of disease or production levels) of a population. Examples might include selecting a sample of dairy cows to estimate the extent of subclinical mastitis in a population and selecting a sample of the dog population to estimate the percentage vaccinated against diseases such as rabies. Descriptive studies such as these are called surveys. The process of collating and reporting information from planned surveys, routinely collected data, or outbreak investigations is termed descriptive epidemiology (see Chapter 4).

The second reason for taking a planned sample is to assess specific associations (e.g., test hypotheses) between events and/or factors in the population. Examples would be a sample designed to look for associations

between the type of milking equipment and milking procedures and the level of mastitis in the herd, or a study designed to test the hypothesis that certain phenotypes of dogs are more susceptible to bone cancer than others. Studies such as these are analytic studies, and the process of collating, analyzing, and interpreting the information is termed analytical epidemiology (see Chapter 6). In practice, the differences between these types of observational studies often become nebulous. For example, it is not uncommon to do some hypothesis testing using data from surveys. Nonetheless, since the main emphasis of surveys differs from hypothesis testing, the distinction is maintained to simplify and add order to the description of the underlying sampling strategies.

Whether the study is a survey or an analytic study, how the study members are obtained from the population (i.e., the method of sampling) will determine the precision and nature of extrapolations from the sample to the population. Planning the sampling strategy is a major component of survey design. Although sampling per se is only a small part of the design of an analytic study, its central importance is indicated by the fact that the three common types of analytic studies are named on the basis of the sample selection strategy.

Further details on sampling are available in a number of texts (Snedecor and Cochran 1980; Cochran 1977; Levy and Lemeshow 1980; Leech and Sellers 1979; Schwabe et al. 1977). An excellent manual on sampling in livestock disease surveys is provided by Cannon and Roe (1982).

2.1 General Considerations

State the objectives clearly and concisely. The statement should include the parameters being estimated and the unit of concern. Usually, it is best to limit the number of objectives, otherwise the sampling strategy and study design can become quite complex.

The investigator usually will have a reference or target population in mind. This population is the aggregate of individuals whose characteristics will be elucidated by the study. The population actually sampled is often more restricted than this target population, and it is important that the sampled population be representative of the target population. It would be inappropriate to attempt to make inferences about the occurrence of disease in the swine population of an entire country (the target population) based on a sample of swine from one abattoir or samples obtained from a few large farms (the sampled population). As another example, data from diagnostic laboratories usually are not representative of problems in the source population and hence would not be appropriate for estimating disease prevalence.

In planning a sample, note the type and amount of data to be col-

lected. If the objectives are straightforward and few in number, this aspect of planning is easy. At this stage of planning, explicit definitions of the outcome must be considered. That is, in a study to estimate the frequency of metritis in dairy cows, the outcome (metritis), must be clearly defined. This increases the scientific validity of the study and allows other workers to compare their results (similarities and differences) to those of the survey. Related to this matter is the data collection method (e.g., personal interview, mailed questionnaire, special screening tests). Identifying the validity and accuracy of data collection methods are discussed in Chapter 3.

Because the results of samples are subject to some uncertainty due to sampling variation, it is important to consider how precise (quantitatively) the answer needs to be. The results of different samples will, in general, not be equal; the greater the precision required (the smaller the sample to sample variation), the larger the sample must be. Factors that influence the number of sampling units required in surveys are discussed in 2.2.8, analytic studies in 2.4.4.

Prior to selecting the sample, the sampled population must be divided into sampling units. The size of the unit can vary from an individual to an aggregate of individuals, such as litters, pens, or herds. The list of all sampling units in the sampled population is called the sampling frame. Often because of practical considerations, although the unit of concern may be individuals, aggregates of individuals are used as the initial sampling unit. For example, although the objective might be to estimate the prevalence of brucella antibodies in cattle (the unit of concern), the initial sampling unit might be the herd, since a list of all cattle in the population would be difficult to construct. In other instances, to estimate the average somatic cell count of milk in dairy herds, the unit of concern is the herd and it also could be the sampling unit (e.g., a convenient way of obtaining a representative sample of milk from the herd would be to take an aliquot portion of milk from the bulk milk tank).

Finally, before proceeding with the full study it is important to pretest the procedures to be used. Such pretesting should be sufficiently rigorous to detect deficiencies in the study design. This would include the sample selection, clarity of questionnaires, and acceptability and performance of screening tests. This pretest should also be used to evaluate whether the data to be collected in the actual study are appropriate to answer the original objectives.

2.2 Estimating Population Characteristics in Surveys

To provide a practical illustration of the different methods of survey sampling, assume that the investigator wishes to estimate the percentage of adult cows (beef and dairy) in a large geographic area that have antibodies

to enzootic bovine leukosis virus. The unit of concern is the cow, and the true but unknown percentage of reactor cows in the target population is the parameter to be estimated. N represents the number of cows in the population and n the number of cows in the sample.

2.2.1 Nonprobability Sampling

Nonprobability sampling is a collection of methods that do not rely on formal random techniques to identify the units to be included in the sample. Some nonprobability methods include judgment sampling, convenience sampling, and purposive sampling.

In judgment sampling representative units of the population are selected by the investigator. In convenience sampling, the sample is selected because it is easy to obtain; for example, local herds, kennels, or volunteers may be used. Using convenience or judgment sampling often produces biased results, although some people believe they can select representative samples. This drawback and the inability to quantitatively predict the sample's expected performance suggest these methods rarely should be used for survey purposes. In purposive sampling, the selection of units is based on known exposure or disease status. Purposive sampling is often used to select units for analytic observational studies, but it is inadequate for obtaining data to estimate population parameters.

Examples of the application of nonprobability sampling to estimate the prevalence of enzootic bovine leukosis virus include the selection of cows from what the investigator thinks are representative herds and the selection of cows from herds owned by historically cooperative or nearby farmers.

The following sampling methods belong to a class known as probability samples. The discussion assumes that sampling is performed without replacement; hence an individual element can only be chosen once.

2.2.2 Simple Random Sampling

In simple random sampling, one selects a fixed percentage of the population using a formal random process; as for example, flipping a coin or die, drawing numbers from a hat, using random number generators or random number tables. ("Random" is often used to describe a variety of haphazard, convenience and/or purposive sampling methods, but here it refers to the formal statistical procedure.) Strictly speaking, a formal random selection procedure is required for the investigator to calculate the precision of the sample estimate, as measured by the standard error of the mean. In practice, formal random sampling provides the investigator with assurance that the sample should be representative of the population being investigated, and for the parameter being estimated, confidence intervals are calculated on this premise. Despite mathematical and theoretical advan-

tages, simple random sampling is often more difficult to use in the field than systematic sampling (described in 2.2.3). Consider the procedure for selecting a sample of 10% of feedlot steers as they pass through a handling facility. In simple random sampling, a list of randomly obtained numbers—representing, for example, the animals' identification (i.e., ear tags) or the order of the animals through a handling facility—would be prepared beforehand to identify the animals for the sample. The practicalities of using such a list in a field situation (e.g., losing count of animals and/or continuously having to refer to a list of numbers) may make this type of sampling inappropriate.

