Glossary of statistical terms

This glossary of statistical terms may be useful to authors when critically reviewing papers and to our reviewers and associate editors when reviewing knowledge summaries

Evidence-based veterinary medicine relies critically on the scientific validity of research. A component of validity is the statistical design and subsequent analysis of data collected during the study. Correct statistical design reduces bias and improves generalizability, and correct analysis leads to appropriate inferences. Inference is the art and science of making correct decisions based on data.

Absolute benefit increase (ABI)

The absolute arithmetic difference in event rates when describing a positive outcome.

Absolute risk increase (ARI)

When the test treatment harms more animals than the control treatment. See ARR below.

Absolute risk reduction (ARR)

The difference in the event rate between untreated control animals (the CER, control event rate), and the treated animals (the EER, experimental event rate).

ANOVA

ANOVA is the acronym for analysis of variance, which is a method of comparing population means of independent groups of independent subjects. If the associated P value is small, at least one group mean is statistically different from the others. A limitation of ANOVA is that it does not tell you which means are different from the others. Therefore, ANOVA is usually followed by post hoc tests, a series of pairwise t tests that describe exactly.

Association

A link between two or more events, characteristics, or other variables for which there is statistical evidence. An association may be detected because of a direct causal link, an indirect causal link, or may have no causal link.

RCVS Knowledge Registered Office
RCVS Knowledge | First Floor | 10 Queen Street Place | London | EC4R 1BE Correspondence
RCVS Knowledge | 3 Waterhouse Square | 138-142 Holborn | London | EC1N 2SW
T 020 7202 0742 E editor@veterinaryevidence.org **W** www.veterinaryevidence.org

Attributable risk

Additional risk of disease is the exposed group over that in the unexposed group. If mammary cancer occurs in 1 of 3,000 spayed bitches (fictitious numbers) and in 21 of 3,000 intact bitches, the attributable risk leaving bitches entire is 20 in 3,000.

Bias

Any factor in the design or execution of a trial, other than the intended intervention, which might affect the interpretation of the results.

Blinded and blinding

A technique used in clinical trials to prevent bias arising from the participants' knowledge of the intervention. In a study described as blinded the authors were deemed to have taken adequate measures to conceal allocation to study groups from those responsible for assessing animals for entry in the trial (e.g., formal randomisation; sequentially numbered, opaque, sealed envelopes; sealed envelopes from a closed bag; numbered or coded bottles or containers; drugs prepared by the pharmacy; or other descriptions that contain elements convincing of concealment).

Blinded study (may also be called a masked study)

In single blinded studies the animal/owner is unaware of which intervention is used. In double-blinded studies neither the observers nor the animal/owners know which intervention is used. In triple-blinded studies the statistical analysis of the results is also carried out without revealing which intervention was used (e.g. the statistician knows that animals received either treatment A or treatment B, but not what they were).

Boolean search

A means of combining search statements or sets using the logical operators "OR" to expand a search and "AND" to restrict a search to articles that contain two or more specified elements together used in searching databases or the Internet.

Box and whisker plot

Box plots are used to compare two or more groups of data visually. The maximum and minimum values, the median value, the first quartile range Lowest 25% of values), the inter quartile range (IQR) which contains the middle 50% of values and the third quartile range which contains the top 25% of values

Case control study

A study in which animals representing cases of a disease are compared to a matched group of animals without this disease in order to see if they were exposed to the putative cause (and the disease-free animals weren't). This type of study is normally retrospective.

Case report or case study

The report of a single case. Although anecdotal they represent the first step in observational epidemiology (when a new disease occurs someone has to point it out).

Case series

A publication, normally a paper in a journal, in which a series of animals with an outcome of interest are described. No control group are used in the analysis of any data presented. They represent a poor source of evidence in scientific terms.

Central tendency measurements

The middle of a distribution. Described by mean, median, and mode.

Chance

Random variation. Difference between the outcomes from a sample of the population and the true value obtained from looking at the outcomes from the entire population. Statistical methods are used to estimate the probability that chance alone accounts for the differences in outcomes.

