Albanian j. agric. sci. ISSN: 2218-2020, (2011), Nr.3 /Vol. 10 © Agricultural University of Tirana

PROBABILITY SAMPLING DESIGNS FOR VETERINARY EPIDEMIOLOGY

XHELIL KOLECI[1*], CHRIS L. S. CORYN[2], KRISTIN A. HOBSON[3], RRUZHDI KEÇI[1]

[1]Asociated Professor of Veterinary Infectious Disease& Epidemiology, Faculty of Veterinary Medicine, Department of Veterinary Public Health, *Agricultural University of Tirana, Tirana ALBANIA*

[2]Assistant Professor of Evaluation, Measurement, and Research (EMR), Director of the Interdisciplinary Ph.D. in Evaluation (IDPE) *Western Michigan University. Michigan USA*

[3]Student of the Interdisciplinary Ph.D. in Evaluation Department, *Western Michigan University. Michigan USA*

[1]Lector of Veterinary Epidemiology, Faculty of Veterinary Medicine, Department of Veterinary Public Health, *Agricultural University of Tirana, Tirana ALBANIA*

[*] Corresponding author; email: xhelil.koleci@ubt.edu.al

Abstract

The objective of sampling is to estimate population parameters, such as incidence or prevalence, from information contained in a sample. In this paper, the authors describe sources of error in sampling; basic probability sampling designs, including simple random sampling, stratified sampling, systematic sampling, and cluster sampling; estimating a population size if unknown; and factors influencing sample size determination for epidemiological studies in veterinary medicine.

Keywords: probability sampling; simple random *sampling; stratified sampling; systematic sampling; cluster sampling; epidemiology; veterinary medicine*

1. Introduction

The objective of sampling is to estimate population parameters, such as incidence or prevalence, from information contained in a sample [6]. That is, to make inferences about a population from information contained in a sample selected from that population [8]. In most instances, such inferences are in the form of an estimate of a population parameter, such as a mean, total, or proportion, with a bound on the error of estimation. Each observation taken from a population contains a certain amount of information about the population parameter or parameters of interest [5] Therefore, the central feature of nearly all sampling designs is determining the necessary sample size or quantity of information in a sample pertinent to a population parameter [11].

If θ is the parameter of interest and $\hat{\theta}$ is an estimator of θ, a bound on the error of estimation is necessary to specify that θ and $\hat{\theta}$ differ in absolute value by less than some value B (e.g., $\pm 5\%$). Stated in notational form, the bound on the error of estimation is expressed as

$$\text{Error of estimation} = \left|\theta - \hat{\theta}\right| < B$$

with a probability $(1-\alpha)$ that specifies the fraction of times in repeated sampling that the error of estimation is less than B. This condition is expressed as

$$P[\text{Error of estimation } < B] = 1 - \alpha$$

Traditionally, B is set to two standard deviations of the estimator, and therefore $(1-\alpha)$ will be approximately $.95$ for standard normal, bell-shaped distributions. Sample means and proportions, for example, exhibit bell-shaped distributions for reasonably large sample sizes, even when the parent population is skewed. Once a specified bound on the error of estimation with its associated probability $(1-\alpha)$ is specified different sampling designs (i.e., methods of selecting a sample) can be compared to determine which yields the desired level of precision most efficiently.

The essential nomenclature related to sampling includes elements, populations, samples, sampling units, and frames, though there are many more. An *element* is an object on which a measurement is taken. A *population* is a collection of elements to which an inference is made from a sample. A *sample* is a collection of sampling units drawn from a frame or frames. *Sampling units* are nonoverlapping collections of elements from the population that cover the entire population. A *frame* is a list of sampling units.

2. Sources of Error

Samples can be afflicted with many types of errors. Some arise because only a sample from a population is intended for measurement and because, even for the sampled elements, data may be incomplete or incorrect. These errors can be divided into two general categories [11]: *errors of nonobservation*, where the sampled elements make up only part of the target population, and *errors of observation*, where recorded data deviate from the truth. Errors of nonobservation can be attributed to sampling, coverage, or nonresponse. Errors of observation can be attributed to the data gatherer, respondent, instrument, or, more generally, method of data collection.

3. Errors of Nonobservation

Generally, the data observed in a sample do not precisely reflect the data in the population from which that sample was selected, even if the sampling and measuring are done with extreme care and accuracy. This deviation between and estimate from an ideal sample and the true population value is the *sampling error* that is produced simply because this is a sample and not a census. Sampling error can be measured theoretically and estimated from sample data for probability samples.

In almost all samples, the sampling frame does not correspond perfectly with the target population, leading to errors of *coverage*. Coverage errors sometime occur because a sampling frame that perfectly matches the target population is unavailable. *Undercoverage* occurs when eligible units are missing from the sampling frame. Elements that are not members of the target population but are members of the sampling frame (i.e., ineligible units) are referred to as *overcoverage*. Therefore, coverage bias can be described as a function of the proportion of the target population not covered by the sampling frame and the difference between the covered and noncovered population.

Coverage error is a property of a frame and a target population on a specific statistical estimate. It exists before a sample is taken. The bias of coverage for a mean can be expressed as

$$\bar{Y}_C - \bar{Y} = \frac{U}{N}(\bar{Y}_C - \bar{Y}_U)$$

where $\bar{Y}$ is the mean of the entire target population, $\bar{Y}_C$ is the mean of the population on the sampling frame, $\bar{Y}_U$ is the mean of the target population not on the sampling frame, N is the total members of the target population, and U is the total number of eligible members not on the sampling frame (i.e., not covered elements).