To obtain a simple random sample of cows for the prevalence of enzootic bovine leukosis antibodies one would obtain a list of n random numbers between 1 and N , each number identifying a cow in the sampling frame. Thus the cows selected would be distributed randomly throughout the sampled population.

2.2.3 Systematic Random Sampling

In systematic sampling the n sampling units are selected from the sampling frame at regular intervals (e.g., every fifth farm or every third animal), thus the interval k is 5 or 3 respectively. If k is fixed initially, n will vary with N ; whereas if n is fixed initially, k becomes the integer nearest to N/n . When systematic methods are used, the starting point in the first interval is selected on a formal random basis.

Systematic sampling is a practical way to obtain a representative sample, and it ensures that the sampling units are distributed evenly over the entire population. There are two major disadvantages of this method. First, it is possible that the characteristic being estimated is related to the interval itself. For example, in estimating the prevalence of respiratory disease in swine at slaughter, one might systematically select a day of the week (e.g., Wednesday) to examine lungs. If swine slaughtered on Wednesdays were not representative of swine slaughtered on the other days of the week (e.g., because of local market customs), a biased result would be obtained. The second disadvantage is the difficulty of quantitatively assessing the variability of estimates obtained by systematic random sampling. In practice, one uses methods appropriate for simple random sampling to obtain these estimates.

If N/k is not an integer, some bias will result in the sample estimate because some animals (elements) will have more impact on the mean than others. This is of little concern if N is large and k is small relative to N . To prevent this bias, select the desired k and draw a random number (RN) between 1 and N ; then divide RN by k and note the remainder. This remainder identifies the starting point between 1 and k (i.e., a remainder of 0 means the starting point is the k th individual, a remainder of 2 the second

individual, and so forth) (Levy and Lemeshow 1980, p 76).

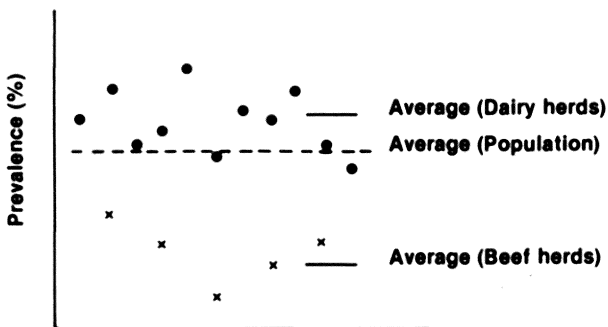
In sampling to estimate the prevalence of antibodies to enzootic bovine leukosis virus, using a list of all N cows in the area in question (the sampling frame), the initial animal to be tested would be selected from the first N/n animals randomly. Subsequently, every k th cow would be tested. In selecting 10% of steers, one could randomly select a number between 1 and 10 (say 6) and then the 6th, 16th, 26th, etc. animal through the facility would be included in the sample.

2.2.4 Stratified Random Sampling

In stratified sampling, prior to selection, the sampling frame is divided into strata based on factors likely to influence the level of the characteristic (e.g., prevalence of antibodies) being estimated. Then a simple random or systematic random sample is selected within each stratum.

Stratified sampling is more flexible than simple random sampling because a different sampling percentage can be used in the various strata (e.g., 2% in one stratum and 5% in another). Also, the precision of the sample estimate may be improved, because only the within-stratum variation contributes to the variation (standard error) of the mean in stratified sampling; whereas in simple random sampling both the within-stratum and the between-stratum variation are present. A graphic illustration of this feature is shown in Figure 2.1.

In simple random sampling, the variability of the estimate of prevalence has components related to both within-herd type and between-herd type variation in prevalence. In stratified random sampling, the variability of the estimate has components related to only the within-herd type variation in prevalence; hence its variability is expected to be less than that



2.1. Prevalence of disease X in population of dairy and beef cattle herds: relationship of sampling design to variability of sample means.

obtained in simple random sampling. For example in Figure 2.1 the variability of the prevalence in beef herds, about the mean for beef herds, and the variation of the prevalence in dairy herds, about the mean for dairy herds, are much smaller than if type of herd is ignored and the variation of herd disease prevalence about the overall mean is calculated. Variation (see Table 2.1) of the mean (estimate of prevalence) is calculated using standard formula for the variance or its square root, the standard deviation. The standard deviation of a mean is referred to as a standard error.

The obvious disadvantage of stratified sampling is that the status of all sampling units, with respect to the factors forming the strata, must be known prior to drawing the sample. In general, the number of factors used for stratification should be limited to those likely to have a major impact on the value of the characteristic (e.g., prevalence of antibodies) being estimated.

As an example of this method and given that dairy cows are likely to have a higher rate of enzootic bovine leukosis antibodies than beef cows, one should obtain a more precise estimate of the population mean (prevalence) if strata were formed based on type of cow. Also, if 60% of the cow population N comprised dairy cows, 60% of the sample n should be dairy cows. This is called proportional weighting, and it keeps the arithmetic involved in calculating the sample statistic simple. Cows would be selected within each stratum by using simple random or systematic random sampling methods.

In the sampling methods discussed, the sampling unit and the unit of concern are the same (i.e., a cow). These methods are well suited for sampling from laboratory files or from relatively small groups of identifiable animals. However, the practical difficulty of obtaining a list (the sampling frame) of all cows in a large geographic area such as a province or state is a drawback. Additionally, with stratified sampling, the appropriate characteristics of each sampling unit must be identified (e.g., as dairy or beef in the previous example). To overcome these problems, allow flexibility in sampling strategy, and decrease the cost of the sampling, it is often easier to initially sample herds or other natural aggregates of animals within the area, although individual animals are the units of concern. Two of the more common sampling methods used for this purpose are cluster and multistage sampling.

2.2.5 Cluster Sampling

In cluster sampling, the initial sampling unit is larger than the unit of concern (e.g., usually the individual). Clusters of individuals often arise naturally (e.g., litters, pens, or herds) or they may be formed artificially (e.g., geographic clusters). Administrative units such as counties may also be used as artificial clusters for sampling purposes. The clusters (sampling units) can be selected by systematic, simple, or stratified random methods;

all individuals within the sampling units are tested.

Sometimes the group, be it a herd, pen, or litter, is the unit of concern, and therefore is not considered to be a cluster. Some examples of this situation are investigations to classify herds as to whether they are infected with enzootic bovine leukosis; estimation of the mean somatic cell count for dairy herds using bulk tank milk samples; and estimation of the mean herd milk production or days to conception.