Chi-squared test

A test used to compare the proportion of subjects in each of two groups having one of two outcomes (dichotomous outcomes).

Clinical significance (as opposed to statistical significance)

Statistical significance means the likelihood that the difference found between groups could have occurred by chance alone. In most clinical trials, a result is statistically significant if the difference between groups could have occurred by chance alone in less than 1 time in 20. This is expressed as a p value < 0.05. Remember that a trivial difference can have a very low p value if the number of subjects is large enough! Clinical significance has little to do with statistics and is a matter of judgment. It answers the question "Is the difference between groups large enough to be worth achieving?" Studies can be statistically significant yet clinically insignificant but not usually the other way around.

Cohort study

A study in which two groups (cohorts) of animals are identified. One group represents a cohort of animals exposed to a putative cause of an outcome, while the other is a cohort free from this exposure. The cohorts are examined for the outcome of interest in order to test the association of the putative cause with the outcome.

Co-morbidity

The existence of disease other than the disease of interest in animals that are the subject of a study.

Comparison group

A group of animals to which the intervention group is compared. In a trial of a new therapy the ideal comparison groups might be a control group receiving no treatment, and a group receiving an established treatment.

Confidence Interval (CI)

Studies are performed on a sample of the population, not the whole population and so confidence intervals give us some idea of how likely the sample mean represents the population mean. Expressed as the sample mean plus and minus a specified amount they are a measure of the precision of the estimate. The 95% CI is the range of values within which we can be 95% sure that the true value lies for the whole population of animals from whom the study animals were selected. Results from a sample population with a wider range of values will have broader CIs than results from a study with a narrower range of values. Increasing the number of results (animals) within a sample population narrows the CIs. The confidence interval quantifies uncertainty and is derived from the sample mean and the standard error. Note that not all error bars shown on graphs of results represent CIs.

Control event rate (CER)

The proportion of animals in which the outcome of interest is seen (e.g. a disease, an adverse reaction to treatment, etc) in the control group of animals (e.g. animals not receiving the treatment).

Control group

The study animals that did NOT receive the experimental intervention (e.g. therapy). In an ideal study both a positive and negative controls are used (e.g. placebo treated, and an existing well documented treatment of known efficacy).

Cost-benefit analysis

Is an analysis performed by converting effects into the same monetary terms as the costs and comparing them.

Crossover study design

The administration of two or more experimental therapies one after the other to the same group of animals.

Cross-sectional study (prevalence study)

Survey of an entire population for the presence or absence of a disease and/or other variable in every member (or a representative sample) and the potential risk factors at a particular point in time or time interval. Exposure and outcome are determined at the same time.

Determinant

A factor that produces a change in the health or disease status of an animal.

Dispersal measurements

Range (max-min values), standard deviation, percentiles

Dose-response relationship

A situation in which the magnitude of the outcome is related to the amount, duration, or intensity of exposure. The change in the outcome may be an increase or a decrease.

Double-blind

Typically used in randomized controlled trials (RCTs). An experimental method in which both the animals and the investigators do not know and cannot work out which animals are receiving treatment and which placebo.

Effectiveness

A measure of the benefit resulting from an intervention for a particular health problem for a group of animals in normal clinical practice (cf efficacy).

Efficacy

A measure of the benefit resulting from an intervention for a particular health problem in ideal (experimental) conditions.

Event rate

The proportion of animals in a group in whom the event is observed. Thus, if out of 100 animals, the event is observed in 18, the event rate is 0.18. Control Event Rate (CER) and Experimental Event Rate (EER) are used to refer to this in control and experimental groups of animals respectively.

Evidence

Evidence is something that serves as proof to support or refute a hypothesis. The strongest evidence (i.e. the evidence with the greatest validity) is provided by studies with a high power to demonstrate a true difference. The value of evidence may range from weak to strong.

Evidence-based medicine

The conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual animals. The practice of evidence-based medicine means integrating individual clinical expertise with the best available external clinical evidence from systematic research.

External validity

Are the results valid outside the animal population studied? Are results from studies done on one breed valid for another breed?

Gold standard

Accepted reference standard, or diagnostic test for a particular illness.