Perhaps the most serious of all nonobservational errors is *nonresponse*. This is a particularly difficult and important problem when information is collected directly from people (e.g., through some form of interview). Nonresponse arises in one of three ways: inability to contact the sampled elements (e.g., person, household), inability of the person responding to answer the question of interest, or refusal to answer. For a mean, nonresponse bias can be expressed as

$$\bar{y}_r - \bar{y}_s = \frac{m_s}{n_s}(\bar{y}_r - \bar{y}_m)$$

where, $\bar{y}_r$ is the mean of respondents within the sth sample, $\bar{y}_s$ is the mean of the entire specific sample selected, $\bar{y}_m$ is the mean of nonrespondents within the sth sample, n_s is the total number of sample members in the sth sample, and m_s is the total number of nonrespondents in the sth sample.

4. Errors of Observation

Errors of observation can be classified as due to the interviewer, the respondent, the measurement instrument, or the method of data collection.

Interviewers can have a direct and dramatic effect on responses to questions. Reading a question with inappropriate emphasis or intonation can force a response in one direction or another. *Respondents* differ greatly in motivation to answer questions correctly and in ability to do so. Each respondent must understand the entire question and be clear about the options for the answer. Inaccurate responses are often caused by errors of definition in questions; that is, characteristics of the *measurement instrument*. In any measurement question, the unit of measurement must be clearly defined, whether it be units on a tape measure or number of glasses of water. A *method of data collection*, such as direct observations of certain variables on crops in sections of fields in order to produce estimates of crop yields or self-administered questionnaires (e.g., where bias can be introduced when those who respond differ from the target population), among many others, can introduce a variety of errors (errors of recording in interviews, poorly trained interviewers who deviate from a prescribed protocol).

5. Basic Probability Sampling Designs

The classical formulation of a statistical estimation problem requires that randomness be built into the sampling design so that properties of estimators can be assessed probabilistically. Here, and throughout, the type of sampling discussed is *sampling without replacement*. In sampling without replacement, a particular element can appear only once in a given sample. Moreover, many important features of sampling designs and methods, such as distinctions between finite and infinite populations, for example, are beyond the scope of this paper.

6. Simple Random Sampling

A *simple random sample* is a sample of n elements from a population of N in which each of the $\binom{N}{n}$ possible samples of n elements has the same probability of selection, namely $\frac{1}{\binom{N}{n}}$. In simple random sampling, the probability of any element being selected is equal to $\frac{n}{N}$, the ratio of the sample size to the population size.

The number of observations needed to estimate a population mean, μ, with a bound on the error of estimation of magnitude B is found by setting two standard deviations (i.e., approximately a 95% confidence interval) of the estimator, $\bar{y}$, equal to B and solving this expression for n. That is,

$$2\sqrt{V(\bar{y})} = B$$

with the estimated variance of $\bar{y}$ given by

$$\hat{V}(\bar{y}) = \frac{s^2}{n}\left(\frac{N-n}{N}\right)$$

Also,

$$V(\bar{y}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N}\right)$$

The required sample size can be found by solving for n

$$2\sqrt{V(\bar{y})} = 2\sqrt{\frac{\sigma^2}{n}\left(\frac{N-n}{N}\right)} = B$$

where

$$n = \frac{N\sigma^2}{(N-1)D + \sigma^2}$$

If N is large, as it usually is, $(N-1)$ can be replaced with N in the denominator. In the equation

$$D = \frac{B^2}{4}$$

Solving for n often presents a problem because the population variance, σ^2, is unknown. Because a sample variance, s^2, is often available from prior studies, an approximate sample size can be by replacing σ^2 with s^2 or estimating a value of σ^2 when little prior information is available. Because the range is often approximately equal to four standard deviations, one-fourth of the range will provide an approximation of σ^2. If one were interested in the average weekly milk production for dairy livestock in liters, μ, for in a community of small farms, and although no prior data are available to estimate the population variance, it is known that most weekly production lies within a 100 liter range, therefore the estimated value of σ would be

$$\sigma \approx \frac{\text{Range}}{4} = \frac{100}{4} = 25$$

and

$$\sigma^2 = (25)^2 = 625$$

If there were $N = 1{,}000$ small farms in the community and the desired bound on the error of estimation were $B = 3$ liters, then

$$D = \frac{B^2}{4} = \frac{(3)^2}{4} = 2.25$$

and

$$n = \frac{N\sigma^2}{(N-1)D + \sigma^2} = \frac{1{,}000(625)}{999(2.25) + 625} = 217.56$$

Therefore, approximately $n = 218$ observations would be needed to estimate μ, the average monthly milk production in liters for small farms, with a bound on the error of estimation of $B = 3$ liters.

In a like manner, the number of observations necessary to estimate a population total, τ, with a bound on the error of estimation of magnitude B can be found by setting two standard deviations of the estimator equal to B and solving for n, where

$$2\sqrt{V(N\bar{y})} = B$$

where

$$n = \frac{N\sigma^2}{(N-1)D + \sigma^2}$$

with

$$D = \frac{B^2}{4N^2}$$

If, for example, one were interested in estimating the total weight gain in 0 to 4 weeks for $N = 1{,}000$ chicks fed on a new ration with a bound on the error of estimation equal to 1,000 grams and previous studies found a population variance, σ^2, approximately equal to 36.00 (grams)^2, then

$$D = \frac{B^2}{4N^2} = \frac{(1{,}000)^2}{4(1{,}000)^2} = 0.25$$

and

$$n = \frac{N\sigma^2}{(N-1)D + \sigma^2} = \frac{1{,}000(36.00)}{999(0.25) + 36.00} = 125.98$$

Therefore, approximately $n = 126$ chicks would need to be weighed to estimate τ, the total weight gain for $N = 1{,}000$ chicks in 0 to 4 weeks, with a bound on the error of estimation equal to 1,000 grams.