In the bovine leukosis example, a cluster sample could be obtained by taking a simple random sample of all herds in the sampled population and testing all cows within the selected herds. From the formula in Table 2.1, note that the variability of the mean of the cluster sample is a function of

Table 2.1. Formulas for estimating simple characteristics of populations

Type of random sample	Estimates of mean	Estimates of precision (standard error of mean)
<i>Simple</i>	$y = \Sigma y_i/n$	$se(y) = (s^2/n)^{1/2}$ where $s^2 = \Sigma (y_i - y)^2/(n - 1)$ $s^2 = pq$ for attributes
	<p>y_i = value of variable y in ith individual. If an attribute (e.g., disease) is being measured, $y_i = 1$ if present and 0 if absent; hence $p = \Sigma y_i/n$ and $q = 1 - p$</p> <p>n = sample size and N = population size</p> <p>If $n/N > 0.1$, then s^2 is adjusted by multiplication by $1 - n/N$</p>	
<i>Stratified</i>	$y_n = \Sigma W_j y_j$ or $y_n = \Sigma y_{ij}/n$	$se(y_n) = [\Sigma (W_j^2 s_j^2/n_j)]^{1/2}$ where $W_j = N_j/N$ and $s_j^2 = \Sigma (y_{ij} - y_j)^2/(n_j - 1)$ $s_j^2 = \bar{p}_j \bar{q}_j$ for attributes
	<p>The subscript j indicates the stratum</p> <p>W_j = population weighting factor: the proportion of the population in the jth stratum, i.e., N_j/N. Second formula for the mean assumes proportional weighting, i.e., $w_j = W_j$</p> <p>Y_{ij} = value of variable y in ith individual in jth stratum. If an attribute is being measured, $Y_{ij} = 1$ if present and 0 if absent</p> <p>y_j = mean of jth stratum</p> <p>n_j = number of individuals in jth stratum in sample</p> <p>N_j = number in jth stratum in population</p> <p>If $n_j/N_j > 0.1$, each s_j^2 may be adjusted by multiplying by $1 - n_j/N_j$</p>	
<i>Cluster (equal sized clusters only)</i>	$y_{cs} = \Sigma y_c/m$ = $\Sigma y_{ic}/mn$	$se(y_{cs}) = (s^2/m)^{1/2}$ where $s^2 = \Sigma (y_c - \bar{y}_{cs})^2/(m - 1)$
	<p>y_c = mean of variable y in cth cluster; it is p_c for attribute variables, but treat these as continuous variables</p> <p>y_{ic} = value of ith individual in cth cluster. If an attribute is being measured, $y_{ic} = 1$ if present and 0 if absent</p> <p>m = number of clusters in sample (M is number of clusters in population) each containing n individuals</p> <p>Adjust s^2 if $m/M > 0.1$ using multiplication by $1 - m/M$</p>	

the between-herd variance and the number of clusters m in the sample, not the number of animals in the sample.

2.2.6 Multistage Sampling

This method is similar to cluster sampling except that sampling takes place at all stages. As an example of two-stage sampling, one would begin as in cluster sampling by selecting a sample of the primary units (e.g., herds) listed in the sampling frame. Then within each primary unit, a sample of secondary units (e.g., animals) would be selected. Thus the difference between cluster and two-stage sampling is that subsampling within the primary units is conducted in the latter method.

Multistage sampling is used because of its practical advantages and flexibility. The number of primary (n_1) and secondary units (n_2) may be varied to account for different costs of sampling primary versus secondary units as well as the variability of the characteristic being estimated between primary units and between secondary units within primary units (see 2.2.9).

To continue with the bovine leukosis example, one could proceed in the same manner as cluster sampling, but after selection of the herds (the primary units), a simple or systematic random sample of cows within each herd (the secondary units) would be selected. This process could be extended to three-stage sampling by selecting small geographic areas as the primary units, selecting herds within these areas as secondary units, and finally selecting animals within the herds as tertiary units. Whenever possible, one should select each stage's sampling units with probability proportional to the number of individuals they contain. This minimizes the error of estimate and stabilizes the sample size. The main disadvantage of cluster and multistage samples is that more individuals may be required in the sample to obtain the same precision as would be expected if individuals could be selected with simple random sampling.

As an illustration of multistage sampling, suppose that in the bovine leukosis example there are M farms (say 120) and N animals (say 8000) in the population. The objective is to estimate the proportion of animals having enzootic bovine leukosis antibodies using a sample size of 800 ($n = 800$). The sampling frame would have the format shown in Table 2.2.

Suppose the number of primary sampling units (farms) to be selected is $40(n_1)$ and, on average, $20(n_2)$ secondary units (animals) will be selected from within each primary unit. (Note that $n_1 \times n_2 = n$.) If the number of animals in each herd was unknown, one could take a simple or systematic random sample of 40 herds and randomly select a fixed percentage (i.e., $30\% = Mn/mN$) of the animals in each herd for testing. When the number of animals in each herd is known, a more optimal procedure is to sample the primary units with probability proportional to their size, and then to select a fixed number of animals from each herd. In this example, the initial step is to randomly select 40 numbers within the range of 1 to 8000. Each of

Table 2.2. Format for a sampling frame for two-stage sampling

Farm number	Number of animals	Cumulative number of animals
1	62	1–62
2	48	63–110
3	74	111–184
4	36	185–220
.	.	.
.	.	.
119	42	7900–7941
120	59	7942–8000

the random numbers will identify a farm according to the cumulative number column. Subsequently, 20 animals may be randomly selected from each farm. Both of these procedures give each individual the same probability of being selected. Since it is assumed that sampling is without replacement, if a farm is identified twice, another should be selected randomly. (Technically it would be better to randomly select twice the number of animals from that herd.) If fewer than 20 animals are present in a specified herd, the practical solution is to test all available animals.

A modification of this method to ensure that each farm may be selected only once is the use of systematic random techniques. For example, the selection interval k is found by dividing the total number of animals N by n_1 (in this case, $k = 8000/40 = 200$). A number is then selected randomly from the range 1 to k (e.g., 151). The remaining 39 numbers (351, 551, etc.) would identify the farms to include in the sample. This process will select a farm only once, providing the interval k is greater than the number of animals on the largest farm.

2.2.7 Calculating the Estimate

The point estimate of the prevalence of reactors in the population, the parameter $P(T+)$, is the test-positive proportion in the sample, the statistic $p(T+)$ or \bar{p} . To calculate this statistic the number of test positives are added together and divided by the sample size. (This assumes a proportionally weighted sample when stratified sampling is used, which is self-weighting in terms of the mean. The same approach is also used for estimates obtained from cluster or multistage samples. See Snedecor and Cochran 1980 for details.) Calculating the estimate of a population mean (say average milk production) is performed in an analogous manner (see 3.6).