Histogram

Plot of the distribution of the data by plotting a bar chart of the data. To make the bars, the data must be arbitrarily grouped into sections. The size of the grouping (large or small) can dramatically affect the interpretation of the histogram.

Incidence Rate

Number of new cases of a disease in a specified period of time.

Inferential Statistics

Determines how likely a given result occurred by chance alone. Since we can rarely study an entire population, we study a sample of the population and by inference apply that result to the entire population. All statistics used in veterinary studies are inferential.

Internal validity

Are the results of the study valid for the animal population studied cf external validity?

Inter-observer variability

Variability between observers. Do two or more radiologists give the same reading from the same radiograph?

Intra-observer variability

Variability by the same observer. Does a radiologist give the same reading of a radiograph when viewed on more than one occasion?

Kaplan-Meier curve

Used for estimating probability of surviving a unit of time. Used to develop a survival curve when not all survival times are exactly known.

Kappa statistic

Used to measure concordance agreement among raters. The value ranges from 0 to 1. Concordance is the proportion of all results on which two or more tests or observers agree. The level of agreement is often expressed as the kappa (k) statistic) – the proportion of potential agreement beyond chance exhibited by two or more tests or observers. Scale of interpretation of Kappa Statistic Values (Smith 2006, p142, Veterinary Clinical Epidemiology) is given below:

Kappa Value	Strength of agreement
0.01-0.20	No better than chance
0.21-0.40	Slight
0.41-0.60	Fair
0.61-0.80	Moderate
0.81-0.99	Almost perfect
1.00	Perfect

Level of significance

The probability of incorrectly rejecting the null hypothesis, i.e. saying that there is a difference between two groups when actually there is none. Otherwise known as the probability of Type I error. By convention, the level of significance is often set to a p value of 0.01 or 0.05.

Likelihood ratios

The likelihood that a given test result would be expected in an animal with the target disorder compared to the likelihood that the same result would be expected in a animal without that disorder.

Mean

The arithmetic average in a set of values. The average.

Measurement bias

Being studied can affect the outcome. If owners are asked to record the amount of feed being given to their animals, they are likely to measure out quantities more carefully for example. The methodology can also affect outcome.

Median

For a set of values arranged in order of magnitude, the median is the middle value for odd numbers of values and the average of the two middle values for an even number of values.

Meta-analysis

A methodically prepared overview of published studies. Meta-analyses typically use statistical analysis to summarise the combined results.

Mode

For a set of values, the mode is the value that occurs most often.

Multivariate analysis

An analysis where the effects of many variables are considered. Can be used to identify a subset of variables that significantly contribute to the variation in outcome.

Negative predictive value (NPV)

The percentage of animals with a negative test that do NOT have the disease.

Non-parametric statistical tests

if the data is not normally distributed then non-parametric test should be used. Chi-squared test is an example of a non-parametric test. Equivalent non-parametric tests are shown below.

	Parametric test	**Non-parametric counterpart:**
Correlation	Pearson correlation coefficient	Spearman's correlation coefficient
Groups	Independent-means t-test	Mann-Whitney test

Normal distribution

Many biological parameters are normally distributed, such as height and weight. Some, but not all, statistical analyses are designed to work on data that is normally distributed. If the mean, median, and mode are roughly equal then a data set is probably normally distributed.

Normal distribution tests

Statistical tests include: Statistical Kolmogorov-Smirnov (K-S) test, Shapiro-Wilk test. Graphical representation-Q-Q probability plots, Cumulative frequency (P-P) plots

Null hypothesis

The proposal that no difference exists between groups or that there is no association between risk indicator and outcome variables. If the null hypothesis is true then the findings from the study are the result of chance or random factors. The overall purpose of a typical study is to "reject the null hypothesis.

Number Needed to Harm (NNH)

The number of animals who would need to be treated to cause one bad outcome (typically an adverse effect of a therapy).

Number Needed to Treat (NNT)

The number of animals who need to be treated to prevent one bad outcome. It is the inverse of the ARR.

