

# **Part IV**

## **Appendices**

**Appendix A: Glossary**

**Appendix B: Statistical Formulae**

**Appendix C: Computer Programs**

**Appendix D: Example Data Collection Forms**

**Appendix E: Contents of CD and Floppy Disks**



# Appendix A

## Glossary of Epidemiological Terms

(Courtesy of Dr Ian Gardner, University of California, Davis)

**Accuracy** - the degree to which a measurement, or an estimate based on measurements, represents the true value of the attribute that is being measured.

**Agent** - a factor such as a microorganism or chemical substance whose presence or excessive presence is necessary for the occurrence of a disease.

**Analytical study** - a hypothesis testing method of investigating the association between a given disease, health state, or other outcome variable, and possible causative factors.

**Benefit-Cost Ratio** - the ratio of the net present values (usually monetary values) of measurable benefits to costs. Used to determine the economic feasibility or probability of success of a time-bounded program.

**Bias** - any effect at any stage of an investigation tending to produce results that depart systematically from the true values i.e. a systematic error.

**Bias (Response bias)** - a systematic error due to differences in characteristics between those who volunteer to participate in a study and those who do not.

**Bias (Selection bias)** - error due to systematic differences in characteristics between those animals or herds which are selected for study and those which are not.

**Categorical Data** - qualitative data which can be allocated to specific groups. May be nominal (ie. named) or ordinal (ie. ordered) or dichotomous (ie. presence/absence).

**Chi-Square Test** - a method of testing to determine whether two or more series of proportions or frequencies are significantly different from one another or whether a single series of proportions differs significantly from an expected distribution. Pearson's Chi-square is used for unmatched data and McNemar's Chi-square for matched data. See definition of association for further explanation.

**Clustering** - a closely grouped series of events or cases of a disease in relation to time or place or both. The term is normally used to describe aggregation of relatively uncommon events or diseases.

**Confidence Limits** - an interval whose end points can be calculated from observational data and has a specified probability of containing the parameter of interest.

**Confounding** - a situation in which the effects of two factors are not separated. The distortion of the apparent effect of an exposure or risk factor brought about by association with other factors that can influence the outcome.

**Confounding Factor** - a confounding factor or variable is one which is distributed non-randomly with respect to the independent (exposure) variable and is associated with the dependent (outcome) variable being studied. The association with the dependent variable is usually established from results of previous studies.

**Contingency Table** - a tabular cross-classification of data such that subcategories of one characteristic are indicated horizontally (in rows) and subcategories of another characteristic are indicated vertically (in columns), and the number of units in each cell is indicated. The simplest contingency table is the fourfold or 2 x 2 table, but a contingency table may include several dimensions of classification.

**Continuous Data** - quantitative data with a potentially infinite number of possible values along a continuum.

**Cost Benefit Analysis** - methods of identifying the losses and gains in monetary terms of the effects of a disease that are incurred by society as a whole.

**Cross-Sectional Study** - (syn: prevalence study) - a study carried out on a representative sample of a population that examines the relationship between a disease or other health-related characteristic and other variables of interest as they exist in a defined population at one particular time.

**Crude Rate** - a rate which applies to a total population irrespective of the attributes of that population (cf. specific rate).

**Data** - facts of any kind. Data are plural, datum is singular.

**Data Base** - a systemized collection of information, commonly on electronic media about a specific subject such as animal disease.

**Denominator** - the population at risk in the calculation of a rate or ratio. See also Numerator

**Dependent Variable** - (syn:outcome/response variable) a variable or factor, the value of which depends on or is hypothesized to depend on the effect of other [causal] variable(s) in the study.

**Endemic Disease** - the constant presence of a disease or infectious agent within a given geographic area or population group. It also implies a prevalence which is usual in the area or in the population.

**Epidemic** - the occurrence in a population or region of cases of disease clearly in excess of normal expectancy - this is frequently taken as more than two standard deviations greater than the mean occurrence.

**Epidemic curve** - a histogram in which the X-axis represents the time of occurrence of disease cases and the Y-axis represents the frequency of disease cases. It is a useful tool to determine the epidemiology of disease occurrence in an outbreak investigation.

**Epidemic, Propagating** - an outbreak or series of outbreaks resulting from animal to animal spread.

**Epidemiology** - the study of the distribution and determinants of health related states and events in populations. It is a term now in common usage for studies in animal populations although epizootiology is still occasionally used.

**Epidemiology, Descriptive** - study of the occurrence of disease or other health related characteristics in populations. Implies general observation rather than analysis.

**Error, Sampling** - after testing a sample from a large population, the mean or any other statistic calculated from the sample will have a different value from the true value if the whole population was measured. The difference between the value for the whole population and its estimate calculated from the sample is called the sampling error.

**Error, Systematic** - that due to factors other than chance, such as faulty measuring instruments.

**False Negative** - when the result of an individual test is negative but the disease or condition is present.

**False Positive** - when the result of an individual test is positive but the disease or condition is not present.

**Frequency** - a count, or number of occurrences, of an event in a specified population and time period.

**Frequency Distribution** - any arrangement of numerical data obtained by measuring a parameter in a population.

**Histogram** - frequency distribution plotted in the form of rectangles whose bases are equal to the class width and whose areas are proportional to the absolute or relative frequencies.

**Hypotheses** - a proposition that can be tested by facts that are known or can be obtained. The assertion that an association between two, or more variables or a difference between 2 or more groups, exists in the larger population of interest.

**Incidence** - the number of new cases of disease or other condition which occur in a specified population during a given period. Mathematically, 2 types of incidence rate can be distinguished. These are incidence density rates and cumulative incidence.

**Incubation Period** - the interval of time between invasion by an infectious agent or contact with a chemical and the appearance of symptoms of the disease or condition in question.

**Independent Variable** - the characteristic being observed or measured that is hypothesized to influence an event. An independent variable is not influenced by the event or manifestation but may cause it or contribute to its variation.

**Index Case** - the first diagnosed case of an outbreak in a herd or other defined group.

**Infectivity** - the ability of an agent to enter, survive and multiply in the host. Epidemiologically, it is measured as the % of the individuals exposed to an agent who become infected.

**Inference** - the process of passing from observations to generalizations.

**Latent Infection** - persistence of an infectious agent within the host without symptoms of disease.

**Linear Regression** - statistical method used to study the relationship between independent and dependent variables when the dependent variable consists of continuous data.

**Longitudinal Study** - a study conducted over a defined period of time which may be either retrospective or prospective. See also Case Control and Cohort Study.

**Mean-Arithmetic** - a measure of central tendency computed by adding all the individual values together and dividing by the number in the group.

**Median** - the median is the middle value of a set of observations arranged in order of magnitude.

**Mode** - the mode is the most frequently occurring value in a set of observations. A given set of observations can have more than one mode. (see also Bimodal Distribution).

**Monitoring** - the performance and analysis of routine measurements aimed at the early detection of changes in the prevalence or incidence of disease, health, or alteration in a production parameter.

**Multistage Sampling** - a term applied to the selection of a sample in two or more stages. eg, selecting a sample of herds and then a sample of livestock within those herds.

**Nominal Data** - a type of data in which there are limited categories but no order, such as breed and eye color.

**Normal** - within the usual range of variation in a given population or population group; or frequently occurring in a given population or group.

**Normal Distribution** - a continuous symmetrical frequency distribution where both tails extend to infinity, the arithmetic mean, mode and median are identical. Graphically it is a bell shaped curve and its steepness or shape is completely determined by the mean and variance.

**Null Hypothesis** - the hypothesis that two variables have no association at all, or two or more population distributions do not differ from each other.

**Numerator** - the upper portion of a fraction used to calculate a rate or ratio.

**Observational Study** - an epidemiological study where nature is allowed to take its course while changes or differences in one characteristic are studied in relation to changes or differences in other(s) without intervention of the investigator (e.g. descriptive, cross-sectional case-control, cohort).

**Occurrence** - a statement indicating the presence of disease without signifying the frequency. This definition describes the use of the word in international animal disease reports.