EXAMPLE CALCULATIONS In the enzootic bovine leukosis example, if 125 of 2000 cows were test-positive, the estimate of the prevalence of reactors in the population would be $\bar{p} = 125/2000 = 0.063$ or 6.3%. If a

simple random sample or systematic random sample were used to obtain the sample, the variability of the point estimate would be:

$$\begin{aligned}\text{Variance } (\bar{p}) &= \bar{p}(1 - \bar{p})/n = 0.063 \times 0.937/2000 \\ V(\bar{p}) &= 0.295 \times 10^{-4} \\ \text{Standard Error } (\bar{p}) &= V(\bar{p})^{1/2} \\ SE(\bar{p}) &= 0.0054 \text{ (0.54\%)}\end{aligned}$$

These estimates could be written as 6.3% \pm 0.5% (*SE*). With moderately large sample sizes, 65% of all possible sample means will be within 1 standard error of the true mean, 95% within 1.96 standard errors, and 99% within 2.6 standard errors. The calculation of a confidence interval as an extension of the above facts is described in 3.6. More complex calculations are required to determine the variability of means obtained from cluster or two-stage samples (see Table 2.1). Since the clusters are rarely of equal size, the reader can use the formula shown in Table 2.1 for the initial calculations, but should consult one of the reference texts for details of more accurate methods.

2.2.8 Sample Size Considerations

Accurate determinations of the sample size required for a survey can be quite detailed, and most complex surveys will require the assistance of a statistician. For less complex surveys one of the following formulas should provide suitable estimates.

To determine the sample size n necessary to estimate the prevalence of reactors $P(T+)$ in a population (the mean of a qualitative variable, morbidity rates or mortality rates, see 3.2 and 3.3), the investigator must provide an educated guess of the probable level of reactors \hat{P} (read “ P hat”), and must specify how close to $P(T+)$ the estimate should be.

EXAMPLE CALCULATIONS Suppose the available evidence suggests that approximately 30% ($\hat{P} = 0.3$) of the cow population will have antibodies to enzootic bovine leukosis. Also, assume the investigator wishes the survey estimate to be within 6% of the true level 95% of the time. (6% is termed the allowable error, or required precision, and is represented in the following formula by L .) Then the required sample size is:

$$\begin{aligned}n &= 4\hat{P}\hat{Q}/L^2 \quad \text{where } \hat{Q} = 1 - \hat{P} \\ &= 4 \times 0.3 \times 0.7/0.06^2 = 0.84/0.0036 = 233\end{aligned}$$

Thus approximately 230 cows would be needed for the survey.

In general, the number of animals in the population has little influence

on the required sample size except when n is greater than $0.1N$. For example, if the herd contained only 200 cows ($N = 200$), the required number of cows is found using the reciprocal of $1/n^* + 1/N$ where n^* is the above sample size estimate. In this instance, the number required to obtain the same precision is the reciprocal of $1/233 + 1/200 = 1/108$; thus the required sample is approximately 108 animals (Cannon and Roe 1982).

When determining the sample size necessary to estimate the mean of a quantitative variable (e.g., production parameters, see 3.6), the investigator needs to supply an estimate of the standard deviation or variance of that variable in the target population and specify how close to the mean the sample estimate should be. Suppose reproductive efficiency as measured by the calving-to-conception interval is the event of interest. Assume that the available evidence suggests that the standard deviation of this interval is 20 days, and the investigator wishes the sample to provide an estimate within 5 days of the true average 95% of the time. Then $\hat{S} = 20$ and $L = 5$, and the required sample size is:

$$n = 4\hat{S}^2/L^2 = 4 \times 20^2/5^2 = 1600/25 = 64$$

Thus approximately 64 cows are required for the survey.

The number 4 in the previous formulas is the approximate square of $Z = 1.96$, which provides a 95% confidence level. If the investigator wished to be 99% certain that the results would be within $\pm L$ of the true level, 6.6 (the approximate square of $Z = 2.56$) should be substituted for 4. The reader is encouraged to experiment with different values in each of the above formulas to assist in understanding the consequences of these changes.

In using the above formulas, it is assumed that the sampling unit is the same as the unit of concern. When using cluster or multistage sampling, an upward adjustment in the sample size may be required to obtain the desired precision in the estimate. If the disease is not very contagious and/or the within-primary-unit correlation coefficient is small, a two to three times increase in the sample size should be appropriate. For very contagious diseases, the necessary sample size may have to be increased five to seven times (Leech and Sellers 1979). These increases are based on rule-of-thumb, and more accurate formulas as described in 2.2.9 should be used when the appropriate information on the within- and between-herd variances is available.

2.2.9 Cost considerations in survey design

Frequently, the investigator must perform the sampling under monetary as well as practical and biologic constraints. Thus, rather than only specifying the precision of the estimate, the investigator may seek to obtain

the highest precision for a specified cost or, conversely, the least cost for a specified precision.

Simple probability sampling procedures are not particularly flexible in terms of meeting monetary constraints, other than altering (usually reducing) the total number of sampling units studied. However, stratified sampling allows the investigator to select different numbers of units from different strata, depending on the relative costs associated with sampling in each stratum. The basic rule is to reduce the number of samples in strata with high sampling costs and to increase the number with lower sampling costs. The optimal stratified sample will have stratum weights proportional to $N_j S_j / C_j^{1/2}$ where N_j is the number in the population in stratum j , S_j is the standard deviation of the parameter being measured in stratum j , and C_j is the cost of sampling in stratum j . If the resulting sample is not proportionally weighted according to the population structure, the calculation of the sample mean should be done using the weighting formula in Table 2.1.

Cluster sampling is often used because of practical difficulties in obtaining a sampling frame in which the individual is the sampling unit. Thus circumventing these "practical difficulties" by using cluster sampling is really a reaction to economic constraints. For example, it may cost less to sample 4000 swine using cluster sampling than to sample 1000 using random sampling, although the precision of the estimate obtained by the latter may be greater than that obtained using cluster sampling with more individuals.

The most flexible sampling method to take account of cost factors is multistage sampling. In two-stage sampling one may vary the number of primary and secondary units selected according to the costs of sampling primary units (e.g., herds) as well as the costs of sampling secondary units (e.g., animals within a herd). In the enzootic bovine leukosis example, the cost of traveling to a herd to obtain samples may be large relative to the cost of obtaining a sample from an individual cow once on the farm. This would suggest an increase in the number of secondary units (cows) and a decrease in the number of primary units (herds) to reduce the total cost of sampling. The balance between primary and secondary sampling units can be investigated formally. If c is the total monies available for sampling, c_1 the cost of sampling primary units, and c_2 the cost of sampling secondary units, the relationship between these costs and the numbers of primary and secondary units is:

$$c = c_1 n_1 + c_2 n_1 n_2$$

The appropriate number of secondary units n_2 to select, minimizing costs for a given precision, or vice-versa (Snedecor and Cochran 1980), is found using:

$$n_2 = (c_1 s_2^2 / c_2 s_1^2)^{1/2}$$

The number of primary units n_1 may then be found using the previous formula, since c , c_1 , c_2 and n_2 are known. If $c_1 = c_2$, then n_2 is merely a function of the respective variances; namely, $n_2 = (s_2^2/s_1^2)^{1/2}$.