Odds

A ratio of events to non-events. If the event rate for a disease is 0.1 (10 per cent), its non-event rate is 0.9 and therefore its odds are 1:9, or 0.111. Note that this is not the same expression as the inverse of event rate. It represents the chance of detecting an event in a single individual from the population.

One-tailed test

Is there a difference between groups A and B is a specified direction e.g. Is the mean value of A greater than the mean value of B. One tailed tests need smaller group sizes to achieve a given power.

p value

The measured probability of a finding occurring, i.e. rejecting the null hypothesis, by chance alone given that the null hypothesis is actually true. By convention, a p value < 0.05 is often considered significant. ("There is less than a 5% probability that the finding [null hypothesis rejected] was due to chance alone.")

Parametric statistical tests

The requirements for the use of parametric tests are: the subjects in the groups are randomly selected from the population; measurements are interval data (continuous variables); the data are normally distributed; and, the variances (the spread of the values) in the groups are equal. When any one of these conditions is not met, need to use a non-parametric test. Parametric tests are more powerful than non-parametric tests (i.e. they have a higher probability of rejecting the null hypothesis when the alternative hypothesis is true).

Pearson's correlation

Measures the strength of linear association between continuous values of two variables
Interpreting the correlation (r): -1 = perfect downhill linear relationship (but does not indicate the gradient); 0 = no linear relationship exists; +1 = perfect uphill relationship (but does not indicate the gradient). Above or below 0.6 considered significant. It is a unit-less measure. Correlation does not necessarily mean cause and effect.

Percentile

proportion of values/observations lying within in a specified range e.g. lower quartile = range of values containing lowest 25% of values

Point prevalence frequency

Disease sign point prevalence frequencies are the expected sign frequencies if the disease is encountered.

Population

Every animal that satisfies the inclusion criteria for the study. It can be a group of animals with a defined characteristic (e.g. neutered male cats), or animals in a defined location (e.g. in the UK). It is the denominator in the calculation of a rate.

Positive predictive value (PPV)

The percentage of animals with a positive test result that actually have the disease. [Positive predictive value = true positives / (true positives + false positives)]

Post-test odds

The odds that the animal has the target disorder after the test is carried out.

Post-test probability

The proportion of animals with that particular test result who have the target disorder (post-test odds/[1 + post-test odds]). Use of a nomogram avoids the need to perform any arithmetic.

Power

The probability of detecting an effect in the treatment vs. control group if a difference actually exists. Must also specify the size of the difference. For example, a paper describing a clinical trial with a new mastitis treatment may contain the following statement - "The study had a power of 80% to detect a difference of 10,000 cells per ml in milk between the treatment and control groups." Typical power probabilities are 80% or greater. Power = 1 - ß (see Type II Error)

Pre-test probability

The probability that the animal has the target disorder before the test is carried out. This is normally the prevalence of disease in the population of animals in which the test is used.

Prevalence

The number of animals with a disease, at a given point or period) divided by the population at risk at a particular point or period. Prevalence = Incidence x duration. Point prevalence = at a specific point in time. Period prevalence = during a specific period of time.

Probability

The likelihood that a particular event will occur or the proportion of animals in which a particular characteristic is present.

Randomisation

A process by which animals are selected for a group by random. This should involve a formal randomisation method such as the use of a random number table (strictly speaking this is pseudo-randomisation), a computer program, or selecting identities from a hat. Some investigators confuse the term with arbitrary assignment (e.g. assigning every other case to one of two groups) which can introduce bias.

Randomised controlled trial (RCT)

A true experiment, in which the researcher randomly assigns some animals to at least one intervention and other animals to a placebo, or conventional treatment. Animals are followed over time (Prospectively). A blinded RCT represents the best/strongest form of evidence.

Range

From lowest to highest value in a distribution

Recall bias

The recall of exposures or events may differ in owners of cases and controls. Questions may be asked more times and more intensively in cases compared to controls. Owners of animals with the disease are more likely to carefully consider whether or not an exposure occurred. Can be avoided by the use of a prospective study.

Referral bias (centripetal bias)

Veterinary schools and specialised referral clinics tend not to see the same range of animals presented to general veterinary practices.

