BIOSTATISTICS

SAMPLING

Sampling is that part of statistical practice concerned with the selection of individual observations intended to yield some knowledge about a population of concern, especially for the purposes of statistical inference. Each observation measures one or more properties (weight, location, etc.) of an observable entity enumerated to distinguish objects or individuals. Survey weights often need to be applied to the data to adjust for the sample design. Results from probability theory and statistical theory are employed to guide practice.

Simple random sampling

In a simple random sample of a given size, all such subsets of the frame are given an equal probability. Each element of the frame thus has an equal probability of selection: the frame is not subdivided or partitioned. It is possible that the sample will not be completely random.

Stratified sampling

Where the population embraces a number of distinct categories, the frame can be organized by these categories into separate "strata." A sample is then selected from each "stratum" separately, producing a stratified sample. The two main reasons for using a stratified sampling design are to ensure that particular groups within a population are adequately represented in the sample, and to improve efficiency by gaining greater control on the composition of the sample. In the second case, major gains in efficiency (either lower sample sizes or higher precision) can be achieved by varying the sampling fraction from stratum to stratum. The sample size is usually proportional to the relative size of the strata. However, if variances differ significantly across strata, sample sizes should be made proportional to the stratum standard deviation. Disproportionate stratification can provide better precision than proportionate stratification. Typically, strata should be chosen to: have means which differ substantially from one another minimize variance within strata and maximize variance between strata.

Cluster sampling

Sometimes it is cheaper to 'cluster' the sample in some way e.g. by selecting respondents from certain areas only, or certain time-periods only. (Nearly all samples are in some sense 'clustered' in time - although this is rarely taken into account in the analysis.)

Random sampling

In random sampling, also known as probability sampling, every combination of items from the frame, or stratum, has a known probability of occurring, but these probabilities are not necessarily equal. With any form of sampling there is a risk that the sample may not adequately represent the population but with random sampling there is a large body of statistical theory which quantifies the risk and thus enables an appropriate sample size to be chosen

Systematic sampling

Selecting (say) every 10th name from the telephone directory is called an every 10th sample, which is an example of systematic sampling. It is a type of probability sampling unless the directory itself is not randomized. It is easy to implement and the stratification induced can make it efficient, but it is especially vulnerable to periodicities in the list. If periodicity is present and the period is a multiple of 10, then bias will result

Mechanical sampling

Mechanical sampling is typically used in sampling solids, liquids and gases, using devices such as grabs, scoops, thief probes, the COLIWASA and riffle splitter.
Care is needed in ensuring that the sample is representative of the frame. Much work in this area was developed by Pierre Gy.

SKEWNESS

In probability theory and statistics, skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable.
Skewness, the third standardized moment, is written as γ1 and defined as

$$\gamma_1 = \frac{\mu_3}{\sigma^3},$$

where μ3 is the third moment about the mean and σ is the standard deviation. Equivalently, skewness can be defined as the ratio of the third cumulant κ3 and the third power of the square root of the second cumulant κ2:

$$\gamma_1 = \frac{\kappa_3}{\kappa_2^{3/2}}.$$

KURTOSIS

In probability theory and statistics, kurtosis (from the Greek word κυρτός, kyrtos or kurtos, meaning bulging) is a measure of the "peakedness" of the probability distribution of a real-valued random variable. Higher kurtosis means more of the variance is due to infrequent extreme deviations, as opposed to frequent modestly-sized deviations.
The fourth standardized moment is defined as

$$\frac{\mu_4}{\sigma^4},$$

where μ4 is the fourth moment about the mean and σ is the standard deviation. This is sometimes used as the definition of kurtosis in older works, but is not the definition used here.

MEAN

Mean has two related meanings: The **arithmetic mean** (and is distinguished from the geometric mean or harmonic mean).The expected value of a random variable, which is also called the population mean.
It is sometimes stated that the 'mean' means average. This is incorrect if "mean" is taken in the specific sense of "arithmetic mean" as there are different types of averages: the mean, median, and mode. For instance, average house prices almost always use the median value for the average.