EXAMPLE CALCULATIONS Suppose a person wished to estimate the blood globulin level in mature dairy cows. Assume that the total money available for the project (c) is \$10,000, that it will cost an average of \$100 per farm (c_1) to sample each herd (this includes travel costs), and that the cost per cow (c_2) is \$10 once at the herd (this includes the cost of blood vials, needles, technician time, and laboratory analysis). Assume also that the between-herd variability (s_1) in globulin concentration is 8g/l and the within-herd (cow-to-cow) variability (s_2) is 4 g/l. On this basis,

$$n_2 = (100 \times 4^2 / 10 \times 8^2)^{1/2} = 2.5^{1/2} = 1.6$$

Since n_2 should be an integer, round 1.6 to 2 cows per herd. Now, solve the initial cost equation for n_1 .

$$\begin{aligned} 10,000 &= 100n_1 + 10 \times 2n_1 = 120n_1 \\ n_1 &= 83 \end{aligned}$$

Thus, approximately 80–85 herds would be used, taking 2 cows per herd.

Despite the high cost per herd, the relatively large between-herd variability dictates that a large number of herds are required. In this instance, if $c_1 = c_2$, the ratio $(s_2^2/s_1^2)^{1/2}$ indicates that one animal (the minimum number) per herd should be selected.

2.3 Sampling to Detect Disease

As part of many disease control or eradication programs, entire herds or flocks are tested to ascertain if the specified disease is present or, conversely, to ensure that the disease is absent. However, testing entire herds or flocks is expensive, and the veterinarian may have to accept the results of testing only a portion of the animals.

When sampling is used for this purpose, a frequently asked question is, What sample size is required so that the veterinarian can be 95% or 99% confident that the herd or flock is disease-free if no animals or birds in the sample give a positive test result? To actually prove (i.e., be 100% certain) that a disease is absent from a population requires testing almost every individual. For example, to prove that atrophic rhinitis was not present in a

5000 pig feeder operation would require the examination of the snout of virtually every pig.

Despite these limitations, sampling can provide valid insight into the health status of the population, because it is rare for only one animal in a herd to have the disease of interest. Infectious diseases tend to spread, and even infrequent noninfectious diseases would be expected to cluster somewhat within a herd, assuming environmental determinants of the disease are present. Thus for many diseases, if the disease is present at all, the herd will be likely to contain more than one diseased individual. This knowledge may be utilized when sampling to detect disease. The sampling strategy is designed to detect disease if more than a specified number or percentage (>0) of animals have the disease. The actual number or percentage of diseased animals to specify when making the sample size calculations should be based on knowledge of the biology of the disease. Often, the results of previous testing campaigns will supply useful information. For example, available data might indicate that the percentage of cattle with bovine tuberculosis in infected herds averages between 5 and 10%. These could be used as starting points to determine the possible range of sample sizes required to detect bovine tuberculosis when it is present.

Table 2.3 contains the sample size required to be 95% or 99% certain that at least one animal in the sample would be diseased if the disease were present at or above the specified level. The minimum number of diseased animals assumed to be present in a herd is one, and for populations of greater than 100 individuals, the number of diseased animals is based on assumed prevalences ranging from 1–50%. Note that a formal random sampling method, with individuals as the sampling units, is required if the desired confidence level shown is to be attained. If no formal random selection is used, the confidence one can have in the result is unknown, at least quantitatively. This circumstance may arise when animals are examined at slaughter for the presence of disease (e.g., in slaughter checks of pigs for respiratory disease). The pigs examined may not be representative of the source population; for example, the disease of interest may have a high case fatality rate and hence only disease-free animals survive to market age and weight. Although sample size requirements may be calculated to assist in evaluating the potential workload, one should be cautious and assign only a judgmental level of confidence if no diseased animals are observed in an informal sample such as this. Sometimes it may be assumed with a high degree of certainty that the level of disease in culled animals is much higher than in the source population; these diseases influencing the withdrawal of the animal in the first instance. If a sufficient number of these animals are examined and are found to be disease-free, the source herd or flock may be deemed disease-free, although no formal sampling was used in selecting the culled animals to be examined. (In fact, if a high

Table 2.3. Sample sizes required to be 95/99% confident disease is present at/or below specified prevalence D/N, if no diseased animals are observed

Population size	Prevalence of disease: (D/N) × 100			
	1%	5%	10%	50%
30	29/30	23/27	19/23	5/7
60	57/60	38/47	23/31	5/7
100	95/99	45/59	25/36	5/7
300	189/235	54/78	28/41	5/7
500	225/300	56/83	28/42	5/7
1,000	258/367	58/86	29/43	5/7
10,000	294/448	59/90	29/44	5/7

*The minimum number of diseased animals is one, at 1% and 5% prevalence in populations of size 30 and 60 respectively.

The above sample size requirements were derived using the following formula from Cannon and Roe (1982):

$$n = [1 - (1 - a)^{D/N}] [N - (D - 1)/2]$$

where *n* is the required sample size

a = probability (confidence level) of observing at least one diseased animal in sample when the disease affects at least D/N in population

D = number of diseased animals in population

N = population size

Note: If the column heading *D/N* is read as the proportion of animals in a population that is tested (*n/N*), the body of the table provides the expected maximum number of cases in the population.

percentage of culled animals are tested at slaughter, the tested animals essentially are a census of all culled animals. The problem in this case is not so much concerned with sampling, but with the amount of information about the population of interest provided by testing the culled animals.)

EXAMPLE CALCULATIONS Assume that in a population of 1000 (*N*) swine, there will be at least 10 (*D*) pigs with atrophic rhinitis, if it is present at all. The sample size required to be 95% (*a* = 0.95) sure of detecting at least one pig with rhinitis is:

$$n = [1 - (1 - 0.95)^{0.1}] [1000 - (9/2)] = 0.259 \times 995.5 = 258$$

To be 99% certain of detecting at least one pig with rhinitis under the conditions in this example, the required sample size is:

$$n = 0.369 \times 995.5 = 367$$

The previous formula may be solved for *D*, rather than *n*, and the following formula results:

$$D = [1 - (1 - a)^{1/n}] (N - [(n - 1)/2])$$

This formula is useful to provide the maximum number of diseased animals (*D*) expected in a population, with confidence *a*, when *n* individuals are examined and found to be free of disease.

EXAMPLE CALCULATIONS If 20 randomly selected layer hens from a flock of 5000 are examined and found to be free of pullorum disease, the maximum expected number of infected birds in that flock would be:

$$\begin{aligned} D &= [1 - (1 - 0.95)^{0.05}][5000 - (19/2)] \\ &= 0.139 \times 4990.5 = 694 \end{aligned}$$

giving a maximum percentage with pullorum disease of 13.9%. If 200 randomly selected hens were all negative, the maximum expected number infected in the flock would be 73, or a maximum prevalence of 1.5%.

As noted, Table 2.3 can be used to obtain the maximum number diseased by changing the column header *D/N* to *n/N* where *n/N* represents the percentage of the population examined and found disease-free. The body of the table will provide the maximum number of diseased individuals expected in a population of size *N*.