**Ordinal data** - a type of data in which there are limited categories with an inherent ranking from lowest to highest (such as severity of disease).

**Outbreak** - the occurrence of disease in a herd or any other identifiable group of animals. For practical purposes, the term is synonymous with epidemic.

**Outliers** - observations differing so widely from the rest of the data as to lead one to suspect that a gross error in recording may have been committed, or suggesting that these values came from a different population.

**Pandemic** - an epidemic occurring over a very wide area, involving many countries and usually affecting a large proportion of the population.

**Parameter** - a summary descriptive characteristic of a population (cf statistic - which is a sample-based measure).

**Pathogenicity** - the ability of an organism to produce disease. Epidemiologically, it is measured as the % of infected individuals who develop clinical disease.

**Power** - probability of finding a difference between two or more groups given that a difference exists. Power = 1-Beta = 1-Probability of a type II error.

**Precision** - the quality of being sharply defined or stated. Refers to the ability of a test or measuring device to give consistent results when applied repeatedly. Sometimes also called repeatability.

**Predictive Value** - in screening or diagnostic tests, the predictive value of a positive test is the proportion of test positive animals that have the disease. The predictive value of a negative test is the probability that an animal with a negative test does not have the disease. The predictive value of a test is determined by the sensitivity and specificity of the test, and by the prevalence of the condition at the time the test is used.

**Prevalence** - the proportion of cases of a disease or other condition present in a population without any distinction between old and new cases. When used without qualification the term usually refers to the number of cases as a proportion of the population at risk at a specified point in time (point prevalence).

$$\text{Prevalence} = \frac{\text{No. cases at specific point in time}}{\text{Population at risk at same point in time}}$$

**Prevalence study** - see cross-sectional study

**Primary Case** - the individual that introduces disease into a herd, flock, or other group under study. Not necessarily the first diagnosed case in that herd. See index case.

**Proportion** - a fraction where the numerator is a subset of the denominator.

**Prospective Study** - see Cohort Study.

**Qualitative data** - that which possess specific qualities such as breed, gender, or color. See nominal data.

**Random** - governed by chance.

**Randomization** - allocation of individuals to groups by chance. Within the limits of chance variation, randomization should make control and experimental groups similar at the start of an investigation and ensure that personal judgement and prejudices of the investigator do not influence allocation. Note that random allocation follows a predetermined plan often devised with the aid of a table of random numbers or by an electronic random number generator.

**Random Sample** - a sample of a population assembled so that each member of the population has an equal and non-zero opportunity to be selected.

**Random Sampling** - procedure for selecting individuals from a population so that each has an equal chance of being selected in the sample.

**Rate** - an expression of the change in one quantity per unit time.

It is a ratio whose essential characteristic is that time is an element of the denominator and in which there is a distinct relationship between numerator and denominator. See also ratio and proportion.

**Ratio** - the expression of the relationship between a numerator and denominator where the two are separate and distinct quantities, i.e the numerator is not included in the denominator.

**Relative Risk** - the ratio of the disease incidence in individuals exposed to a hypothesized factor to the incidence in individuals not exposed; a measure of association commonly used in cohort studies. See also odds ratio.

	Diseased	Not diseased
Exposed	a	b
Unexposed	c	d

The Relative Risk is  $[a/(a+b)] / [c/(c+d)]$

**Repeatability** - the ability of a test to give consistent results in repeated tests. See precision.

**Response Rate** - the number of completed or returned survey instruments (questionnaires, interview etc.) divided by the total number of individuals selected for study.

**Retrospective Study** - a study that collects and utilizes historical data. A case-control study is retrospective because it looks back from the point of known effects to determine causative factors.

**Robust** - a statistical test is described as robust if the inferences hold true even when assumptions inherent in the tests are violated.

**Sampling** - the process of selecting a number of representative subjects from all the subjects in a particular group. Conclusions based on sample results may be attributed only to the population sampled. See also random sample and selection bias.

**Screening** - implies subjecting a population or sample of a population to a diagnostic test or procedure, with the objective of detecting disease. Tests used for this purpose are usually cheap, easily performed, sensitive but often not very specific.

**Sensitivity** - is the proportion of truly diseased animals in the screened population which are identified as diseased by the test. It is a measure of the probability that a diseased individual will be correctly identified by the test.

**Sentinel Herds** - herds that are reasonably representative of the population as a whole and which are tested at regular intervals for infectious disease to determine whether and to what extent the diseases are occurring in the population.

**Seroepidemiology** - epidemiological studies based on an examination of sera taken from the population or a sample of the population.

**Significance, Level of** - also known as alpha ( $\alpha$ ) or type I error rate. The probability of saying a difference exists when none does.

**Spatial distribution** - the relationship of disease events to location of individual animals or clusters of animals.

**Specificity** - is the proportion of truly non-diseased animals correctly identified by the test. Like sensitivity, specificity is a conditional probability.

**Specific Rate** - expresses the frequency of a characteristic per unit size of a specific population.

**Sporadic** - a disease occurring irregularly and generally infrequently and without any apparent underlying pattern.

**Standard Deviation** - a measure of dispersion or variation. Equal to the positive square root of the variance. The mean indicates where the values for a group are centered. The standard deviation is a measure of how widely values are dispersed around the mean in the population.

**Standard Error** - measure of the variability of a sample statistic that specifically relates an observed mean to the true mean of the population.

**Statistic** - a summary value calculated from a sample of observations usually to estimate a population parameter.

**Statistical Significance** - statistical methods allow an estimate to be made of the probability of the observed degree of association between independent and dependent variables being exceeded under a null hypothesis. From this estimate the statistical "significance" of a result can be stated. Usually the level of statistical significance is stated by the "P" value or probability value. See also Significance, Level of.

**Statistics** - the science and art of dealing with variation in data through collection, classification, and appropriate analysis.

**Stratified Sample** - involves dividing the population into distinct subgroups according to some important characteristic, eg herd size, and selecting a random sample out of each subgroup.

**Surveillance** - a system or measurement technique to gain knowledge about a population by collection, analysis, and interpretation of data with a view to the early

detection of cases of disease or changes in the health status of the population. The goal of surveillance is directed action in the treatment or prevention of the condition.

**Survey** - an investigation in which information is systematically collected.

**Systematic Sample** - the procedure of selecting according to some simple systematic rule, such as every 5th cow in the herd as they enter the milking parlor. A systematic sample may lead to errors that invalidate generalizations.

**Temporal Distribution** - the relationship of disease events to time.

**Trend** - a long-time movement in an ordered series (e.g. a time series). An essential feature is that the movement, whilst possibly irregular in the short term, shows movement consistently in the same direction over a long term.

**Type I Error** - an error which occurs when using data from a sample that demonstrates a statistically significant association when no such association is present in the population. Equals the level of significance or alpha.

**Type II Error** - an error that occurs from failure to demonstrate a statistically significant association when one exists in a population. Equals Beta. The power of a study equals 1-Beta.

**Validity** - the extent to which a study or test measures what it sets out to measure.

**Variable** - see Dependent variable, Independent variable.

**Variance** - the variance of a set of observations is the sum of squares of the deviation of each observation from the arithmetic mean of the observations, divided by one less than the number of observations.

**Virulence** - it is the degree of pathogenicity and indicates the potential severity of the disease produced by an agent in a given host. Epidemiologically, it is measured as the % of individuals with disease who become seriously ill or die. Sometimes, the case-fatality rate is considered an indicator for the virulence of disease.



# Appendix B

## Statistical Formulae

### Definition of symbols

The symbols used are based on the terminology of Levy and Lemeshow (1991) and Yamane (1967).

$M$  = number of clusters or PSUs in the population (villages).

$m$  = number of clusters sampled

$N$  = number of SSUs in the population (animals).