Determining the sample size necessary for estimating a population proportion, p, to within B units is analogous to determining a sample size necessary for estimating μ with a bound on the error of estimation B, given that p can be regarded as the average (μ) of 0 and 1 values for a population. Therefore,

$$n = \frac{N\sigma^2}{(N-1)D + \sigma^2}$$

and

$$D = \frac{B^2}{4}$$

The corresponding sample size needed to estimate p can be found by replacing σ^2 with the quantity pq, where

$$n = \frac{Npq}{(N-1)D + pq}$$

with

$$q = 1 - p$$

However, p is typically unknown. An approximate sample size can be found by replacing p with an estimated value. If no prior information is available, $p = 0.50$ is used to obtain a conservative sample size (one that is likely larger than required). Assuming that no prior information was available to estimate p, and one were interested in determining the proportion of Holstein cattle from a population of $N = 2{,}000$ that have Johne’s disease, with a bound on the error of estimation of magnitude $B = 0.05$, the necessary sample size would be

$$D = \frac{B^2}{4} = \frac{(0.05)^2}{4} = 0.000625$$

with

$$n = \frac{Npq}{(N-1)D + pq}$$
$$= \frac{(2{,}000)(0.5)(0.5)}{(1{,}999)(0.000625) + (0.5)(0.5)}$$
$$= \frac{500}{1.499} = 333.56$$

Therefore, a sample size of $n = 334$ Holstein cattle would be needed to estimate the proportion having Johne's disease with bound on the error of estimation of $B = 0.05$.

7. Stratified Random Sampling

A *stratified random sample* is one obtained by separating the population elements into discrete, nonoverlapping groups, called *strata*, and then selecting a simple random sample from each stratum. The principle reasons for using stratified random sampling rather than simple random sampling are:

Stratification may produce a smaller bound on the error of estimation than would be produced by a simple random sample of the same size. This is particularly true if measurements within strata are homogenous.

The cost per observation may be reduced by stratification of the population elements into convenient groupings.

Estimate of population parameters may be desired for subgroups of the population. These subgroups should then be identifiable strata.

In stratified random sampling

$$L = \text{Number of strata}$$
$$N_i = \text{Number of sampling units in stratum } i$$
$$N = \text{Number of sampling units in the population}$$
$$= N_1 + N_2 + N_3 + \cdots + N_L$$

In stratified random sampling, the number of observations needed to estimate a population mean, μ, where the estimator is denoted $\bar{y}_{st}$, where the st subscript indicates that stratified random sampling was used, of population total, τ, with a bound on the error of estimation of magnitude B, is

$$n = \frac{\sum_{i=1}^{L} N_1^2 \sigma_i^2 / a_i}{N^2 D + \sum_{i=1}^{L} N_i \sigma_i^2}$$

Where a_i is the fraction of observations allocated to stratum i, σ_i^2 is the population variance for stratum i. Approximations of the population variances σ_1^2, σ_2^2, . . . σ_L^2 can be obtained by using sample variances s_1^2, s_2^2, . . . s_L^2 from previous estimates. When estimating μ (or p)

$$D = \frac{B^2}{4}$$

and

$$2\sqrt{V(\bar{y}_{st})} = B$$

When estimating τ

$$D = \frac{B^2}{4N^2}$$

and

$$2\sqrt{V(N\bar{y}_{st})} = B$$

A scientist intends to investigate the average time spent monthly milking livestock, $\bar{y}_{st}$, in small-sized farms (n_1), medium-sized farms (n_2), and large-sized farms (n_3), with an error of estimation of $B = 2$ hours and equal allocation fractions given by $a_1 = \frac{1}{3}$, $a_2 = \frac{1}{3}$, and $a_2 = \frac{1}{3}$. The population of small-sized farms is $N_1 = 155$, medium-sized farms is $N_1 = 62$, and large-sized farms is $N_1 = 93$. A prior study indicates that the stratum variances are approximately $\sigma_1^2 \approx 25$, $\sigma_2^2 \approx 225$, and $\sigma_3^2 \approx 100$. A bound on the error of estimation of 2 hours means

$$2\sqrt{V(\bar{y}_{st})} = 2$$

and, therefore, $D = 1$. Given the population of small-sized farms is $N_1 = 155$, medium-sized farms is $N_1 = 62$, and large-sized farms is $N_1 = 93$, then

$$\sum_{i=1}^{3}\frac{N_i^2\sigma_i^2}{a_i}=\frac{N_1^2\sigma_1^2}{a_1}+\frac{N_2^2\sigma_2^2}{a_2}+\frac{N_3^2\sigma_3^2}{a_3}$$
$$=\frac{(155)^2(25)}{\left(\frac{1}{3}\right)}+\frac{(62)^2(225)}{\left(\frac{1}{3}\right)}+\frac{(93)^2(100)}{\left(\frac{1}{3}\right)}$$
$$=(24{,}025)(75)+(3{,}844)(675)+(8{,}649)(300)$$
$$=6{,}991{,}275$$

where

$$\sum_{i=1}^{3}N_i\sigma_i^2=N_1\sigma_1^2+N_2\sigma_2^2+N_3\sigma_3^2$$
$$=(155)(25)+(62)(225)+(93)(100)=27{,}125$$
$$N^2D=(310)^2(1)=96{,}100$$

Then,

$$n=\frac{\sum_{i=1}^{L}N_1^2\sigma_i^2/a_i}{N^2D+\sum_{i=1}^{L}N_i\sigma_i^2}=\frac{6{,}991{,}275}{96{,}100+27{,}125}$$
$$=\frac{6{,}991{,}275}{123{,}225}=56.7$$

Therefore, $n = 57$ with

$$n_1=n(a_1)=57\left(\frac{1}{3}\right)=19$$
$$n_2=n(a_2)=57\left(\frac{1}{3}\right)=19$$
$$n_3=n(a_3)=57\left(\frac{1}{3}\right)=19$$

To calculate the sample size necessary to estimate the total time spent milking per month, τ, with a bound on the error of estimation of $B = 400$ hours, the same procedure would be used except that

$$D=\frac{B^2}{4N^2}$$

and

$$2\sqrt{V(N\bar{y}_{st})}=B$$

would be substituted for

$$D=\frac{B^2}{4}$$

and

$$2\sqrt{V(\bar{y}_{st})}=B$$

yielding $n \approx 105$, with $n_1 = n_2 = n_3 = 35$.