2.4 Hypothesis Testing in Analytic Observational Studies

The three sampling methods – each denoting a type of analytic study – described in this section differ in the amount of information they provide with respect to the population. Cross-sectional studies are based on a single sample of the population, whereas, in principle, cohort and case-control studies are based on two separate often purposive samples (Fleiss 1973).

To assist the description of these sampling methods, the basic population structure with respect to one exposure factor (often called the independent variable) and one disease (often called the dependent variable) both with two levels, present or absent, is shown below. The letters *A*, *B*, *C*, and *D*, represent the number of individuals (sampling units) in each factor-disease category in the population.

		Diseased (<i>D</i> +)	Not diseased (<i>D</i> -)	
Exposed	(<i>F</i> +)	<i>A</i>	<i>B</i>	<i>A</i> + <i>B</i>
Not exposed	(<i>F</i> -)	<i>C</i>	<i>D</i>	<i>C</i> + <i>D</i>
		<i>A</i> + <i>C</i>	<i>B</i> + <i>D</i>	<i>N</i> = <i>A</i> + <i>B</i> + <i>C</i> + <i>D</i>

A variety of rates and proportion can be calculated if the numbers in each of the four cells (factor-disease combination) are known. The objective of analytic studies is to estimate these rates, although not all may be estimated from each study design. See Table 2.4.

For purposes of nomenclature, lowercase characters indicate that the values are derived from a sample, whereas uppercase characters indicate population values. Thus p indicates an estimate, that is a statistic, from a sample, whereas P indicates the corresponding population value or parameter. In discussing numbers of individuals as opposed to proportions, n will be substituted for p . For example, $n(F+)$ is the number of exposed units in the sample which may also be indicated as $(a + b)$.

Table 2.4. Method of calculating major population parameters

Parameter (rate or proportion)	Notation	Calculated using
Exposed	$P(F+)$	$(A + B)/N$
Diseased	$P(D+)$	$(A + C)/N$
Diseased and exposed	$P(F+ \text{ and } D+)$	A/N
Diseased in exposed group	$P(D+ / F+)$	$A / (A + B)$
Diseased in nonexposed group	$P(D+ / F-)$	$C / (C + D)$
Exposed in diseased group	$P(F+ / D+)$	$A / (A + C)$
Exposed in nondiseased group	$P(F+ / D-)$	$B / (B + D)$

To clarify the sampling strategy in each of the three analytic study methods, assume the investigator wishes to test if vaccination against selected viruses alters the risk of pneumonia in feedlot cattle. Although it is rare that the structure of the population to be sampled is known, a numerical example is given in Table 2.5. Although based on fictitious data, the example demonstrates the information that would be provided by each of the sampling methods, in comparison to the information that would be available if the population structure was known. With a few modifications, the same approaches to sampling could be used if disease was the independent variable and production the dependent variable (e.g., if the intention were to test the hypothesis that the presence of a disease alters the level of production).

2.4.1 Cross-Sectional Sampling

A sample, usually obtained by one of the previous probability sampling methods, is selected from the population, and each member (sampling unit) is classified according to its current status for the factor and the disease. All of the disease rates in the population may be estimated, based on the results of a cross-sectional sample. Thus this method allows the investigator to learn about the population structure, as well as to test the null hypothesis that the factor (vaccination) and disease (pneumonia) are

Table 2.5. Demonstration of the anticipated results of sampling a population using cross-sectional, cohort, and case-control methods

Suppose the factor is vaccination and the disease is pneumonia. Further, assume the population has the following structure:

		Pneumonia <i>D+</i>	No pneumonia <i>D-</i>	Total
Vaccinated	<i>F+</i>	12,000	48,000	60,000
Not vaccinated	<i>F-</i>	18,000	22,000	40,000
		30,000	70,000	100,000

If 1000 animals were sampled from this population using *cross-sectional* methods, the anticipated results, ignoring sampling error, would be:

		<i>D+</i>	<i>D-</i>	<i>p(D+ / F)</i>
Vaccinated	<i>F+</i>	120	480	600 (20%)
Not vaccinated	<i>F-</i>	180	220	400 (45%)
		300	700	1000
<i>p(F+ / D)</i>		(40%)	(69%)	

All the population characteristics including those shown in parentheses may be estimated from these data.

If *cohort* sampling were used with 500 individuals per group the results would be:

		<i>D+</i>	<i>D-</i>	<i>p(D+ / F)</i>
Vaccinated	<i>F+</i>	100	400	500 (20%)
Not vaccinated	<i>F-</i>	225	275	500 (45%)

Only the two characteristics (shown in parentheses) of the population may be estimated from these data.

Finally, if *case-control* sampling were used with 500 individuals per group, the results would be:

		<i>D+</i>	<i>D-</i>
Vaccinated	<i>F+</i>	200	343
Not vaccinated	<i>F-</i>	300	157
		500	500
<i>p(F+ / D)</i>		(40%)	(69%)

Again, only the two population characteristics (shown in parentheses) may be estimated from these data.

independent events in the population. However, this method of sampling may be impractical when disease frequency is low, because large sample sizes would be required to obtain a sufficient number of cases. In the example in Table 2.5, 120 vaccinated cattle with pneumonia were observed; whereas 180 would be expected if vaccination and pneumonia were independent events. The expected number is derived by multiplying the first row total by the first column total, and dividing by *n* (i.e., $600 \times 300 / 1000$). This calculation is based on statistical theory regarding probabilities

of independent events and is the basis of the chi-square test, see 5.2. Since there are fewer observed vaccinated animals with pneumonia than expected, it appears that vaccination may protect against pneumonia.

An example of a cross-sectional study is presented in Table 2.6. This northern California study was designed to estimate the frequency of acute bovine pulmonary emphysema and to identify factors associated with this disease (Heron and Suther 1979). A list of all herds in three counties (the sampling frame) was obtained from the California Bureau of Animal Health. Then a stratified random sample was used—each county constituted a separate stratum—and a 10% random sample of herds (the sampling unit and the unit of concern) was selected within each county.

Farm owners were interviewed about their husbandry methods, particularly forage management practices. Based on the results of this study, it appeared that approximately 10% of the farms experienced an outbreak of acute bovine pulmonary emphysema during the 4-year study, and that approximately 35% of farm managers used pasture rotation but did nothing specific to prevent the problem. Approximately 2.5 farms ($24 \times 7/68$) or 3.6% of farms would be expected to use pasture rotation and experience the disease if these were independent events; whereas 7 (10.3%) actually did. This suggested a strong association between pasture rotation with no preventive measures and the occurrence of pulmonary emphysema. Additional data indicated that about 3% of the cattle at risk on the affected farms developed pulmonary emphysema. The case fatality rate was 53.8%.