$n$  = number of SSUs sampled

$N_i$  = number of SSUs in the  $i$ th cluster

$n_i$  = number of SSUs sampled in the  $i$ th cluster

$\bar{n}$  = average number of SSUs sampled per cluster, which is equal to  $n_i$  if a constant number of SSUs is used

$\Pi$  = proportion of population with the characteristic of interest

$\pi_i$  = proportion of the  $i$ th cluster with the characteristic of interest

$s_{wtb}^2$  = sample variance between clusters

$s_{wti}^2$  = sample variance within clusters

$x_{ij}$  = the observation of the  $j$ th SSU in the  $i$ th cluster, coded as 1 for individuals with the characteristic of interest, and 0 for those without

$x_i$  = the sum of the  $j$  values in cluster  $i$ .

$V(\hat{\Pi})$  = variance of the estimated proportion

# Prevalence Surveys

## PPS Sampling

### Prevalence Estimate

The estimate of the proportion of the population with the characteristic of interest (for example, the proportion of animals with protective titres to FMD) (Yamane, 1967) is:

$$\hat{\Pi} = \frac{\sum_{i=1}^m \sum_{j=1}^{n_i} x_{ij}}{m \bar{n}} \quad (1)$$

The sampling scheme produces a self-weighting sample, which means that each SSU has an equal probability of being selected. An unbiased estimate of the population proportion is therefore simply the proportion of positive SSUs in the sample, or the total number of positive SSUs in the sample divided by the total sample size.

### Variance Estimate

An unbiased estimator of the variance of the estimate is given by:

$$V(\hat{\Pi}) = \frac{1}{m} \frac{1}{m-1} \sum_{i=1}^m (\hat{\pi}_i - \hat{\Pi})^2 \quad (2)$$

where  $\hat{\pi}_i$  is defined as:

$$\hat{\pi}_i = \frac{1}{\bar{n}} \sum_{j=1}^{\bar{n}} x_{ij} = \frac{x_i}{\bar{n}} \quad (3)$$

### Sample Size Calculation

The usual method of calculating the optimal value of  $m$  and  $\bar{n}$  is based on the use of a cost function (e.g. Levy and Lemeshow, 1991 p262). The choice of cost function depends on the nature of the survey work, and can be quite complex. For instance, the curve produced may not be continuous. This would occur if, at some point, the number of animals per village exceeded that which could be examined in one day. After this point, extra travel or accommodation costs may be incurred. The nature of travel costs also depend on the way in which field sites (clusters) are visited: they may be visited one at a time, returning to a central base each time, or two or more may be visited on each trip. Despite these complications, a simple cost function will usually be adequate, taking the general form:

$$C = C_0 + C_1 m + C_2 m \bar{n} \quad (4)$$

where  $C$  = total costs,  $C_0$  = fixed costs,  $C_1$  = cost per PSU, and  $C_2$  = cost per SSU. The value of  $\bar{n}$  which will minimise the variance subject to the constraint of the cost function can be found by (Yamane, 1967):

$$\bar{n} = \sqrt{\frac{C_1 \frac{S_{wti}^2}{N}}{C_2 \left( S_{wtb}^2 - \left( \frac{S_{wti}^2}{N} \right) \right)}} \tag{5}$$

If the number of animals sampled per village is much less than the village population, this simplifies to:

$$\bar{n} = \sqrt{\frac{C_1 \frac{S_{wti}^2}{N}}{C_2 S_{wtb}^2}} \tag{6}$$

where the sample estimate of the within-cluster variances is:

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 = \frac{\bar{n} \hat{\pi}_i (1 - \hat{\pi}_i)}{\bar{n} - 1} \tag{7}$$

the mean of the within-cluster variances is:

$$\bar{s}_i^2 = \frac{1}{m} \sum_{i=1}^m s_i^2 = s_{wti}^2 \tag{8}$$

which is equal to the weighted mean of the within-cluster variances, as the weights  $n_i$  are all equal. The sample estimate of the variation between clusters is:

$$s_{wtb}^2 = \frac{1}{m - 1} \sum_{i=1}^m (\hat{\pi}_i - \hat{\Pi})^2 \tag{9}$$

Formula 5 can be used to calculate the optimal number of animals (resulting in minimum total variance) to be sampled per village for specified per village and per animal costs, and variance estimates. When a fixed budget is available for the survey, the optimal number of villages required for a survey of a given cost can be calculated by substituting this value of  $\bar{n}$  into the cost function 4. For ongoing surveillance, ensuring that survey estimates achieve the necessary level of precision is more important. The number of villages that yield an estimate of a given accuracy may be calculated in the following way.

The confidence interval for the prevalence estimate is given by (Levy and Lemeshow, 1991 p53):

$$CI = \hat{\Pi} \pm z_{(1-\frac{\alpha}{2})} \sqrt{V(\hat{\Pi})} \tag{10}$$

where  $z_{(1-\frac{\alpha}{2})}$  is the standard normal deviate (1.96 for a 95% confidence interval). The variance Formula 2 can be rewritten as:

$$V(\hat{\Pi}) = \frac{s_{wtb}^2}{m} \quad (11)$$

By substituting Formula 9 into Formula 2 and rearranging, we get

$$m = \frac{s_{wtb}^2}{\left(\frac{u}{z_{(1-\frac{\alpha}{2})}}\right)^2} \quad (12)$$

where  $u$  is half the width of the confidence interval. This approach is adequate when the estimated proportion is neither very large or very small. However, if the prevalence is high or low, then using a fixed-width target confidence interval may be inappropriate. An alternative approach involves the use of relative error,  $R$ , defined as:

$$R = \frac{\sqrt{\text{Var}(\hat{\Pi})}}{\hat{\Pi}} = \frac{s_{wtb}}{\sqrt{m} \hat{\Pi}} \quad (13)$$

In order to ensure that the desired relative error is achieved for proportions down to  $\Pi_0$ , the appropriate number of villages to sample using this approach can then be calculated as:

$$m = \frac{s_{wtb}^2}{\Pi_0^2 R^2} \quad (14)$$

### SRS Sampling

PPS sampling provides estimates of relatively low variance and the selection of a fixed number of animals per village makes field work more predictable. However to achieve these benefits, a sampling frame which includes reliable data on village livestock populations is required. When a sampling frame with no such data available, simple random sampling (SRS) must be used at the first stage. In order to achieve a self-weighting sample (in which every animal in the population has the same probability of selection), a fixed proportion of the village population must be sampled at the second stage. This sampling scheme therefore requires simple random sampling with replacement of  $m$  villages from a total of  $M$ , and then simple random sampling (without replacement) of  $n_i$  animals from the  $i$ th village total of  $N_i$ , such that  $n_i/N_i$  (the second-stage sampling fraction, or  $f_2$ ) is constant (or nearly so).

### Prevalence Estimate

As  $N$ , the total number of animals in the population, is unknown, this value must be estimated. Using the ratio-to-size estimate (Cochran, 1977 p303), the estimated proportion is:

$$\hat{\Pi} = \bar{x} = \frac{\frac{M}{m} \sum_{i=1}^m \frac{N_i}{n_i} \sum_{j=1}^{n_i} x_{ij}}{\frac{M}{m} \sum_{i=1}^m N_i} = \frac{\sum_{i=1}^m N_i \bar{x}_i}{\sum_{i=1}^m N_i} \tag{15}$$

This mean per village is equal to the mean per animal when  $f_2$  is constant. Ratio estimates of this nature are subject to some bias, but are required when an estimate of the population size is not available.