8. Systematic Sampling

A *systematic sample* is a sample in which elements are randomly selected from the first k elements in a frame and every kth element is thereafter called a 1-in-k systematic sample with a random start. Systematic sampling is a useful alternative to simple random sampling for the following reasons:

Systematic sampling is easier to perform in the field and hence is less subject to selection errors by field-workers than are either simple random samples or stratified random samples, especially if a good frame is not available.

Systematic sampling can provide greater information per unit cost than simple random sampling can provide for certain populations with certain patterns in the arrangement of elements.

In general, systematic random involves random selection of one element from the first k elements and then selecting every kth element thereafter. For a systematic sample of n elements from a population of size N, k must be less than or equal to $\frac{N}{n}$ (i.e., $k \leq \frac{N}{n}$).

To determine the number of observations necessary to estimate μ to within B units, the required sample size is found by solving for n where

$$2\sqrt{V(\bar{y}_{sy})}=B$$

Where the sy subscript indicates that the sampling design is systematic. The solution involves both σ^2 and ρ, which must be known (at least approximately) in order to solve for n. Although these parameters can sometimes be estimated, this method is not discussed here as it exceeds the scope of this paper. Instead, the formula for n is the same as for simple random sampling. This formula could give an extra-large sample for ordered populations and too small a sample for periodic populations. If the population is random, the variances of $\bar{y}_{st}$ and $\bar{y}$ are equivalent and the necessary sample size is estimated as

$$n = \frac{N\sigma^2}{(N-1)D + \sigma^2}$$

where

$$D = \frac{B^2}{4}$$

9. Cluster Sampling

Cluster sampling is a less costly alternative to simple or stratified random sampling if the cost of obtaining a frame that lists all population elements is very high or if the cost of obtaining observations increases as the distance separating elements increases. Cluster sampling is an effective design for obtaining a specified amount of information under the following conditions:

A good frame listing all population elements is not available or is very costly to obtain, but a frame listing clusters is easily obtained.

The cost of obtaining observations increases as the distances separating the elements increases.

Clusters typically consist of herds, households, or other units of clustering (e.g., an orange tree forms a cluster of oranges for investigating insect infestations). A farm herd contains a cluster of livestock for estimating proportions of diseased animals. Elements within a cluster are often physically close together and hence tend to have similar characteristics and the measurement on one element within a cluster may be correlated with the measurement on another. In cluster sampling

$$N = \text{Number of clusters in a population}$$

$$n$$
$$= \text{Number of clusters selected in a simple random sample}$$

$$m_i = \text{Number of elements in cluster } i, i = 1, \dots N$$

$$\bar{m} = \frac{1}{n}\sum_{i=1}^{n} m_i = \text{Average cluster size}$$

$$M = \sum_{i=1}^{N} m_i = \text{Number of elements in population}$$

$$\bar{M} = \frac{M}{N} = \text{Average cluster size for population}$$

$$y_i = \text{Total of all observations in } i\text{th cluster}$$

The quantity of information contained in a cluster sample is affected by the number of clusters and the relative cluster size. Assuming that the cluster size (sampling unit) is known or has been selected, the number of clusters for estimating population means and totals, n, to be selected can be estimated from

$$\hat{V}(\bar{y}) = \left(\frac{N-n}{Nn\bar{M}^2}\right) s_r^2$$

where

$$s_r^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y}m_i)^2}{n-1}$$

with the variance $\bar{y}$ of approximated as

$$\hat{V}(\bar{y}) = \left(\frac{N-n}{Nn\bar{M}^2}\right)(\sigma_r^2)$$

where the population quantity σ_r^2 is approximated by s_r^2, and assuming that estimates of σ_r^2 and $\bar{M}$ are available, with

$$2\sqrt{V(\bar{y})} = B$$

and

$$D = \frac{B^2 M^2}{4}$$

Where

$$n = \frac{N\sigma_r^2}{ND + \sigma_r^2}$$

Table 1 Number of Residents and Per-Capita Income

Cluster	Number of Residents (m_i)	Total Annual Income per Cluster (y_i)
1	8	\$96,000
2	12	\$121,000
3	4	\$42,000
4	5	\$65,000
5	6	\$52,000
6	6	\$40,000
7	7	\$75,000
8	5	\$65,000
9	8	\$45,000
10	3	\$50,000
11	2	\$85,000
12	6	\$43,000
13	5	\$54,000
14	10	\$49,000
15	9	\$53,000
16	3	\$50,000
17	6	\$32,000
18	5	\$22,000
19	5	\$45,000
20	4	\$37,000
21	6	\$51,000
22	8	\$30,000
23	7	\$39,000
24	3	\$47,000
25	8	\$41,000
	$\sum_{i=1}^{25} m_i = 151$	$\sum_{i=1}^{25} y_i = \$1{,}329{,}000$

Supposing that the data in Table 1 represent a preliminary sample of agricultural incomes in a region (in United States dollars) and a researcher was interested in estimating the average per-capita annual agricultural income, μ, with a bound on the error of estimation of $B = \$500$, where $N = 415$, the estimate of s_r^2 is

$$s_r^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y}m_i)^2}{n-1} = (25{,}189)^2$$

The quantity $\bar{M}$ can be estimated by $\bar{m} = 6.04$ from Table 1, and D is

$$D = \frac{B^2 M^2}{4} = \frac{(500)^2(6.04)^2}{4} = (62{,}5000)(6.04)^2$$

where

$$n = \frac{N\sigma_r^2}{ND + \sigma_r^2} = \frac{415(25{,}189)^2}{415(6.04)^2(62{,}500) + (25{,}189)^2} = 166.58$$

Therefore, $n = 167$ clusters should be sampled for a bound on the error of estimation of $B = \$500$ from $N = 415$ clusters.