Relative risk increase (RRI)

The increase in rates of bad outcomes, comparing experimental animals to control animals in a trial. RRI is also used in assessing the effect of risk factors for disease.

Relative risk or risk ratio (RR)

The event rate in the treatment group divided by the event rate in the control group. Also known as risk ratio. RR is used in randomised trials and cohort studies. When the outcome of interest is rare in the population studied then the Odds Ratio approximates the Relative Risk.

Relative risk reduction (RRR)

The proportional reduction in rates of bad events between the experimental animals and the control animals in a trial, calculated as (EER – CER)/CER and accompanied by a 95% confidence interval (CI).

Regression

Association two continuous variables (can be linear or curvilinear). Sign of value indicates direction of slope (+/- 0-1) with 0 indicating no relationship. Small P values for the regression indicate that the coefficients are not statistically different from zero. If the correlation is moderate to strong can create an equation which allows you to predict the value of one variable from the value of the other variable. Extrapolation outside the range of values should not be done. Coefficient of determination (r2) is the closeness of fit of the data to the regression line. It expresses the amount of variation in the data that is accounted for by the linear relationship between the two variables. The value is between 0 and 1. A value of 1 indicates all the values lie on the regression line.

Research question

The best research question should specify a single measurable outcome, as well as all the conditions and important variables. The question contains the population, the intervention or conditions affecting the study population, and the outcomes.

Scatter plots

Should not be used for inference but for data description. This is because the arrangement of the axis and the choice of scale of the plot data can affect the appearance of the plot. Scatter plots are common to regression and correlation analysis. They are a plot of two matched outcomes and appear as a cloud of points on the graph. By "stretching" one of the axes, it is possible to correctly plot but distort the appearance of the data.

Sample

The animals who satisfied the study's inclusion criteria and who actually entered the study, a subset of the population.

Selection or sampling bias

The sample population chosen is not representative of the population at risk (e.g. animals with advanced disease were compared with healthy non-diseased animals.

Sensitivity analysis

The value(s) of a parameter(s) within a model is (are) varied while the remaining parameter values are kept constant. Changes in the outcome are monitored. This process allows parameters that do, and do not, contribute to the problem solving in modelling.

Sensitivity

The probability of the test finding disease among those who have the disease or the proportion of animals with disease who have a positive test result. Sensitivity = true positives / (true positives + false negatives)

Specificity

The probability of the test finding NO disease among those who do NOT have the disease or the proportion of animals free of a disease who have a negative test. Specificity = true negatives / (true negatives + false positives)

Standard deviation

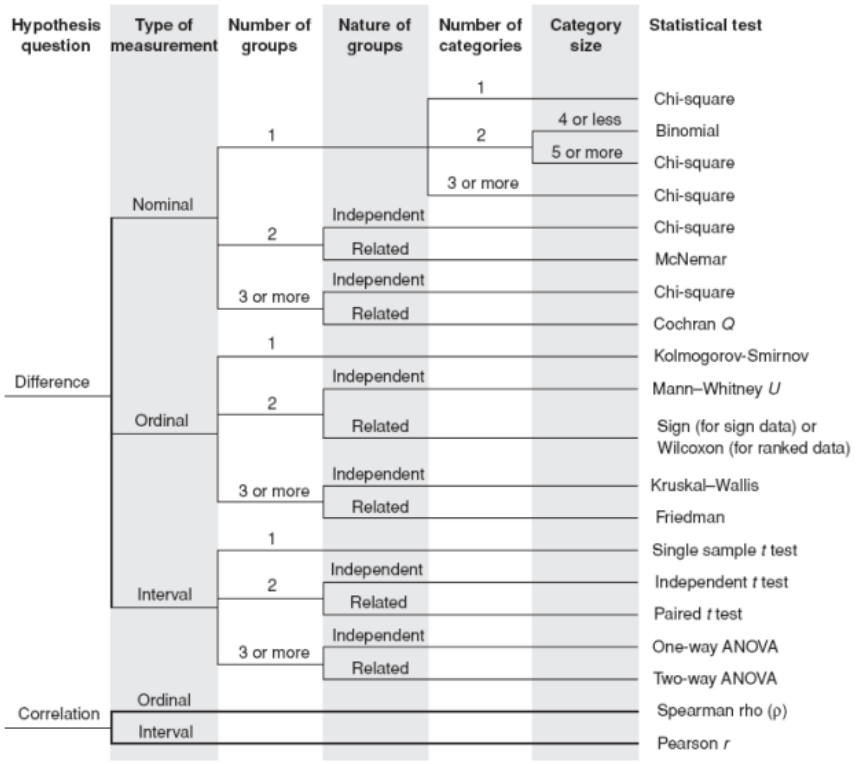
A measure of variability. The standard deviation quantifies how much the values vary from each other. A measure of the spread of individual observations around the mean value of the sample. A normally distributed, unskewed curve will have 34% of the cases between the mean and 1 standard deviation above or below the mean; 68% of cases between 1 standard deviation above and 1 below the mean; 95.5% of cases will be within two standard deviations of the mean. (see normal distribution above).