**Variance Estimate**

When biased estimators are used, estimates of the mean square error (mse) are preferable to the variance, as they take this bias into account. An estimator of the mse for the above ratio estimate is given by:

$$var(\bar{x}) = mse(\hat{\Pi}) = \frac{1}{\hat{N}^2} \left[ \frac{M^2}{m} \left(1 - \frac{m}{M}\right) \frac{\sum_{i=1}^m N_i^2 (\bar{x}_i - \bar{x})^2}{m-1} + \frac{M}{m} \sum_{i=1}^m \frac{N_i^2 \left(1 - \frac{n_i}{N_i}\right) s_{2i}^2}{n_i} \right] \tag{16}$$

**Sample Size Calculation**

For optimal allocation of number of villages and number of animals based on costs assuming the same cost function as formula 4 above, except that the per-village costs will now include the cost of developing a sampling frame for the village. The optimal average number of animals sampled per village ( $\bar{n}$ ) can be calculated (Cochran, 1977 p314):

$$\bar{n}_{opt} = \frac{S_2}{\sqrt{S_b^2 - \frac{S_2^2}{N}}} \sqrt{\frac{c_1}{c_2}} \tag{17}$$

where the weighted variance among villages per animal, and a slightly biased (upwards) estimate are:

$$S_b^2 = \frac{\sum_{i=1}^M N_i^2 (\bar{X}_i - \bar{\bar{X}})^2}{\bar{N}^2(M-1)} \quad \hat{S}_b^2 = \frac{\frac{M}{m} \sum_{i=1}^m N_i^2 (\bar{x}_i - \bar{\bar{x}})^2}{\bar{N}^2(M-1)} \tag{18}$$

and the weighted mean of within village variances and its unbiased sample estimate are:

$$S_2^2 = \sum_{i=1}^M \frac{N_i}{N} S_{2i}^2 \quad \hat{S}_2^2 = \frac{M}{m} \sum_{i=1}^m \frac{N_i}{N} s_{2i}^2 \tag{19}$$

Calculation of the optimal second-stage sampling fraction ( $f_2$ ) requires an estimate of the average village livestock population:

$$f_2 = \frac{\bar{n}}{N} \quad (20)$$

Calculation of the optimal number of villages to sample is based on the following formula:

$$m_{se}(\hat{\Pi}) = \frac{1}{m} \left( S_b^2 - \frac{S_2^2}{N} + \frac{1}{n} S_2^2 \right) - \frac{S_b^2}{N} \quad (21)$$

To calculate the optimal number of villages based on a fixed width confidence interval (size  $2 \times u$ ),

$$m_{opt} = \frac{S_b^2 - \frac{S_2^2}{N} + \frac{1}{n} S_2^2}{\left( \frac{u}{z_{(1-\frac{\alpha}{2})}} \right)^2 + \frac{S_b^2}{N}} \quad (22)$$

Alternatively, the relative error can be used as in formula 14 above:

$$m_{opt} = \frac{S_b^2 - \frac{S_2^2}{N} + \frac{1}{n} S_2^2}{\Pi_0^2 R^2 + \frac{S_b^2}{N}} \quad (23)$$

## RGCS Sampling

Random geographical coordinate sampling (RGCS), described in Chapter 3, allows the random selection of villages in the absence of a village sampling frame. When using RGCS for the first stage of a two-stage sampling scheme, calculation of an unbiased estimate requires village proportions to be weighted by the number of villages within the selection radius for the randomly selected point.

### Prevalence Estimate

The estimate of the proportion is:

$$\hat{\Pi} = \frac{\sum_{i=1}^m N_i \bar{x}_i w_i}{\sum_{i=1}^m N_i w_i} \quad (24)$$

where  $m$  is the total number of villages and  $w_i$  is the number of villages around the  $i$ th random point. If the proportion of circles which straddle the border of the study region is large (say greater than 10%) it is advisable to modify Equation 24 to take the area of the circles outside the study region into account. This is achieved by replacing  $w_i$  with  $w_i/c_i$ , where  $c_i$  is the proportion of the  $i$ th circle lying within the study region.

**Variance Estimate**

An estimator for the variance of the estimated proportion is

$$var(\hat{\Pi}) = \left( \frac{1}{m(m-1)\hat{N}^2} \right) \left( \frac{A}{\pi r^2} \right)^2 \sum_{i=1}^m N_i \frac{w_i}{c_i} (\pi_i - \hat{\pi})^2 \tag{25}$$

where  $\hat{N}$  is the estimated total number of animals in the population, A is the total area of the study region and r is the selection radius used.  $\hat{N}$  may be estimated as:

$$\hat{N} = \frac{A}{a} \sum_{i=1}^m N_i w_i \tag{26}$$

where a is the total area inside the selection radii around the random coordinates. This is equal to  $(m_t \pi r^2) - a_{\text{external}}$ , where  $m_t$  is the total number of circles used (including those with no villages), r is the selection radius, and  $a_{\text{external}}$  is the sum of the area of the parts of circles lying outside the study area. These values are most easily calculated with a GIS.

The variance given by Formula 25 may be biased if the variability of the quantity of interest is greater in high density areas than in lower density areas. If there is reason to suspect that this is the case, then simple random sampling would be preferable. However, when no sampling frame can be constructed, RGCS may be the only alternative despite this potential bias in the variance.

**Sample Size Calculation**

The optimal second stage sampling fraction is found from:

$$f_2 = \sqrt{\frac{B_2 C_1}{(B_1 - B_2) C_2 \bar{N}'}} \tag{27}$$

where the values and their corresponding estimates are:

$$\begin{aligned} B_1 &= \sum_{i=1}^M N_i^2 \frac{w_i}{c_i} (\hat{\pi}_i - \hat{\pi})^2 & \hat{B}_1 &= \frac{A}{a} \sum_{i=1}^m N_i^2 \frac{w_i}{c_i} (\hat{\pi}_i - \hat{\pi})^2 \\ B_2 &= \sum_{i=1}^M N_i \frac{w_i}{c_i} S_{2i}^2 & \hat{B}_2 &= \frac{A}{a} \sum_{i=1}^m N_i \frac{w_i}{c_i} \hat{S}_{2i}^2 \\ C_1 &= C_u + C_l \bar{N}' & \hat{S}_{2i}^2 &= \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \hat{\pi}_i)^2 \\ \bar{N}' &= \frac{1}{W} \sum_{i=1}^M N_i \frac{w_i}{c_i} & \hat{\bar{N}}' &= \frac{1}{W} \frac{A}{a} \sum_{i=1}^m N_i \frac{c_i}{w_i} \\ W &= \frac{A}{\pi r^2} \end{aligned} \tag{28}$$

The approximate relationship between  $B_1, B_2, \hat{\bar{N}}'$  and  $s_b^2, s_2^2$  and  $\bar{N}$  (using average weighting values) is as follows:

$$\begin{aligned}
\hat{B}_1 &\approx \bar{N}^2 (m-1) \frac{A}{a} \hat{S}_b^2 \left( \frac{w_i}{c_i} \right) \\
\hat{B}_2 &\approx \frac{1}{N} \hat{S}_2^2 \left( \frac{w_i}{c_i} \right) \\
\hat{N}' &\approx \left( \frac{c_i}{w_i} \right) \bar{N}
\end{aligned}
\tag{29}$$

To calculate  $m$ , these equations can be combined with the desired confidence interval half-width,  $u$ , or desired relative error,  $R$ , using:

$$V = \frac{u^2}{z_{1-\frac{\alpha}{2}}^2} \quad \text{or} \quad V = R^2 \Pi_0^2
\tag{30}$$

The number of villages to be sampled is then found from:

$$m = \left( \frac{W}{N^2 V} \right) \left[ B_1 - B_2 + \frac{B_2}{f_2} \right]
\tag{31}$$

### Stratified Sampling

Stratification of the sample will almost always improve the variance of the estimate (Levy and Lemeshow, 1991 p105), as well as providing some practical advantages. Stratification is also important if separate estimates need to be made for the different strata. The OIE recommends stratification for random geographical coordinate sampling for these reasons (OIE, 1990). Ideally strata should be created such that the variation within strata is minimised, and variation between strata is maximised. However, for surveys of this type, very little population data is available. Stratification by geographical area is therefore perhaps the only practical option. It can generally be expected that serological response will be variable between regions, and this stratification would lead to lower variance. Administrative subdivisions are the most readily available geographical areas for stratification.

A conceptually simple approach to stratified sampling is to use proportionate sampling of elements. This means that within each stratum, the sampling fraction is equal to the overall sampling fraction for the whole sample. Each stratum's contribution will be proportional to its size.