In a like manner, the number of observations necessary to estimate a population total, τ, with a bound on the error of estimation of magnitude B can be found by

$$n = \frac{N\sigma_r^2}{ND + \sigma_r^2}$$

with

$$D = \frac{B^2}{4N^2}$$

Again, using Table 1 as a preliminary sample, a researcher intends to determine how large a sample is necessary to estimate the total annual per-capita income of all regional residents, τ, with a bound on the error of estimation of $B = \$1{,}000{,}000$ where there are $M = 2{,}500$ residents. The estimate of s_r^2 is

$$s_r^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y}m_i)^2}{n-1} = (25{,}189)^2$$

with

$$D = \frac{B^2}{4N^2} = \frac{(1{,}000{,}000)^2}{4(415)^2}$$

$$ND = \frac{(1{,}000{,}000)^2}{4(415)} = 602{,}409{,}000$$

where

$$n = \frac{N\sigma_r^2}{ND + \sigma_r^2} = \frac{415(25{,}189)^2}{602{,}409{,}000 + (25{,}189)^2} = 212.88$$

Which gives $n = 213$ clusters to be sampled to estimate total annual per-capita income with a bound on the error of estimation of $B = \$1{,}000{,}000$.

For a population proportion, p, estimated as $\hat{p}$, using cluster sampling, the proportion of elements in cluster i that possesses the characteristic of interest is given by a_i, which replaces y_i when estimating μ or τ,

$$\hat{p} = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} m_i}$$

where m_i is the number of elements in the ith cluster, with variance

$$\hat{V}(\hat{p}) = \left(\frac{N-n}{Nn\bar{M}^2}\right) s_p^2$$

where

$$s_p^2 = \frac{\sum_{i=1}^{n}(a_i - \hat{p}m_i)^2}{n-1}$$

When an estimation of the population proportion, p, with a bound on the error of estimation of B units is desired

$$2\sqrt{V(\hat{p})} = B$$

where

$$n = \frac{N\sigma_p^2}{ND + \sigma_p^2}$$

with

$$D = \frac{B^2\bar{M}^2}{4}$$

To estimate the necessary sample size for a population proportion, p, using cluster sampling from $N = 415$, with σ_p^2 estimated as $s_p^2 = 0.527$, $\bar{M}$ estimated as $\bar{m} = 6.04$, with a bound of error where $B = 0.04$, for example, D is

$$\frac{B^2\bar{M}^2}{4} = \frac{(0.04)^2(6.04)^2}{4} = 0.0146$$

where

$$n = \frac{N\sigma_p^2}{ND + \sigma_p^2} = \frac{(415)(0.527)}{(415)(0.0146) + 0.527} = 33.20$$

Therefore, $n = 34$ clusters should be sampled from $N = 415$ to obtain a bound on the error of estimation of $B = 0.04$ for a population proportion, p.

10. Estimating an Unknown Population Size

Frequently, the population size is not known and is important to the goals of a study. Samples can be afflicted with many types of errors. The study of the growth, evolution, and maintenance of wildlife populations crucially depends on accurate estimates of population sizes. Several methods are available, including direct sampling, inverse sampling, density and size from quadrant samples, density and size from stocked quadrants, and adaptive sampling. Due to space limitations, only direct and inverse sampling is discussed here. For presentations of other methods see Scheaffer, Mendenhall III, & Ott (2006).

11. Estimation of a Population Using Direct Sampling

Direct sampling can be used to estimate the size of a mobile population. First, a random sample of size t is drawn from the population, tagged, and released. At a later date, a second sample of size n is drawn. Using these data, often called recapture data, the population size, N, can be estimated. Letting s be the number of tagged wildlife in the second sample, the proportion of tagged wildlife is

$$\hat{p} = \frac{s}{n}$$

An estimate of N is given by

$$\hat{N} = \frac{t}{\hat{p}} = \frac{nt}{s}$$

with variance

$$\hat{V}(\hat{N}) = \frac{t^2 n(n-s)}{s^3}$$

Before posting a schedule for an upcoming hunting season, the game commission for a particular country wishes to estimate the size of the deer population. A random sample of $t = 300$ deer are captured, tagged, and released. A second sample of $n = 200$ deer is taken two weeks later. If 62 tagged deer are recaptured in the second sample, $s = 62$, N can be estimated by

$$\hat{N} = \frac{200(300)}{62} = 967.74$$

with a bound on the error of estimation of

$$2\sqrt{\hat{V}\left(\bar{\hat{N}}\right)} = 2\sqrt{\frac{t^2 n(n-s)}{s^3}} = 2\sqrt{\frac{(300)^2(200)(138)}{(62)^3}}$$
$$= 204.18$$

Thus, the game commission estimates that the total number of deer is $N = 968$ with a bound on the error of estimation of $B = 205$.

12. Estimation of a Population Using Inverse Sampling

Inverse sampling is similar to direct sampling, except the second sample size is not fixed. That is, sampling continues until a fixed number of tagged animals are observed. An initial sample of t animals is drawn, tagged, and released. Later, random sampling is conducted until exactly s tagged animals are recaptured. If the sample contains n animals, the proportion of tagged animals in the sample is given by

$$\hat{p} = \frac{s}{n}$$

And *N* is estimated by

$$\hat{N} = \frac{t}{\hat{p}} = \frac{nt}{s}$$

with variance

$$\hat{V}(\hat{N}) = \frac{t^2 n(n-s)}{s^2(s+1)}$$

Authorities of a large wildlife preserve are interested in the total number of birds of a particular species that inhabit the preserve. A random sample of $t = 150$ birds is trapped, tagged, and then released. In the same month, a second sample is drawn until $s = 35$ birds are recaptured. In total, $n = 100$ birds are recaptured in order to locate $s = 35$ tagged ones. The population, N, is estimated by

$$\hat{N} = \frac{100(150)}{35} = 428.57$$

with a bound on the error of estimation of

$$2\sqrt{\hat{V}\left(\bar{\hat{N}}\right)} = 2\sqrt{\frac{t^2 n(n-s)}{s^2(s+1)}} = 2\sqrt{\frac{(150)^2(100)(62)}{(35)^2(36)}}$$
$$= 115.17$$

Therefore, the population of birds is estimated as $N = 429$ with a bound on the error of estimation of $B = 116$.