MEDIAN

Median is described as the number separating the higher half of a sample, a population, or a probability distribution, from the lower half. The median of a finite list of numbers can be found by arranging all the observations from lowest value to highest value and picking the middle one. If there is an even number of observations, the median is not unique, so one often takes the mean of the two middle values. At most half the population have values less than the median and at most half have values greater than the median. If both the population, then some of the population is exactly equal to the median.

STANDARD DEVIATION

The standard deviation is a measure of the dispersion of a set of values. It can apply to a probability distribution, a random variable, a population or a multiset. The standard

deviation is usually denoted with the letter σ (lower case sigma). It is defined as the root-mean-square (RMS) deviation of the values from their mean, or as the square root of the variance

STANDARD ERROR

The standard error of a method of measurement or estimation is the estimated standard deviation of the error in that method. Specifically, it estimates the standard deviation of the difference between the measured or estimated values and the true values. Notice that the true value of the standard deviation is usually unknown and the use of the term standard error carries with it the idea that an estimate of this unknown quantity is being used. It also carries with it the idea that it measures, not the standard deviation of the estimate itself, but the standard deviation of the error in the estimate, and these can be very different.

ADDITION RULE

The rule of addition applies to the following situation. We have two events from the same sample space, and we want to know the probability that either event occurs.

Rule of Addition If events A and B come from the same sample space, the probability that event A and/or event B occur is equal to the probability that event A occurs plus the probability that event B occurs minus the probability that both events A and B occur.

P(A □ B) = P(A) + P(B) - P(A ∩ B)

Note: Invoking the fact that P(A ∩ B) = P(A)P(B | A), the Addition Rule can also be expressed as

P(A □ B) = P(A) + P(B) - P(A)P(B | A)

PROBABILITY

Probability is the likelihood or chance that something is the case or will happen. Probability theory is used extensively in areas such as statistics, mathematics, science and philosophy to draw conclusions about the likelihood of potential events and the underlying mechanics of complex systems.

STANDARD DEVIATION

n probability and statistics, the standard deviation is a measure of the dispersion of a set of values. It can apply to a probability distribution, a random variable, a population or a multiset. The standard deviation is usually denoted with the letter σ (lower case sigma). It is defined as the root-mean-square (RMS) deviation of the values from their mean, or as the square root of the variance.

POISSON DISTRIBUTION

In probability theory and statistics, the Poisson distribution is a discrete probability distribution that expresses the probability of a number of events occurring in a fixed period of time if these events occur with a known average rate and independently of the time since the last event. The Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume.

NORMAL DIDTRIBUTION

The normal distribution, also called the Gaussian distribution, is an important family of continuous probability distributions, applicable in many fields. Each member of the family may be defined by two parameters, location and scale: the mean ("average", μ) and variance (standard deviation squared) σ2, respectively. The standard normal distribution is the normal distribution with a mean of zero and a variance of one (the red curves in the plots to the

right). Carl Friedrich Gauss became associated with this set of distributions when he analyzed astronomical data using them,[1] and defined the equation of its probability density function. It is often called the bell curve because the graph of its probability density resembles a bell.

T-Test

A t-test is any statistical hypothesis test in which the test statistic has a Student's t distribution if the null hypothesis is true. It is applied when sample sizes are small enough that using an assumption of normality and the associated z-test leads to incorrect inference. A t-test is any statistical hypothesis test in which the test statistic has a Student's t-distribution if the null hypothesis is true. It is applied when sample sizes are small enough that using an assumption of normality and the associated z-test leads to incorrect inference Among the most frequently used t tests are: A test of the null hypothesis that the means of two normally distributed populations are equal

HYPOTHEIS TEST

Statistical hypothesis test is a method of making statistical decisions from and about experimental data. Null-hypothesis testing just answers the question of "how well the findings fit the possibility that chance factors alone might be responsible."[1] This is done by asking and answering a hypothetical question. One use is deciding whether experimental results contain enough information to cast doubt on conventional wisdom.