A cross-sectional design was used in a study of factors influencing morbidity and mortality in feedlot calves (Martin et al. 1982). However,

Table 2.6. Results of a cross-sectional study of the relationship between pasture changes and the occurrence of acute bovine pulmonary emphysema (ABPE) during a four-year period

	Number of herds			$p(D + / F)$
	Affected	Non-affected		
Pasture rotated and no preventive measures taken	7	17	24	(29.2%)
Pasture not rotated or preventive measures taken if pasture rotated	0	44	44	(0.0%)
	$\bar{7}$	$\bar{61}$	$\bar{68}$	
$p(F + / D)$	(100.0%)	(27.9%)		

Source: Heron and Suther 1979, with permission.

Note: The prevalence of pasture rotation with no preventive measures taken was $24/68 = 35.3\%$ of farms.

ABPE occurred during at least one of four years in $7/68 = 10.3\%$ of farms.

ABPE and pasture rotation with no preventive measures taken occurred together in $7/68 = 10.3\%$ of farms.

Other estimates of rates applying to the source population are shown in parentheses.

since no formal sampling was used to select collaborators, it is not known how closely the distribution of various risk factors or the prevalence of disease found in the study might be to population values. Thus, although the associations found in the study may be valid, it is difficult to extrapolate certain results beyond the sample (i.e., beyond the groups of cattle under study).

2.4.2 Cohort Sampling

In cohort sampling, a sample of exposed ($F+$) and a sample of unexposed ($F-$) sampling units are selected and observed for a period of time, and the rate of disease in each sample is used to estimate the corresponding rates of development of disease in the two populations. Usually when cohort sampling is used, one does not gain information about the frequency of the factor or of the disease in the population. Testing whether the rate of disease in the exposed group is equal to the rate in the unexposed group evaluates the null hypothesis that the factor and disease are independent events in the population. In the example in Table 2.5, a sample of 500 vaccinated animals and a comparison cohort of 500 unvaccinated animals were identified and observed for a specified time to determine the respective rates of pneumonia. In this fictitious data, since only 20% of vaccinated animals and 45% of nonvaccinated animals developed pneumonia, it appears vaccination helped prevent the development of pneumonia.

The two cohorts (i.e., the two exposure groups) are only infrequently selected by a formal random sampling process. Usually they are purposively sampled specifically because of their exposure or nonexposure to the factor of interest. As long as the two groups are comparable in other respects, the effect of the exposure factor can still be evaluated. However, the groups should be demonstratively representative of the exposed and unexposed segments of the population if the results are to be extrapolated beyond the sampling units in the study.

An example of the use of cohort sampling is shown in Table 2.7. The

Table 2.7. Results of a cohort study of the relationship between the place of residence and the extent of pulmonary damage in 7-12-year-old dogs

	Pulmonary tract damage		Total	Rate of lesions $p(D+/F)$
	Severe lesions	No severe lesions		
Urban dogs*	224	82	306	(73.3%)
Rural dogs	50	150	200	(25.0%)
	274	232	506	

Source: Reif and Cohen 1970.

*This classification was based on known levels of air pollutants in the area, as well as housing density.

objective was to contrast the rate of pulmonary disease in rural ($F-$) and urban ($F+$) dogs in an attempt to estimate the impact of living in a relatively unpolluted (rural) versus a polluted (urban) environment (Reif and Cohen 1970). No differences were noted in young dogs. However, significant differences were seen in dogs 7–12 years of age; the highest rates being in urban dogs, suggesting a harmful effect of the polluted environment.

2.4.3 Case-Control Sampling

In case-control sampling, samples of diseased ($D+$) and nondiseased ($D-$) individuals are selected, and the proportion of each that has been exposed to the factor of interest is used to estimate the corresponding population proportion. Testing whether these two sample proportions are equal evaluates the null hypothesis that the factor and disease are independent events in the population. In the example in Table 2.5, a group of 500 animals with pneumonia and a sample of 500 animals without pneumonia would be selected, and the proportion vaccinated in each group would be contrasted. If the proportion of cases that were vaccinated (40%) was significantly different than the proportion of controls that were vaccinated (69%), vaccination would be associated with pneumonia. Since the former proportion is smaller, it appears that vaccination protected against the development of pneumonia in this hypothetical example.

Only infrequently are the two groups ($D+$ and $D-$) obtained by a formal random sampling procedure. Usually the cases are obtained from one or more sources and essentially represent all of the available cases from the purposively selected sources. Often, the comparison group consists of all animals not having the disease of interest from the same source, be that a set of clinic or farm records. Sometimes, however, formal sampling is used. In a study of feline urological syndrome, the cases represented all cats with the disease in the clinic records; whereas the controls were obtained by taking a 10% systematic random sample of cats without the urologic syndrome (Willeberg 1975). In another example, the characteristics of herds with reactors to brucellosis were contrasted with those with no reactors. The data were obtained from the records of a diagnostic laboratory. Since a large number of herd records were available, a 10% random sample of herds having reactors and a 6% random sample of herds not having reactors to bovine brucellosis were selected. (These sampling fractions were selected because initial estimates indicated that they would provide the required number of reactor and nonreactor farms.) (S. W. Martin, pers. comm.)

In a study of factors associated with mastitis in dairy cows (Goodhope and Meek 1980), the case herds were the 550 with the highest milk-gel index in the province of Ontario. Each was matched to the closest herd in the same county with the lowest milk-gel index (i.e., the controls). (The latter

selection method helped ensure that the case and control herds were comparable since they were geographically matched.)

An example of case-control sampling is presented in Table 2.8 (Willeberg 1980). Herds with high levels (>5%) of enzootic pneumonia in swine at slaughter (cases) were compared to herds with low levels (<5%) of enzootic pneumonia in their pigs (controls). While a number of characteristics of these herds were contrasted, Table 2.8 demonstrates the association of one factor (herd size) with level of pneumonia. Note that the sampling units are herds, not individual pigs. It is obvious from these data that larger herds (the exposure factor) occur much more frequently among herds with pneumonia problems than in herds with low levels of pneumonia. This suggests a harmful effect of the factor "large herds" on the level of pneumonia.

Table 2.8. Results of a case-control study of the relationship between herd size and pneumonia level in swine herds

Herd size	Level of pneumonia	
	High (>5%)	Low (<5%)
Large (>400 pigs)	67	22
Small (<400 pigs)	49	111
	116	133
$p(F+/D)$	(57.8%)	(16.5%)

Source: Willeberg 1980, with permission.

Note: The unit of concern and of analysis is the herd, not the pig.

2.4.4 Sample Size Considerations

Because of the time and expense required to conduct a valid analytic study, careful consideration should be given to determining the number of animals or sampling units required. The formulas given in Table 2.9 provide a basis for estimating sample sizes when the study is designed to contrast two groups.