13. Some Factors Influencing Sample Size Determination

13.1 Type I and Type II Errors

In a scientific method one of the primary tasks is identifying and defining a research question. Without questions, science would cease to exist. Therefore, it is of utmost importance that scientists identify, define, and explicitly state the problem under investigation, including the particular question or questions of interest related to that problem and the specific strategy employed for resolving, partially resolving, or simply studying the problem. What are the questions to be investigated? How will those questions be answered? These, and many others, are the essential concerns that should be identified at the outset of any social inquiry. What is more, the process of identifying and defining research questions, in conjunction with gathering information and resources (e.g., existing knowledge about the phenomenon being investigated), typically serves as a guide to formulating one or more specific research hypotheses.

For the majority of scientists formulating hypotheses is the sine qua non (i.e., something that is essential or necessary) of all inquiry. In essence, a research hypothesis is a deductive guess that states the expected outcome of a study. When formulating hypotheses, the researcher deduces, through a literature review process, experience, or observation, an anticipated result. Research hypotheses can be expressed in numerous ways, but typically are formulated first as a null or nil (literally meaning zero difference or zero relationship) hypothesis, then as either an alternative non-directional hypothesis (two-tailed, two-sided) or an alternative directional hypothesis (one-tailed, one-sided). Alternative, non-directional hypotheses imply that a difference is anticipated, but does not express the direction of that difference. Directional hypotheses, however, state the expected direction of an expected difference. Each of these types of hypotheses are presented and defined in Table 2.

Table 2 Common Types of Hypotheses

Type of Hypothesis	Definition
Null or nil hypothesis	States that no difference is expected
Non-directional hypothesis	States that a difference is expected but does not state the direction of the expected difference
Directional hypothesis	States that a difference is expected and the direction of the expected difference

In general, most social scientists are interested in one of the alternative hypotheses, whether directional or non-directional, not the null or nil hypothesis (though the null or nil nearly always serves as the basis for the majority of statistical tests).

Null or nil hypotheses are implicit in nearly all forms of research, whether explicitly stated or not. And, nearly all statistical tests are tests of the null or nil hypotheses rather than tests of the alternative hypothesis. Such hypotheses can be expressed in numerous ways, and the methods for doing so vary by disciplinary traditions, norms, and standards. Some biostatisticians, for example, refer to these types of tests or hypotheses as tests of equivalence (e.g., Is a new vaccine *equally* effective as an old vaccine?) or superiority (e.g., Is a 500 mg dose *more* effective than a 250 mg dose of a new drug or higher dose of vaccine?). If an epidemiologist, for instance, were interested in determining whether the average adult body temperature of Alpine goats managed in Albania is actually 38.9° Celsius, the epidemiologist might express the research question (i.e., What is the average adult body temperature of Alpine goats in Albania?) in the form of a null and alternative hypothesis. In notational form, where H_0 is the null hypothesis, H_A is the alternative hypothesis (where A represents alternative), sometimes written H_1, and μ is the population mean, this hypothesis would be represented as:

$$H_0 \colon \mu \ = \ 38.9^\circ$$
$$H_A \colon \mu \ \neq \ 38.9^\circ$$

Using the same hypothesis, the epidemiologist might formulate a directional alternative hypothesis rather than a hypothesis simply suggesting that the population mean, μ, does not equal 38.9° Celsius, such as the average adult body temperature of Alpine goats in Albania is less than 38.9° Celsius. This directional hypothesis would be expressed as:

$$H_0 \colon \mu \ = \ 38.9^\circ$$
$$H_\mathrm{A} \colon \mu \ < \ 38.9^\circ$$

Two concepts are important considerations for understanding the practice of null hypothesis significance testing: Type I and Type II errors. A Type I error is the conditional prior probability of rejecting H_0 when it is true, where this probability is typically expressed as alpha (α). Alpha is a prior probability because it is specified before data are collected, and it is a conditional prior probability, p, because H_0 is assumed to be true. This conditional prior probability is usually expressed as

$$\alpha = p(\text{Reject } H_0 | H_0 \text{ true})$$

where | means assuming or given. Both p and α are derived from the same sampling distribution and are interpreted as long-run, relative-frequency probabilities. Unlike α, however, p is not the conditional prior probability of a Type I error (often referred to as a false-positive) because it is estimated for a particular sample result. Conventional levels of α are either .05 or .01 in most of the sciences [2]. Alpha sets the risk of a Type I error rate, akin to a false-positive because the evidence is incorrectly taken to support the hypothesis, for a single hypothesis only (sometimes referred to as a primary or focal outcome). When multiple statistical tests are conducted, there is also a familywise probability of Type I error (sometimes referred to as multiplicity), which is the likelihood of making one or more Type I errors across a set of statistical tests. If each test is conducted at the same level of α, then

$$\alpha_{\text{FWE}} = 1 - (1 - \alpha)^c$$

where c is the number of tests performed, each at a specified α level. In this equation, the term $(1 - \alpha)$ is the probability of not making a Type I error for any individual test, $(1 - \alpha)^c$ is the probability of making no Type I errors across all tests, and the whole expression represents the probability of making at least one Type I error among all tests. So, for example, if 10 statistical tests were performed, each at $\alpha = .05$, the familywise Type I error rate would be

$$\alpha_{\text{FWE}} = 1 - (1 - \alpha)^{10} = .40$$

Thus, the Type I error rate across all 10 statistical tests would be 40%. This result indicates the probability of committing one or more Type I errors, but does not indicate how many errors have been committed or which specific statistical test, or tests, the error occurred in.