Under stratified sampling, each of the strata can be treated as independent samples, and the proportions and standard errors calculated separately. To combine these strata estimates into an overall estimate, each stratum is assigned a weight,  $W_h$ . These weights are generally the proportion of the population contained within the stratum. In this case, an estimate of the proportion of villages in each stratum is used, as the number of livestock in each stratum is usually unknown. Where it is impossible to estimate the number of villages in a stratum, the area of the geographical subdivision could be used to weight strata. Kish (1995 p80) points out that when estimates are used as weights, the formulae used to combine strata no longer strictly hold, and will simply be approximations. In practice, good estimates of the number of villages in each area will often be available.

Under proportionate sampling, a constant sampling fraction is used, so that (Kish, 1995 p80)

$$f = \frac{n_h}{N_h} = \frac{n}{N} \quad (32)$$

where  $f$  is the sampling fraction,  $n_h$  is the number of villages chosen per district,  $N_h$  is the total number of villages per district,  $n$  is the sample size and  $N$  is the population total number of villages. In this case, the stratum weights will be

$$W_h = \frac{n_h}{n} = \frac{N_h}{N} \quad (33)$$

The estimate of the population proportion,  $P_t$ , across strata is then simply the weighted sum of the strata proportions, using the calculated weights:

$$P_t = \sum W_h P_h \quad (34)$$

The estimate of the variance is equal to the sum of the strata variances weighted by the square of the strata weights:

$$\text{var}(P_t) = \sum W_h^2 \text{var}(P_h) \quad (35)$$

Calculation of the final estimates is therefore simply a matter of calculating the strata estimates and combining them with the above formulae. The strata estimates may be derived in any way, but in this case are all calculated based on the two-stage sampling design used to collect them.

When a ratio estimate is being used (as in the formulae above for simple random sampling and random geographic coordinate sampling), and either the number of villages in each stratum is small, or the number of strata is large, this approach may lead to significant bias in the estimates. In this case, the combined ratio estimate  $\hat{\Pi}_c$  is preferable (Cochran, 1977 p320):

$$\hat{\Pi}_c = \frac{\sum_{h=1}^L \hat{X}_h}{\sum_{h=1}^L \hat{N}_h} \quad (36)$$

An estimator of the variance of the combined estimate is:

$$v(\hat{\Pi}_c) = \frac{1}{N^2} \sum_{h=1}^L \frac{1}{m_h(m_h-1)} \sum_{i=1}^{m_h} (d'_{hi} - \bar{d}'_h)^2 \quad (37)$$

where

$$\begin{aligned}
 d'_{hi} &= \frac{N_{hi} \bar{d}_{hi}}{Z_{hi}} \\
 \bar{d}'_h &= \frac{1}{m_h} \sum_{i=1}^{m_h} d_{hi} \\
 \bar{d}_{hi} &= \bar{x}_{hi} - \hat{\Pi}_c
 \end{aligned} \tag{38}$$

and the selection probabilities are:

$$\begin{aligned}
 Z_{hi} &= \frac{1}{M_h} && \text{for simple random sampling} \\
 &= \frac{\pi r^2}{A} \frac{c_i}{w_i} && \text{for random geographic coordinate sampling.}
 \end{aligned} \tag{39}$$

## Apparent Prevalence to True Prevalence

### Prevalence Calculation

Prevalence estimates based on the use of an imperfect test must be corrected to take account of test performance. The formula to convert Apparent Prevalence (AP) to True Prevalence is:

$$\text{True Prevalence} = \frac{\text{AP} + \text{Sp} - 1}{\text{Se} + \text{Sp} - 1} \tag{40}$$

### Confidence Interval

If the prevalence estimate has been calculated using simple random sampling, the confidence interval can be calculated from the variance estimate, given by:

$$\text{var}(\hat{p}) = \frac{p(1-p)}{n(\text{Se} + \text{Sp})^2} \tag{41}$$

The confidence interval is

$$(\hat{p} - (Z_{\frac{\alpha}{2}} \times \sqrt{\text{var}(\hat{p})}), \hat{p} + (Z_{\frac{\alpha}{2}} \times \sqrt{\text{var}(\hat{p})})) \tag{42}$$

# Incidence Rate Surveys

## Capture / Recapture

### Estimate of Total

Seber (1970) uses an unbiased modification of the original formula shown in the text:

$$\hat{N} = \frac{(n_A + 1)(n_B + 1)}{(n_{11} + 1)} - 1 \quad (43)$$

### Confidence Interval

The variance can be calculated as (Seber, 1970)

$$\text{var}(\hat{N}) = \frac{(n_A + 1)(n_B + 1)(n_A - n_{11})(n_B - n_{11})}{(n_{11} + 1)^2(n_{11} + 2)} \quad (44)$$

McCarty *et al* (1993) calculate the 95% confidence interval by assuming a normal distribution and using  $\hat{N} \pm 1.96\sqrt{\text{var}}$ . Yip *et al* (1995) warn that this approach is not reliable for log-linear modelling, and that a likelihood interval (Hook and Regal, 1982) or bootstrap procedure (Buckland, 1984) should be used. In this two-sample example, the variance Formula 44 has been used and a normal distribution assumed.

## Survival Analysis

### Sample Size

The sample size required to detect a difference between two populations may be calculated using the following formula (Lee 1992 p 341). The formula indicates the number of uncensored observations required in each group.

$$n_d = \frac{2\tau(1, \alpha, \beta)}{(\log_e a)^2} \quad (45)$$

Where  $a = \frac{\mu_1}{\mu_2}$  and  $\mu_1$  is the larger of the expected median survival times of the two groups and  $\mu_2$  is the smaller and  $\tau(1, \alpha, \beta)$  is a non-centrality parameter for 2 groups, with alpha and beta being type I and type II errors.

### Log-Rank Test

The formula for the chi-square test used to determine if the survival experience of two populations is different is (Lee 1992)

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \quad (46)$$

### Hazard Ratio

The hazard ratio,  $R$ , is a measure of the relative survival experience of the two groups (Altman, 1991). It is defined as:

$$R = \frac{\left(\frac{O_1}{E_1}\right)}{\left(\frac{O_2}{E_2}\right)} \quad (47)$$

where  $O$  represents the observed number of failures and  $E$  the expected number of failures, and the subscripts represent groups 1 and 2. Altman (1991) also presents an alternative approximation for the log hazard ratio, based on the same variance formula as used in the log-rank test which allows the calculation of approximate confidence intervals for the ratio.

# Freedom From Disease

## Probability Formula

In this discussion, the meaning of the symbols used is as follows:

- p prevalence
- Se sensitivity;
- Sp specificity
- D<sup>+</sup> disease<sup>1</sup> positive animals (true positives);
- D<sup>-</sup> disease negative animals (true negatives)
- T<sup>+</sup> test positive animals (positive reactors);
- T<sup>-</sup> test negative animals (negative reactors)
- P() the probability of an event with the event of interest described in the brackets
- x the number of T<sup>+</sup> in a sample
- y the number of D<sup>+</sup> in a sample
- n sample size
- N population size
- d number of diseased D<sup>+</sup> animals in the population
- $\binom{n}{x}$  the number of ways that x objects can be drawn from n, equal to  $\frac{n!}{x!(n-x)!}$

### Infinite population (or sampling with replacement)

The probability of observing x reactors when testing n animals from an infinite population is given by the binomial distribution modified to take account of test sensitivity and specificity:

$$P(T^+ = x) = \binom{n}{x} [pSe + (1 - p)(1 - Sp)]^x [p(1 - Se) + (1 - p)Sp]^{n-x} \tag{48}$$

### Finite population (sampling without replacement)

To overcome the limitations of other commonly used formulae (the assumption of an infinite population or sampling with replacement), the hypergeometric distribution can be modified for imperfect tests. The number of D<sup>+</sup> in the sample has a hypergeometric distribution. Given y D<sup>+</sup> in the sample, the number of true positives is Bin(y, Se), and the number of false positives is Bin(n-y, 1-Sp). We will have x T<sup>+</sup> if we have j true positives and x-j false positives. By considering the possible values of y and j, we can write down:

$$P(T^+ = x) = \sum_{y=0}^d \frac{\binom{d}{y} \binom{N-d}{n-y}}{\binom{N}{n}} \sum_{j=0}^{\min(x,y)} \binom{y}{j} Se^j (1-Se)^{y-j} \binom{n-y}{x-j} (1-Sp)^{x-j} Sp^{n-x-y+j} \tag{49}$$

---

<sup>1</sup>Disease in this context is defined in its broadest context: possessing the abnormality or state of interest. In surrogate tests for disease, it may mean, for example, the presence of antibodies.