SIGNIFICANCE

In statistics, a result is called statistically significant if it is unlikely to have occurred by chance. "A statistically significant difference" simply means there is statistical evidence that there is a difference; it does not mean the difference is necessarily large, important, or significant in the common meaning of the word.The significance level of a test is a traditional frequentist statistical hypothesis testing concept

CHISQURE TEST

"Chi-square test" is often shorthand for Pearson's chi-square test.

A chi-square test (also chi-squared or $\chi 2$ test) is any statistical hypothesis test in which the test statistic has a chi-square distribution when the null hypothesis is true, or any in which the probability distribution of the test statistic (assuming the null hypothesis is true) can be made to approximate a chi-square distribution as closely as desired by making the sample size large enough.

REGRESSION ANALYSIS

Regression analysis is a technique used for the modeling and analysis of numerical data consisting of values of a dependent variable (response variable) and of one or more independent variables (explanatory variables). The dependent variable in the regression equation is modeled as a function of the independent variables, corresponding parameters ("constants"), and an error term. The error term is treated as a random variable. It represents unexplained variation in the dependent variable. The parameters are estimated so as to give a "best fit" of the data. Most commonly the best fit is evaluated by using the least squares method, but other criteria have also been used

CORRELATION

Correlation, (often measured as a correlation coefficient), indicates the strength and direction of a linear relationship between two random variables. In general statistical usage, correlation or co-relation refers to the departure of two variables from independence. In this broad sense

there are several coefficients, measuring the degree of correlation, adapted to the nature of data

STUDENT T DISTRIBUTION

In probability and statistics, the standard deviation is a measure of the dispersion of a set of values. It can apply to a probability distribution, a random variable, a population or a multiset. The standard deviation is usually denoted with the letter σ (lower case sigma). It is defined as the root-mean-square (RMS) deviation of the values from their mean, or as the square root of the variance.

FACTOR ANALYSIS

In probability theory and statistics, correlation, (often measured as a correlation coefficient), indicates the strength and direction of a linear relationship between two random variables. In general statistical usage, correlation or co-relation refers to the departure of two variables from independence. In this broad sense there are several coefficients, measuring the degree of correlation, adapted to the nature of data.

NULL HYPOTHESIS

In statistics, a null hypothesis (H0) is a hypothesis set up to be nullified or refuted in order to support an alternative hypothesis. When used, the null hypothesis is presumed true until statistical evidence, in the form of a hypothesis test, indicates otherwise — that is, when the researcher has a certain degree of confidence, usually 95% to 99%, that the data does not support the null hypothesis. It is possible for an experiment to fail to reject the null hypothesis. It is also possible that both the null hypothesis and the alternate hypothesis are rejected if there are more than those two possibilities

HISTOGRAM

In statistics, a histogram is a graphical display of tabulated frequencies. It shows what proportion of cases fall into each of several categories. A histogram differs from a bar chart in that it is the area of the bar that denotes the value, not the height, a crucial distinction when the categories are not of uniform width (Lancaster, 1974). The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent.

The word histogram is derived from Greek: histos 'anything set upright' (as the masts of a ship, the bar of a loom, or the vertical bars of a histogram); gramma 'drawing, record, writing'. The histogram is one of the seven basic tools of quality control, which also include the Pareto chart, check sheet, control chart, cause-and-effect diagram, flowchart, and scatter diagram. A generalization of the histogram is kernel smoothing techniques. This will construct a very smooth probability density function from the supplied data.

BAR DIAGRAM

A graphical presentation in which the values of the dependent variable are represented by vertical or horizontal bars, drawn at coordinates on the other axis of the corresponding values of the independent discrete variable (see, e.g., Figure 2.4, in which two independent variables are used).

PIE DIGRAM

A pie chart (or a circle graph) is a circular chart divided into sectors, illustrating relative magnitudes or frequencies or percents. In a pie chart, the arc length of each sector (and consequently its central angle and area), is proportional to the quantity it represents. Together, the sectors create a full disk. It is named for its resemblance to a pie which has been sliced.

DOSE RESPONSE

Observable effect, a large amount is fatal), or to populations (e.g.: how many people are affected at different levels of exposure).