There are two basic ways to control familywise Type I error. Either reduce the number of tests (or only test the primary or focal outcome) or lower α to a tolerable rate for each test. The former reduces the total number of tests to those with the greatest substantive meaning, whereas the latter can be determined by a number of methods, including the Bonferroni correction. The Bonferroni correction simply requires dividing the target value of α_{FWE} by the number of tests, and setting the corrected level of statistical significance at α_B where

$$\alpha_B = \frac{\alpha_{\text{FWE}}}{c}$$

If 10 statistical tests were conducted and the tolerable Type I error rate was 5%, then α_B = .05/10 = .005 for each individual test.

Although formal tests of statistical significance largely originated from the works of Fisher (1925) and Neyman and Pearson (1933), statistical power, and the concept of Type II error, however, is largely derived from the work of Cohen [1, 2, 3]. Power is the conditional prior probability of making the correct decision to reject H_0 when it is actually false, where

$$\text{Power} = p(\text{Reject } H_0 | H_0 \text{ false})$$

A Type II error (often referred to as a false-negative) occurs when the sample result leads to the failure to reject H_0 when it is actually false. The probability of a Type II error is usually represented by β, and it is also a conditional prior probability where

$$\beta = p(\text{Fail to reject } H_0 | H_0 \text{ false})$$

Because power and β are complimentary

$$\text{Power} + \beta = 1.00$$

Therefore, whatever increases power decreases the probability of a Type II error and vice versa. Several factors affect statistical power, including α levels, sample size, score reliability, design elements (e.g., within-subject designs, covariates), and the magnitude of an effect, among many others [3, 9]. By lowering α, for example, statistical power is lost, thus reducing the likelihood of a Type I error, which simultaneously increases the probability of a Type II error. Conversely, increasing sample size generally increases power. The relationship between Type I and Type II decision errors arising from statistical hypothesis testing is summarized in Table 3.

Table 3: Accept-Reject Dichotomy and Decisions for Hypotheses

	H_o ***True***	***H_o*** ***falce***
Fail to Reject	Correct decision 1- α	Type II error β
Fail to Accept	Type I error - α	Correct decision 1- β

Table 4: Diagnostic Decisions Relative to a Test Result and True Status

		Test Result (T) ***Positive (+)***	***Negative (-)***
True Status (S)	Disease (+)	a	b
	No Disease (-)	c	d

Null and nil hypothesis significance testing, in most disciplines, has been widely misused and misinterpreted (e.g., a p-value is the probability that a result is due to sampling error, a p-value is the probability that a decision is wrong). The correct interpretation of p-values, for $p < .05$, essentially includes only the following (Kline, 2004):

The odds are less than 1 in 20 of getting a result from a random sample even more extreme than the observed sample when H_0 is true.

Less than 5% of test statistics are further away from the mean of the sampling distribution under H_0 than the one for the observed result.

Assuming H_0 is true and the study is repeated many times, less than 5% of these results will be even more inconsistent with H_0 than the observed result.

13.2 Sensitivity and Specificity

The concepts of *sensitivity* and *specificity* have their origins in diagnostic tests for diseases or other conditions. When a single test is performed, an animal may in fact have the focal disease or the animal may be disease free. The test result may be positive, indicating the presence of disease (T^+), or the test result may be negative (T^-), indicating the absence of the disease as shown in Table 4.

Sensitivity is the conditional probability that a test correctly identifies the presence of a disease or condition when the subject has the disease or condition and is expressed as

$$p(T^+|S^+) = \frac{a}{a+b}$$

Specificity is the conditional probability of a test giving a negative result when the subject does not have a disease or condition and is expressed as

$$p(T^-|S^-) = \frac{d}{c+d}$$

A diagnostic test should have both high sensitivity and specificity otherwise the probability of false positive and false negative diagnoses are substantially increased. In the context of sensitivity and specificity, a false positive occurs when the test result is positive for a subject that is free of a disease or condition. The false positive rate for a diagnostic test is

$$p(S^-|T^+) = \frac{c}{a+c}$$

Table 5 Mastitis Diagnoses in a Sample of 100,000 Jersey Cattle

		Test Result (T) Positive (+)	**Negative (-)**	**Total**
True Status (S)	Disease (+)	475	25	500
	No Disease (-)	4.975	94.525	99.500
	Total	5.450	94.550	100.000

Ideally, the value of c would be 0. However, this is generally impossible in a diagnostic test involving a large population. A false negative occurs when the test result if negative for a subject that has the disease or condition and is

$$p(S^{+}|T^{-}) = \frac{b}{b+d}$$

In a sample of $n = 100{,}000$ Jersey cattle screened for mastitis (see Table 5), for example, where the prevalence rate was low, the sensitivity of the test was $\frac{475}{500} = 0.95$ and specificity was $\frac{94{,}525}{99{,}500} = 0.95$.

Although sensitivity and specificity of the test was high, in large populations where the incidence rate is low the false positive rate increases significantly, and this increase is not strictly a function of sensitivity and specificity but also of the incidence rate in the population. In this case, the false-positive rate was $\frac{4{,}975}{5{,}450} = 0.91$, whereas the false-negative rate was $\frac{25}{94{,}550} = 0.0003$. Therefore, indiscriminately applying a diagnostic test to a large population where the prevalence rate of a disease or condition is very low can be problematic as sensitivity and specificity are related to correct decisions corresponding to Type I and Type II errors.

