Sample size considerations

Compiled by Dr Joachim Otte (Animal Health Service, FAO)

At an early stage of study design the question of how large a sample one needs must be considered. Clearly, one wants to avoid making the sample so small that the estimate is too inaccurate to be useful. Equally, one wants to avoid a sample that is too large so that the estimate is more accurate than is required, thus wasting valuable resources. Accurate determination of the sample size required for a study can be quite complicated and most complex studies will require the assistance of a statistician. Experience shows that the theoretical calculations provide background information rather than sample sizes for practical use as other than purely statistical considerations are likely to determine sample size.

For less complex studies based on simple random sampling one of the following formulae should provide suitable estimates of the required size of the sample. The formula to use will depend on whether a categorical (discrete, qualitative) or numerical (continuous, quantitative) variable is being measured.

Sample sizes for estimating proportions in large populations

To determine the sample size necessary to estimate a population proportion, for example the prevalence of a given disease in a population or the proportion exposed to a suspected risk factor, the investigator must provide an educated guess on the anticipated proportion of diseased or exposed, P, and must specify how close to the true prevalence the estimate, p, should be.

The average of a series of estimates of the prevalence of a disease, p1, p2, etc. will be almost exactly equal to the true prevalence, P, and the distribution of the estimates will tend to normality. Thus, based on the values from the normal distribution (z-values) 68 percent of the estimates will differ from the true prevalence by less than the quantity *(PQ/n) (where Q=100-P and n=size of the sample), called the standard error of the estimated prevalence (SE). Similarly, 95 percent of the estimates would differ from the true value by less than 1.96 times the standard error and 99 percent of the estimates would be expected to lie within 2.58 standard errors of the true value. Therefore, we might, for example, say that we would like to be 95 percent sure that our estimate lies within 1 percent of the true prevalence, P, i.e. we are requesting a 95 percent confidence of

achieving a precision of *1 percent.

By transformation of the above formula for the standard error of a proportion to yield n, and by substituting d / z for SE we receive the following formula for the calculation of sample sizes:

n = P × (100 - P) × z2/d2

where
P is the anticipated prevalence
d is the desired precision
z is the appropriate value from the normal distribution for the desired confidence

level of confidence

90 percent	95 percent	99 percent
z = 1.645	z = 1.960	z = 2.576

Thus, suppose the available evidence suggests that approximately 40 percent of the cow population will have antibodies to leptospira and the investigator wishes to be 95 percent sure that the estimate will lie be within 4 percent of the true prevalence, sample size would be:

n = 40 × (100 - 40) × 1.962/42 = 576.24

The resulting value should be rounded up at least to the nearest integer as some samples are likely not to yield useable results.

If the sample size is a large proportion of the population, say greater than 10 percent, then sample size can be readjusted to n' using the following formula:

$$n' = \frac{n}{1 + n/N}$$

where N is the total size of the population.

Usually one can make a rough estimate of what the highest expected prevalence of the disease might be in the population of interest. The appropriate sample size then is likely to be too big. If P is likely to lie between 35 and 65 percent, the advance estimate can be quite rough since the product PQ varies little for P lying between these limits. If, however, P is near 0 or 100 percent, accurate determination of n requires a close guess about P. If we do not have the slightest idea what prevalence to expect, one can use the sample size corresponding to the least favourable case, a prevalence of 50 percent, which has the largest sampling variation.

Sample sizes for various levels of precision and expected prevalence

Expected prevalence	**95 percent confidence**		
	* 10 percent	* 5 percent	* 1 percent
10 percent	35	138	3.457
20 percent	61	246	6.147
30 percent	81	323	8.067
40 percent	92	369	9.220
50 percent	96	384	9.064

Sample sizes needed to estimate a population mean

In analogy to what has been said for the calculation of the sample size needed to obtain an estimate of the prevalence of a disease at a predetermined level of precision, the investigator needs to supply an estimate of the standard deviation, s, or variance, s2, of the variable of interest in the target population and specify how close to the true mean the sample estimate, i.e. the tolerated difference, d, should be. In the absence of previous estimates, s can roughly be estimated from a knowledge of the highest and lowest values in the population and a rough idea of the shape of the distribution. If h = highest - lowest, then s = 0.29 h for a uniform (*) distribution, s = 0.25 h for a symmetrical distribution shaped like an isosceles triangle (D), and s = 0.21 h for a skew distribution shaped like a right triangle (*).

The laws of probability allow us to be 95 percent sure that the true population mean lies within the interval:

Sample mean ± 1.96 × standard error of the sample mean.

In other words, we can be 95 percent sure that the difference between the sample mean and the true mean is not greater than 1.96 × SE. For a simple random sample, the standard error of the sample mean is:

SE = s/*n

where: s = Standard deviation
n = Sample size

We thus have to solve the following equation for n:

z × s/*n = d

n = (z2 × s2)/d2

Thus, if we wanted to estimate the average weight of mature cows in a given region, and want to be 95 percent sure that our estimate is within ± 5 kg of the true mean (i.e. d = 5) and available evidence suggests that the standard deviation is 20 kg, the sample size required, assuming we will have to sample less than 10 percent of the population, would be:

n = (1.962 × 202)/52

n = 61

Thus, 61 cows would have to be weighed in order to obtain an estimate of the mean cow weight with a 95 percent chance of the true mean being within the interval of the estimate ± 5 kg. (For a 99 percent probability, 1.96 should be replaced by 2.58. To be 99 percent sure, 105 cows would be needed.)

If the calculated sample size is a large proportion of the population, say greater than 10 percent, then sample size can be readjusted to n' using the formula given in the preceding section.

In using the above formulae, it is assumed that the sampling unit is the same as the unit of

concern. When using cluster or multistage sampling, an upward adjustment of sample size may be required to obtain the desired precision of the estimate. The size of the adjustment required depends on the design effect, D, which is in turn determined by the within cluster correlation coefficient and the number of sampling units per cluster.

D = 1 + (b - 1) roh

where:
b is the average number of animals per cluster
roh is the intra-cluster correlation coefficient

Analysis of survey results based on cluster sampling have shown that for most animal diseases the intra-cluster correlation coefficient does not exceed 0.2 and as can be seen from the following table 27 clusters of ten animals each should ensure a prevalence estimate with a 95 percent confidence of achieving a precision of 10 percent at an expected prevalence of 50 percent.

The effect of intra-cluster correlation and cluster size on sample size:

roh	Animals per cluster	D	Animals required	Clusters required
..05	10	1.45	140	14
..05	20	1.95	188	10
..05	30	2.45	236	8
..10	10	1.9	183	19
..10	20	2.9	279	14
..10	30	3.9	375	13
..20	10	2.8	267	27
..20	20	4.8	461	23
..20	30	6.8	653	22