## References

- Altman, D.G. (1991) Analysis of survival times. In: *Practical Statistics for Medical Research*. Chapman & Hall, London, 365-395
- Buckland, S.T., (1984) Monte Carlo confidence intervals. *Biometrics* **40**:811-817
- Cochran, W.G. (1977) *Sampling Techniques*, 3rd ed. John Wiley & Sons, Inc., New York
- Hook, E.B. and Regal, R.R., (1982) Validity of Bernoulli census, log-linear, and truncated binomial models for correcting for underestimates in prevalence studies. *American Journal of Epidemiology* **116**:168-176
- Kish, L. (1995) *Survey Sampling*. John Wiley and Sons, New York
- Lee, E.T. (1992) *Statistical methods for survival data analysis*, 2nd ed. John Wiley & Sons, Inc, New York
- Levy, P.S. and Lemeshow, S. (1991) *Sampling of Populations: Methods and Applications*, 2nd ed. John Wiley & Sons, Inc, New York
- McCarty, D.J., Tull, E.S., Moy, C.S., Kwoh, C.K. and LaPorte, R.E., (1993) Ascertainment corrected rates: Applications of capture-recapture methods. *International Journal of Epidemiology* **22**:559-565
- OIE (1990) *Guide to epidemiological surveillance for rinderpest*. Office International des Épizooties, Paris
- Seber, G.A.F., (1970) The effect of trap response on tag recapture estimates. *Biometrics* **26**:13-22
- Yamane, T. (1967) Cluster Sampling (II):Probability proportional to size. In: *Elementary Sampling Theory*. Prentice-Hall, Englewood Cliffs, N.J., 237-271
- Yip, P.S.F., Bruno, G., Tajima, N., Seber, G.A.F., Buckland, S.T., Cormack, R.M., Unwin, N., Chang, Y.-F., Fienberg, S.E., Junker, B.W., LaPorte, R.E., Libman, I.M. and McCarty, D.J., (1995) Capture-recapture and multiple record systems estimation II: Applications in human diseases. *American Journal of Epidemiology* **142**:1059-1068

# **Appendix C**

## **Computer Programs**

## Random Village

---

**Purpose**

Selection of a random sample from a sampling frame, using simple random sampling (SRS) or probability proportional to size sampling (PPS), optional replacement and stratification.

**Input**

- Data file in dBASE or Paradox format
- File should contain identifier for each element, and optionally, fields for size (population) and stratification.

**Output**

- List of randomly selected elements
- May be printed or saved as a new table

**Dos Version?** No

**Page Reference:** 52

## Random Animal

---

**Purpose**

Simple random sampling of animals from a village livestock sampling frame.

**Input**

- Number of animals owned by each livestock owner

**Output**

- Randomly selected animals identified by ID number of livestock owner, and sequential animal number.

**Dos Version?** Yes

**Page Reference** 60

## **RGCS (Win95)**

---

### **Purpose**

Selection of random coordinates for random geographic coordinate sampling

### **Input**

- Coordinates of a rectangle bounding the study area (Cartesian or Decimal degrees format)
- Number of points to select

### **Output**

- Random coordinates
- May be printed or saved to a new table

**Dos Version?** No

**Page Reference** 70

## **RGCS (ArcView GIS)**

---

### **Purpose**

Selection of random coordinates within one or more irregular polygons for random geographic coordinate sampling.

### **Input**

- Digital map (ArcInfo coverage or ArcView Shapefile format) with polygon showing study area.
- Number of points to select
- Polygons representing the study area
- Selection radius

### **Output**

- Random coordinates, displayed on screen with bounding circles determined by the selection radius
- Coordinates may be printed or saved to database file.
- Map display may be manipulated or superimposed over remotely sensed data to screen selected points.

**Dos Version?** No

**Page Reference** 71

## Prevalence

---

### Purpose

Calculation of sample sizes for two stage prevalence surveys using one of three survey designs (SRS, PPS, RGCS), and calculation of prevalence and other estimates from the results of such surveys.

### Input

#### For analysis of survey data:

- Data file in Paradox or dBASE format with disease status, first stage sampling unit (village) ID.
- Optionally, (depending on survey design) a second file with stratum ID, village population, weights, selection radius, and the size of the study area.

#### For sample size calculation:

- Estimated prevalence
- Within- and between-village variance estimates
- Cost per village and cost per animal
- Total number of villages and average village population
- Confidence level and desired accuracy

### Output

#### Data analysis:

- Estimate of prevalence, with variance and confidence interval.

#### Sample size calculation:

- Optimal (minimum cost) first and second stage sample sizes.

**Dos Version?** No

**Page Reference** 162

## Compare Prevalence

---

### Purpose

Compare prevalence estimates from two surveys to determine if the difference is likely to be real or just due to chance.

### Input

- Prevalence and variance estimates from two surveys.

### Output

- P value: Probability that the two observations came from the same population.

**Dos Version?** No

**Page Reference** 170

## True Prevalence

---

**Purpose**

Calculate the true prevalence based on the apparent prevalence and test performance.

**Input**

- Apparent prevalence
- Test sensitivity
- Test specificity

**Output**

- True prevalence
- Confidence interval (assuming single stage simple random sampling)

**Dos Version?** Yes

**Page Reference** 168

## Survival

---

**Purpose**

Perform survival analysis. Specifically to analyse data from retrospective disease outbreak surveys, create a Kaplan-Meier survival curve and statistics.

**Input**

- Data file in dBASE or Paradox format, with time, censoring and optionally weight and grouping fields.

**Output****Single group analysis:**

- Kaplan-Meier survival curve
- Mean and median survival time

**Two-group analysis**

- Kaplan-Meier survival curve
- Log rank test statistic and P value
- Hazard ratio and confidence interval

**Dos Version?** Yes

**Page Reference** 181

## Survive Size

---

### Purpose

Sample size calculation for survival analysis

### Input

- Mean or median survival times for two groups representing the minimum difference that can be distinguished

### Output

- Number of non-censored observations (villages recalling outbreaks) required for each group.

**Dos Version?** No

**Page Reference** 176

## CapRecap

---

### Purpose

Calculate population total based on two data sources

### Input

- Total number in data source 1
- Total number in data source 2
- Total appearing in both data sources

### Output

- Estimate of total in population with confidence interval

**Dos Version?** Yes

**Page Reference** 187

## FreeCalc

---

### Purpose

Calculate sample sizes for surveys to demonstrate freedom from disease, and analyse the results of such surveys

### Input

- Test sensitivity
- Test specificity
- Population size
- Minimum expected (maximum acceptable) prevalence
- Type I and II error levels
- Additionally, for analysis of results, sample size and number of positive reactors

### Output

- Sample size and cutpoint number of reactors
- Probability that the population is diseased

**Dos Version?** Yes

**Page Reference** 204



# Appendix D

## Example Data Collection Forms

The forms on the following pages can be copied or used as a model for developing your own data collection forms for a survey. In addition to the information shown, all sheets should have a few lines at the top saying where the data came from, who it was collected by, and when.

### **Village Livestock Sampling Frame**

This sheet is used during village interview to record the number of animals kept by each of the livestock owners. If more than one species is being sampled, extra columns can be added. Extra copies of this sheet (without pre-printed numbers) should be available for larger villages. When using them, make sure to number the livestock owners on the extra sheets sequentially, and keep the sheets together.