EXAMPLE CALCULATIONS Two hypothetical examples will be presented to demonstrate the use of sample size formulas. In the first example, assume that the study is intended to compare the milk production of cows with clinical mastitis to cows not having mastitis (i.e., comparing the means of two quantitative variables). Suppose cows not experiencing clinical mastitis will produce 160 BCM units of milk with a standard deviation of 40 BCM units. (BCM is the breed class average for milk; see 3.6.1.) Further, assume clinical mastitis will reduce milk production by 10% to 144 BCM

Table 2.9. Formulas for calculating the sample size in observational studies or field trials involving two treatments

If the outcome is measured as a proportion use:

$$n = [Z_{\alpha}(2\bar{P}\bar{Q})^{1/2} - Z_{\beta}(P_e Q_e + P_c Q_c)^{1/2}]^2 / (P_e - P_c)^2$$

If the outcome is expressed as a mean use:

$$n = 2[(Z_{\alpha} - Z_{\beta})S / (\bar{X}_e - \bar{X}_c)]^2$$

n = estimated sample size for each of the exposed (cases) and unexposed (control) groups. The above formulas are based on large sample size theory; thus, if $n < 10$, double it, and if $n < 25$ increase n by about 1.5 times.

Z_{α} = value of Z which provides $\alpha/2$ in each tail of normal curve if a two-tailed test is used or α in one tail if a one-tail test is used. If α , the type I error, is 0.05 then the two-tailed Z is 1.96. α specifies the probability of declaring a difference to be statistically significant when no real difference exists in the population.

Z_{β} = value of Z which provides β in the lower tail of normal curve (Z_{β} is negative if $\beta < 0.5$). If β , the type II error, is 0.2, the Z value is -0.84 . β specifies the probability of declaring a difference to be statistically nonsignificant when there is a real difference in the population.

P_e = estimate of response rate in exposed (or case) group

P_c = estimate of response rate in unexposed (or control) group

$$\bar{P} = (P_e + P_c) / 2$$

$$\bar{Q} = 1 - \bar{P}$$

S = estimate of standard deviation common to both exposed (cases) and unexposed (control) groups

\bar{X}_e = estimate of mean of outcome in the exposed (or case) group

\bar{X}_c = estimate of mean of outcome in the unexposed (or control) group

Note: Since P , Q , S , and \bar{X} are estimates of population parameters, they should be written with a caret (^); however, the syntax becomes complicated and thus for clarity the caret is omitted.

units. How many cows are required in a cohort study to be 80% (1 - type II error) certain of detecting a difference as large as this, if it exists? Substitution of the above estimates into the second formula for sample size determinations gives:

$$\begin{aligned} n &= 2[(1.96 + 0.84)40 / (144 - 160)]^2 = 2(112 / -16)^2 \\ &= 2(-7)^2 = 2 \times 49 = 98 \end{aligned}$$

Thus, the investigator should use approximately 100 mastitic and 100 non-mastitic cows for the study.

As a second example, suppose a newly identified organism is present in 40% (P_e) of nasal swabs of feedlot calves with pneumonia, and it is thought to occur in about 15% (P_c) of swabs from feedlot calves without pneumonia. How many calves would have to be examined in a case-control study to be 80% sure of detecting this difference (or greater) if it existed? Note that $\bar{P} = 0.275$ and $\bar{Q} = 0.725$. (This is contrasting the means of two qualitative variables, the means being expressed as rates or proportions.)

$$\begin{aligned}
 n &= \frac{[1.96(2 \times 0.275 \times 0.725)^{1/2} + 0.84(0.4 \times 0.6 + 0.15 \times 0.85)^{1/2}]^2}{(0.4 - 0.15)^2} \\
 &= (1.24 + 0.51)^2 / 0.25^2 \\
 &= 3.06 / 0.063 \\
 &= 49
 \end{aligned}$$

The investigator should plan to include approximately 50 calves with pneumonia (cases) and 50 calves without pneumonia (controls) in the study.

2.4.5 Cost Considerations in Analytic Studies

Under most practical field conditions, it can be shown that case-control studies require the fewest sampling units of all analytic observational studies to evaluate a specified hypothesis (Fleiss 1973). This and other features of study design make case-control studies a popular choice when selecting a study method (see Chapter 6).

In the previous discussions of sampling for hypothesis testing, equal size groups were used (i.e., the $F+$ and $F-$ groups were of equal size in cohort studies and the $D+$ and $D-$ groups were of equal size in case-control studies). If the costs of obtaining study subjects differ between unexposed and exposed, or cases and controls, the study design can be modified to take this feature into consideration. Although straightforward in principal, the formulas are somewhat complex, and the interested reader should consult the appropriate references for details and examples (Mey-drech and Kupper 1978; Pike and Casagrande 1979).

References

- Cannon, R. M., and R. T. Roe. 1982. *Livestock Disease Surveys: A Field Manual For Veterinarians*. Canberra: Australian Bureau of Animal Health.
- Cochran, W. G. 1977. *Sampling Techniques*. Toronto, Canada: John Wiley & Sons.
- Fleiss, J. L. 1973. *Statistical Methods for Rates and Proportions*. Toronto, Canada: John Wiley & Sons.
- Goodhope, R. G., and A. H. Meek. 1980. Factors associated with mastitis in Ontario dairy herds: A case-control study. *Can. J. Comp. Med.* 44:351-57.
- Heron, B. R., and D. E. Suther. 1979. A retrospective investigation and a random sample survey of acute bovine pulmonary emphysema in Northern California. *Bov. Pract.* 14:2-8.
- Leech, F. B., and K. C. Sellers. 1979. *Statistical Epidemiology in Veterinary Science*. New York, N. Y.: Macmillan Co.
- Levy, P. S., and S. Lemeshow. 1980. *Sampling for Health Professionals*. Belmont, Calif.: Wadsworth.
- Martin, S. W., A. H. Meek, D. G. Davis, J. A. Johnson, and R. A. Curtis. 1982. Factors associated with mortality and treatment costs in feedlot calves: The Bruce County beef project, years 1978, 1979, 1980. *Can. J. Comp. Med.* 46:341-49.

- Meydrech, E. F., and L. L. Kupper. 1978. Cost considerations and sample size requirements in cohort and case-control studies. *Am. J. Epidemiol.* 107:201-5.
- Pike, M. C., and J. T. Casagrande. 1979. Re: Cost considerations and sample size requirements in cohort and case-control studies. *Am. J. Epidemiol.* 110:100-2.
- Reif, J. S., and D. Cohen. 1970. Canine pulmonary disease. II. Retrospective radiographic analysis of pulmonary disease in rural and urban dogs. *Arch. Environ. Health* 20:684-89.
- Schwabe, C. W., H. P. Riemann, and C. E. Franti. 1977. *Epidemiology in Veterinary Practice*. Philadelphia, Penn.: Lea & Febiger.
- Snedecor, G. W., and W. G. Cochran. 1980. *Statistical Methods*. 7th ed. Ames: Iowa State Univ. Press.
- Willeberg, P. 1975. A case-control study of some fundamental determinants in the epidemiology of the feline urological syndrome. *Nord. Vet. Med.* 27:1-14.
- _____. 1980. The analysis and interpretation of epidemiological data. *Proc. 2nd Int. Symp. Vet. Epidemiol. Econ.*, May 1979, Canberra, Australia.