Standard error of the mean (SEM)

Another measure of variability. The standard error of the mean quantifies how accurately the true population mean is known. It is a measure of the variability of the mean of the sample as an estimate of the true value of the population mean. The larger the sample size, the smaller the standard error of the mean will be. It is used in computing confidence intervals. In a clinical trial, the larger the sample size the tighter the 95% CI is around the point estimate of the study.

Statistical tests

an algorithm (Holmes and Cockcroft, 2006)



Student's t test (t-test)

used to compare the difference between the mean of a continuous outcome (interval data) derived from one group with the mean from another group.

Survival analysis

Statistical procedures for estimating survival (prognosis) in a population under study.

T-test

also called the Student's t-test (named after the pen name of the person who developed the test), are used when there are two groups of independent study subjects and the data are continuous (eg, body weight) rather than discrete (eg, lameness score). The objective of using a t test is to compare the means of two populations of subjects.

Type I Error

Mistakenly rejecting the null hypothesis when it is actually true. The maximum probability of making a Type I error that the researcher is willing to accept is called alpha (a). Alpha is determined before the study begins. It leads to a false positive conclusion. Studies commonly set alpha to 1 in 20 (=0.05).

Type II Error

Mistakenly accepting (not rejecting) the null hypothesis when it is false. The probability of making a Type II error is called beta (b). Power = 1 - b (see above). It leads to a false negative conclusion. For trials the probability of a b error is usually set at 0.20 or 20% probability (i.e. a 20% chance of missing a true difference).

Type I error (α) :	Mistakenly rejecting the null hypothesis when it is actually true
Type II error (β):	Mistakenly accepting (not rejecting) the null hypothesis when it is actually true
Power (1- β) :	indicates the probability of finding a statistically significant difference **when one** exists (sensitivity)
α :	indicates the probability of finding a statistically significant difference **when one do**es not exist

Two tailed test

Is there a difference between groups A and B?

Volunteer bias

Owners who volunteer to participate in a trial may treat their animals differently to how non-volunteers do (e.g. volunteers' animals to be better looked after).

Withdrawal bias

Animals which are withdrawn from studies may differ systematically from those who remain.

References

This glossary has been adapted from the following publications:

- Holmes, M.A, & Cockcroft, P.D. (2008). *Handbook of Veterinary Clinical Research*. Wiley-Blackwell.
- Cockcroft, P.D. & Holmes, M.A. (2003). *Handbook of Evidence-based Veterinary Medicine*. Blackwell Publishing
- Evans, R.B. & O'Connor, A. (2007). Statistics and evidence-based veterinary medicine: answers to 21 common statistical questions that arise from reading scientific manuscripts. *Veterinary Clinics of North America: Small Animal Practice* 37, 477–486. DOI: https://doi.org/10.1016/j.cvsm.2007.01.006
- Smith, R.D. (2022). *Veterinary Clinical Epidemiology: From Patient to Population*, 4th ed. CRC Press.

Further reading

- Ellen, M. & Boggis H. The Statistics Tutor's Quick Guide to Commonly Used Statistical Tests. http://www.statstutor.ac.uk/resources/uploaded/tutorsquickguidetostatistics.pdf