### **Random Number Table**

This table can be used for any random selection exercise, but in particular for selecting animals from the sampling frame made during the village interview;

### **Specimen Collection Sheet**

The species column is only necessary for multiple species surveys. Extra columns may be added for other relevant information (eg vaccination history).

### **Disease Outbreak Questionnaire**

This form was designed for FMD. Substitute the disease of interest.

### **Disease Ranking and Seasonal Patterns**

Be sure to accurately record a description of the disease.

### **RGCS Village Sheet (Sheet 1) and RGCS Random Point Sheet (Sheet 2)**

Sheets to be used for random geographic coordinate sampling fieldwork.

## Village Livestock Sampling Frame

No	Name of Livestock Owner	Total Animals	Cumulative Total	Selected Animals
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				
21				
22				
23				
24				
25				
26				
27				
28				
29				
30				
31				
32				
33				
34				
35				
36				
37				
38				

## Random Number Table

9537	7654	2531	7467	2873	5885	5154	6419	9346	9458	2281	4520	1241	6730	4263
3014	7669	2948	7241	0139	3841	1369	1123	8300	7790	3632	9154	4698	3874	2423
2682	4082	3359	0932	6215	9668	4282	7428	2833	7014	0217	2737	6768	4218	3007
5531	1283	5400	7610	3466	2697	0649	2159	4803	7655	3325	7537	5885	1465	4746
4534	4703	1566	8974	8989	3953	5752	4976	1253	1041	2678	0067	1001	1802	8224
4202	9222	0395	0882	0406	5696	4204	7995	0571	0744	6751	8284	7202	2610	2531
4783	0798	7713	5203	3246	9008	1017	6802	5738	9416	0092	3831	4662	7819	5152
1515	3328	4102	2777	3867	8974	0632	1175	6051	8063	2795	5037	2319	6941	0285
7824	5298	1243	0754	4284	9480	4027	6284	1251	7275	9796	9015	0199	7321	3200
3894	3231	2288	0103	7834	2159	6589	7655	4435	2457	0141	3600	6792	1631	0840
4495	1477	3933	1570	7080	6521	1885	5664	2691	7577	8866	2425	0383	5134	1282
2495	0365	0326	0856	7851	0801	9001	7861	6828	4483	6681	8913	5735	9767	7244
6941	7266	1482	6315	5838	5539	3608	9895	4136	7294	5075	7471	0057	4551	1275
7136	7584	1352	4940	4637	4448	5390	8329	0559	3921	7029	2652	4622	4366	2786
6602	5200	3213	4913	6662	9579	7025	1113	1206	9229	5973	9585	0994	1648	9597
3346	4427	2525	5519	0821	0334	2335	4005	0598	6894	8161	1447	3213	7990	9132
5327	7977	9909	7696	3362	8331	3798	3732	6549	9457	6097	2249	9890	5228	6924
2541	7991	9425	0987	0809	2695	2051	1145	4111	8633	3193	5735	2601	8008	2604
9611	9655	9767	5203	6374	2752	2562	0175	8457	0393	2300	3658	9471	2385	6007
5322	9436	8575	7562	3770	7711	7100	0856	8138	1847	3270	9227	5393	7474	8566
7959	2467	2482	8581	4816	5323	0199	7210	2602	9477	7211	4004	2738	9695	7642
7906	6113	8081	2517	9752	4073	3221	3255	0388	0730	7586	9013	9009	1631	3952
1374	9257	1451	0624	1662	5929	1230	2935	6900	3504	0815	3387	5632	0377	4424
1676	9319	6404	8020	8916	9174	0284	2252	3169	0590	1531	6276	1788	3408	6972
6970	1559	4110	7432	2041	3362	5336	4365	9501	8548	0159	0352	4491	4694	4804
5850	7679	9254	5612	3905	0924	1378	0962	0437	3103	2957	7646	5019	2527	1399
4712	3274	0387	0697	4663	2449	3002	5661	9899	5543	7188	1043	6954	0520	5805
3291	6142	4611	1300	5324	5192	0015	7741	7972	7192	6577	7169	8827	3935	9888
7277	9996	9284	0611	6375	6807	9284	6975	3175	1465	4700	8996	3251	8478	7923
9425	0618	5866	1284	0362	8875	5458	2846	6681	5532	6480	8909	7075	4222	1831
3045	3952	3590	9404	9828	7222	5711	3926	7353	6153	0426	5545	9608	9806	7823
4299	8225	3096	8302	4524	8587	6188	5714	9020	6674	6780	0167	8418	4586	2754
2207	4564	2702	5504	4287	5653	0294	7690	3897	4751	9238	0857	4756	8867	0935
7750	3178	2451	8603	6500	4976	1476	2884	8548	2806	0380	5326	3127	4905	4731
6009	4643	3594	8319	9547	4857	5677	5734	1317	5770	3484	5591	2051	3796	4675
7711	8280	3680	9546	6147	2663	1095	6521	2602	3125	5871	0333	5523	5951	7422
9115	2208	9888	3651	2995	3651	5409	3153	1912	4784	1442	3188	7233	5272	2297
1634	2060	5774	7820	5607	5813	3150	3583	8092	2846	2552	7785	2049	9719	9730
5092	7923	9073	9726	9775	7783	8331	4648	1630	3745	3901	2776	1808	3408	7362
1041	1523	0736	8295	3543	9323	0040	5601	0440	7831	3570	2664	4956	7887	2088
5022	8169	7826	3863	6097	6440	1104	7124	3058	5921	8873	2708	2044	2776	8838
9198	0531	5469	3493	2502	5640	2531	9095	5617	4837	7192	8672	1628	8392	9365
9246	3728	5474	9748	5657	4377	8841	2910	1538	6470	4421	4721	3605	5547	6820
1925	3806	1808	3684	9405	7201	1973	6606	5327	7402	6204	5216	9511	0145	4407
2225	4105	5575	5354	9190	9667	3896	3610	4398	8622	9613	8722	7660	8141	8922
1507	6559	4651	7610	9162	4502	0623	8353	5306	7346	5421	4992	6490	0868	7323
0525	7467	5629	1470	7150	7088	2736	4571	3323	5504	3615	8199	0720	6842	1583
3757	9743	8240	3837	1403	9785	0110	4526	6744	1897	7339	2223	2982	0299	4867
3934	6211	4903	0863	5501	7117	0980	9984	9837	7574	2885	6252	6631	9876	7689
8185	0935	0549	2719	0349	6359	8011	8187	0842	6450	5905	1492	0645	8788	4341
9698	8154	0394	8064	4653	0565	6530	8610	3923	5696	6513	2257	8723	5929	5121



## Disease Outbreak Questionnaire

Village Name		Date	
District Name		Vill No	

<b>Question 1</b>	
Has there ever been an outbreak of FMD in this village? (✓)	
Yes <input type="checkbox"/>	No <input type="checkbox"/>

**If Yes**

<b>Question 2</b>	
When was the last outbreak of FMD in the village? (Month and year that the first animal got sick.)	
Date	

<b>Question 3</b>	
At the time of the last outbreak, how many cattle and buffalo were there in the village?	
Pigs	
Cattle	
Buffalo	

**If No**

<b>Question 2</b>	
What is the earliest date since which you are sure there has been no FMD in the village?	
Date	

<b>Question 3</b>	
At that time, how many cattle and buffalo were there in the village?	
Pigs	
Cattle	
Buffalo	





# RGCS Random Point Sheet (Sheet 2)

Point ID	Visited	Number of villages
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		
15		
16		
17		
18		
19		
20		
21		
22		
23		
24		
25		
26		
27		
28		
29		
30		
31		
32		
33		
34		

# Appendix E

## Contents of CD

Survey Toolbox Win95 Installation Version

Survey Toolbox Win95 Runtime Version

Survey Toolbox DOS Runtime

Acrobat copy of text

EpiInfo 2000 Win95

EpiInfo 6.04 DOS

Adobe Acrobat (Install and RunTime)