The basic methodology to account for sensitivity and specificity of a diagnostic device involves taking a sample of size n from a population of N and applying the diagnostic procedure to each of the sampled elements. If k of the sampled elements are screened as positive, then the maximum likelihood estimate $\hat{\pi}$ of the unknown prevalence π is given by

$$\hat{\pi} = \frac{\hat{p} + S_p - 1}{S_e + S_p - 1}$$

where $p = \frac{k}{n} =$ the proportion having positive diagnoses, S_e = the sensitivity of the test, and S_p = the specificity of the test. The standard error of the estimated prevalence is given by

$$\mathrm{SE}(\hat{\pi}) = \frac{\mathrm{SE}(\hat{p})}{\left(S_e + S_p - 1\right)}$$

where the $\mathrm{SE}(\hat{p})$ of $\hat{p}$ would depend on the particular sample design and estimation procedure used. From a random sample of $n = 150$ Alpine goats taken from a population of $N = 2{,}560$ to estimate the prevalence rate of ketosis, using a diagnostic method with sensitivity of 96% and specificity of 89%, 23 were diagnosed positive. Assuming a simple random sample

$$\hat{p} = \frac{23}{150} = 0.153$$

and

$$\begin{aligned}\widehat{\mathrm{SE}}(\hat{p}) &= \left(\frac{N-n}{n}\right)^{1/2}\left(\frac{\hat{p}(1-\hat{p})}{n-1}\right)^{1/2} \\ &= \left(\frac{2{,}560-150}{2{,}560}\right)^{1/2}\left(\frac{0.153 \times 0.847}{149}\right)^{1/2} \\ &= 0.029\end{aligned}$$

where

$$\hat{\pi} = \frac{1.53 + 0.89 - 1}{0.96 + 0.89 - 1} = \frac{0.043}{0.85} = 0.051$$

with

$$\widehat{\mathrm{SE}}(\hat{\pi}) = \frac{0.029}{0.85} = 0.034$$

Considering diagnostic sensitivity and specificity of less than 100%, which produce estimates of test prevalence rather than true prevalence, the sample size necessary for a given bound on the error of estimation, B, using slightly different notation and assuming and infinitely large population, is given by Thrusfield [12].

$$n = \left(\frac{1.96}{d}\right)^2 \times \frac{\{(S_e \times P_{exp}) + (1 - S_p)(1 - P_{exp})\}\{(1 - S_e \times P_{exp}) - (1 - S_p)(1 - P_{exp})\}}{(S_e + S_p - 1)^2}$$

where d represents the desired precision, S_e represents sensitivity, S_p represents specificity, and P_{exp} represents an expected prevelance.

If, for instance, an expected herd or flock prevalence of 20% was to be estimated with a desired absolute precision of $\pm 5\%$ $(d = 0.05)$, with a diagnostic procedure having a sensitivity and specificity of 95% $(S_e = 0.95)$ and 90% $(S_p = 0.95)$, respectively, then

$$n = \left(\frac{1.96}{0.05}\right)^2 \times \frac{\{(0.95 \times 0.20) + (1 - 0.90)(1 - 0.20)\}\{(1 - 0.95 \times 0.20) - (1 - 0.90)(1 - 0.20)\}}{(0.95 + 0.90 - 1)^2} = 419$$

Therefore, approximately $n = 419$ herds or flocks would need to be sampled.

14. Conclusions

Probability sampling designs are at the center of epidemiological studies. The primary aim of sampling is to make inferences from a sample to a target population and, therefore, the information that arises from samples must be representative of the entire population. In this paper basic probability sampling designs, including simple random sampling, stratified sampling, systematic sampling, and cluster sampling, are presented and applications for veterinary medicine are illustrated.

Despite the method of sampling, the distribution of values in any sample will differ from the distribution in sample chosen by chance alone. The number of sampled elements sampling affect the results of all studies, and larger samples are much more likely to reflect the characteristic of interest in the target population. Even under ideal conditions, sampling is prone of several types of errors, including errors of observation and nonobservation and errors of estimation. In addition, direct and inverse sampling methods were presented when population sizes are unknown and important factors influencing sample size determination, including Type I and Type II errors, formulation of hypotheses (e.g., directional, nondirectional, superiority, equivalence), and sensitivity and specificity of diagnostic tests, were presented. Finally, systematic application of the equations and formulas, under consideration of the focal research question or hypothesis, presented in this paper can help mitigate sampling errors as well be used to determine the most efficient sampling method for a particular purpose.

15. References

1. Cohen, J: **Statistical power analysis for the behavioral sciences**. New York, NY: Academic Press; 1969.
2. Cohen, J: **Some historical remarks on the Baconian conception of probability**. *Journal of the History of Ideas*1980, **41**(2), 219-231.

3. Cohen, J. **Statistical power analysis for the behavioral sciences** (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum; 1988.

4. Fisher, R. **Statistical methods for research workers. Edinburgh**, UK: Lover & Boyd; 1925.

5. Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R: **Survey methodology**. New York, NY: Wiley; 2004.

6. Kish, L. **Survey sampling.** New York, NY: Wiley; 1965.

7. Kline, R. B. **Beyond significance testing: Reforming data analysis methods in behavioral research**. Washington, DC: American Psychological Association; 2004.

8. Levy, P. S., & Lemeshow, S. **Sampling of populations: Methods and applications** (4th ed.). New York, NY: Wiley; 2008.

9. Lipsey, M. W., & Hurley, S. M. **Design sensitivity: Statistical power for applied experimental research**. In L. Bickman & D. J. Rog (Eds.), **The Sage handbook of applied social research methods** (2nd ed.) Thousand Oaks, CA: Sage pp. 44-76: 2009.

10. Neyman, J., & Pearson, E. S. **On the problem of the most efficient tests of statistical hypotheses**. Philosophical Transactions of the Royal Society of London, Series A, **231**. 289-337;1933.

11. Scheaffer, R. L., Mendenhall III, W., & Ott, R. L. **Elementary survey sampling** (6th ed.). Belmont, CA: Thompson; 2006.

12. Thrusfield, M. **Veterinary epidemiology** (3rd ed.). Oxford, UK: Blackwell; 2007.