Studying dose response, and developing dose response models, is central to determining "safe" and "hazardous" levels and dosages for drugs, potential pollutants, and other substances that humans are exposed to. These conclusions are often the basis for public policy. When the agent is radiation instead of a drug, this is called the exposure-response relationship.

DOSE RESPONSE CURVE

A dose-response curve is a simple X-Y graph relating the magnitude of a stressor (e.g. concentration of a pollutant, amount of a drug, temperature, intensity of radiation) to the response of the receptor (e.g. organism under study). The response is usually death (mortality), but other effects (or endpoints) can be studied.The measured dose (usually in milligrams, micrograms, or grams per kilogram of body-weight) is generally plotted on the X axis and the response is plotted on the Y axis. Commonly, it is the logarithm of the dose that is plotted on the X axis, and in such cases the curve is typically sigmoidal, with the steepest portion in the middle.

LC 50 AND LD 50

LC stands for "Lethal Concentration". LC values usually refer to the concentration of a chemical in air but in environmental studies it can also mean the concentration of a chemical in water.

For inhalation experiments, the concentration of the chemical in air that kills 50% of the test animals in a given time (usually four hours) is the LC50 value.

LD stands for "Lethal Dose". LD50 is the amount of a material, given all at once, which causes the death of 50% (one half) of a group of test animals. The LD50 is one way to measure the short-term poisoning potential (acute toxicity) of a material.

LD/LC50 tests

In nearly all cases, LD50 tests are performed using a pure form of the chemical. Mixtures are rarely studied.

The chemical may be given to the animals by mouth (oral); by applying on the skin (dermal); by injection at sites such as the blood veins (i.v.- intravenous), muscles (i.m. - intramuscular) or into the abdominal cavity (i.p. - intraperitoneal).

The LD50 value obtained at the end of the experiment is identified as the LD50 (oral), LD50 (skin), LD50 (i.v.), etc., as appropriate.

In general, if the immediate toxicity is similar in all of the different animals tested, the degree of immediate toxicity will probably be similar for humans. When the LD50 values are different for various animal species, one has to make approximations and assumptions when estimating the probable lethal dose for man. Tables 1 and 2 have a column for estimated lethal doses in man. Special calculations are used when translating animal LD50 values to possible lethal dose values for humans. Safety factors of 10,000 or 1000 are usually included in such calculations to allow for the variability between individuals and how they react to a chemical, and for the uncertainties of experiment test results.

F-RATIO

In oceanic biogeochemistry, the f-ratio is the fraction of total primary production fuelled by nitrate (as opposed to that fuelled by other nitrogen compounds such as ammonium). This fraction is significant because it is assumed to be directly related to the sinking (export) flux of organic marine snow from the surface ocean by the biological pump. The ratio was originally defined by Richard Eppley and Bruce Peterson in one of the first papers estimating global oceanic production[1].

ANALYSIS OF VARIANCE

In statistics, analysis of variance **(ANOVA)** is a collection of statistical models, and their associated procedures, in which the observed variance is partitioned into components due to different explanatory variables. The initial techniques of the analysis of variance were developed by the statistician and geneticist R. A. Fisher in the 1920s and 1930s, and is sometimes known as Fisher's ANOVA or Fisher's analysis of variance, due to the use of Fisher's F-distribution as part of the test of statistical significance.

MULTIPLE RANGE TEST

In statistics, **Duncan's new multiple range test (MRT)** is a multiple comparison procedure developed by David B. Duncan in 1955. Duncan's MRT belongs to the general class of multiple comparison procedures that use the studentized range statistic qr to compare sets of means. Duncan's new multiple range test (MRT) is a variant of the Student Newman Keuls method that uses increasing alpha levels to calculate the critical values in each step of the Newman Keuls procedure. Duncan's MRT attempts to control family wise error rate (FWE) at $\alpha ew = 1 - (1 - \alpha pc)k-1$ when comparing k, where k is the number of groups

FREQUENCY CURVE

A smooth curve which corresponds to the limiting case of a histogram computed for a frequency distribution of a continuous distribution as the number of data points becomes very large.