# Index

- active surveillance, 3, 17
- animal health information, 3, 14, 84, 113
  - need for, 14
- animal restraint
  - cattle and buffalo, 87
  - goats, 94
  - horses, 93
  - Pigs, 92
  - sheep, 94
- anticoagulant, 101
- attitude, 112
  
- backing up, 132
- bias, 21, 24
  - measurement, 21
  - selection, 21
- bleeding pole, 89
- blood collection, 96
  - cattle and buffalo, 97
  - caudal vein, 97
  - chickens, 99
  - goat, 98
  - guidelines, 96
  - horses, 98
  - jugular vein, 97
  - pigs, 98
  - sheep, 98
- blood samples, 95
  - transporting, 100
- brainstorming, 217
  
- CapRecap, 187
- capture - recapture, 183
  - analysis, 187
  - background, 183
  - data collection, 186
  - data management, 186
  - data sources, 185
  - matching, 186
- censoring, 121
- census, 18
- centrifuge, 101
- children, 105
- clinical prevalence, 31
- clustering, 199
- Compare Prevalence, 167
- competitions, 216
- computer programs, 4
- computers, 124
- confidence interval, 23
- confidence level, 156
- cooperation, 111
- crush, 87
  
- data, 124, 126
- accuracy, 128
- categorical, 126
- checks, 129
- coding, 128
- completeness, 128
- continuous, 126
- conversion, 140
- dates, 126
- dichotomous, 126
- dictionary, 128
- entry, 131
- error checking, 133
- integers, 126
- linking, 139
- management, 175
- missing, 129
- Nominal, 126
- numbers, 126
- ordinal, 126
- processing, 127
- real numbers, 126
- saving, 132
- storage, 127
- text, 126
- types, 126
- yes/no, 126
- data analysis, 124, 141, 160
- database, 127
- dBASE, 140
- diagnostic tests, 32
- disease
  - impact of, 2
  - measures of, 26
  - patterns of, 2
- disease control, 2
- disease names, 109
- disease priorities, 116
- double entry system, 132
  
- encouraging participation, 106
- Epi Info, 4, 43, 51, 125, 129, 176
  - Analysis, 133
  - counting records, 134
  - creating a table, 130
  - data entry, 132
  - frequency tables, 134
  - histogram, 136
  - means, 135
  - recoding, 139
  - relate, 139
  - scatter, 138
  - select, 137
  - setting up data checks, 131
  - tables, 137
- error

- random, 24
- systematic, 21, 24
- Type I, 195
- Type II, 195
- estimation, 23
- evacuated tubes, 95
  
- faecal sample, 95
- field, 127, 129
- field trips, 218
- field work, 219
- FreeCalc, 196
  - analysis of results, 204
  - formulae, 206
  - Infinite Population Size, 207
  - Maximum Sample Size, 207
  - sample size, 196, 201
- freedom from disease, 190
  - data analysis, 204
  - herd or village, 194
  - Large-area surveys, 199
  - proving, 191
  - sample size, 194, 200
- frequency tables, 134
- fun, 113
  
- games, 216
- geographical information system, 67
- georeference, 73
- GIS, 67
- global positioning system , 67
- GPS, 67
- group discussions, 216
  
- halter, 88, 93
  - nose, 88
  - rope, 88
- hardware, 124
- Hazard Ratio, 182
- histograms, 136
  
- import, 140
- Incidence rate, 28
  - analysis of two data sources, 183
  - village-level, 170
- inference, 20
- information sources, 84
  - collecting specimens, 85
  - examining animals, 85
  - existing records, 84
  - interviews, 84
  
- landmarks, 120
- language, 109
- learning land marks, 214
- lesson planning, 213
- listening, 106
- Logrank test, 182
  
- maximum acceptable prevalence, 192
- measures of disease, 26
- median survival time, 180
- minimum expected prevalence, 192
- monitoring, 3
  
- nose grip, 92
  
- OIE, 3
- outbreak, 118
- outbreak history, 118
  
- passive surveillance, 3, 14
- payment, 112
- physical randomisation, 40
- pipette, 101
- plasma, 100
- population, 18, 195
- population size, 155
- practical activities, 219
- precision, 24, 156
- Prevalence, 26, 155
  - apparent, 34
  - Clinical, 31
  - comparison, 167
  - seroprevalence, 31
  - true, 34, 165
- Prevalence Analysis, 160
  - data inputs, 160
- Prevalence surveys, 150
  - conducting, 152
  - design, 152
  - first-stage sampling, 159
  - PPS, 153
  - probability proportional to size, 153
  - random geographic coordinate sampling, 154
  - RGCS, 154
  - sample size, 154, 156
  - second-stage sampling, 160
  - simple random sampling, 153
  - SRS, 153
- proportional allocation, 158
- proportions, 16
- prospective study, 170
  
- questions, 110
  
- race, 87
- Random Animal, 58, 160
- random error, 24
- random number tables, 42
- random numbers, 40, 41
  - selecting, 41
  - selecting with a computer, 43
- random sampling, 22
- Random Village, 50, 159, 175, 186, 198, 204
- ranking, 116, 217

- Rates, 16
- recoding, 138
- record, 127
- recording data, 86
- relative error, 156
- remote sensing, 72, 73
- replacement, 44
- representative samples, 20
- resolution, 73
- restraint, 87
- retrospective disease outbreak surveys, 171
  - activities, 172
  - complex analysis, 179
  - data analysis, 177
  - interview, 175
  - prerequisites, 171
  - sample size, 173
- retrospective study, 170
- RGCS, 68, 154
  - field procedures, 74
  - for ArcView GIS v.3, 69
  - for Windows 95, 68
- role playing, 218
  
- sample, 18
- sample size, 23, 154, 173
  - minimum cost, 201
- sampling, 38
  - convenience, 38
  - haphazard, 38
  - non-probability, 38
  - PPS, 48
  - probability, 38, 39
  - probability proportional to size, 48, 153
  - purposive, 38
  - random, 39
  - random geographic coordinate, 65
  - RGCS, 65
  - simple random, 153
  - stratified, 45
  - systematic, 44
  - techniques, 40
  - two-stage, 64
  - with replacement, 44
  - without a sampling frame, 65
  - without replacement, 44
- sampling frame, 115, 150
- sampling frames, 49
  - building, 55
  - selecting a sample from, 50
  - sources, 50
- sampling interval, 45
- seasonal differences, 178
- selecting animals, 54
  - identifying selected animals, 60
  - random number table, 56
  - with a computer, 58
- selecting subgroups, 137
  - selecting villages, 52
  - selection radius, 71
  - sensitivity, 33, 194
- Seroprevalence, 31
- serum, 101
- snare, 92
- social status, 108
- software, 4, 125
  - installing, 6
  - Requirements, 5
  - running, 6
- specificity, 33, 194
- specimen collection, 95
- stratification, 45, 158
- student-centred learning, 214
- surveillance, 3
  - active, 3
  - passive, 3
- survey
  - objectives, 111
- survey design, 152
- Survey Toolbox, 4
  - how to use, 4
- surveys, 18
  - costs, 154
  - Incidence rate, 170
  - retrospective disease outbreak, 171
  - to demonstrate freedom from disease, 190
  - which to use, 34
- Survive Size, 173
- syringe, 95
- systematic error, 21
  
- table, 129
  - creation, 129
- tables, 137
- teaching techniques, 214
- Tests, 32
  - combining, 205
  - in parallel, 205
  - in series, 205
- trainers
  - who should be a trainer, 212
- training
  - activities, 213
  - courses, 222
  - skills, 212
- True Prevalence, 165
- twitch, 93
- two-stage sampling, 64, 150
  
- under-reporting, 15
- uninterruptable power supply, 132
- units of interest, 18
- UPS, 132
  
- vacuum tubes, 95
- variance, 155

## 330 Index

- village calendar, 120
- village history, 120
- village interviews, 55, 104
  - activities, 114
  - introduction, 114
  - leaders, 106
  - order, 114
  - organising, 105
  - outputs, 114
  - sampling frame, 115
  - who should attend, 104
  
- warmers, 215
- weighting, 65
- women, 104, 108