

Definition of Terms Iss2

Central Limit Theorem

The central limit theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately normally distributed.

Inductive-Deductive Reasoning

Inductive reasoning is a method of drawing conclusions by going from the specific to the general. It's usually contrasted with deductive reasoning, where you proceed from general information to specific conclusions. Inductive reasoning is also called inductive logic or bottom-up reasoning.

Law of Large Numbers

In probability theory it is a mathematical theorem that states that the average of the results obtained from a large number of independent and identical random samples converges to the true value, if it exists. More formally, the LLN states that given a sample of independent and identically distributed values, the sample mean converges to the true mean.

The LLN is important because it guarantees stable long-term results for the averages of some random events. For example, while a casino may lose money in a single spin of the roulette wheel, its earnings will tend towards a predictable percentage over a large number of spins. Any winning streak by a player will eventually be overcome by the parameters of the game.

Populations and Samples

A **population** is the entire set of similar items or events that are of interest to some question or experiment. Populations can be finite or infinite.

Variables are denoted by capital letters X Y N

A **sample** is a manageable subset.

A **parameter** describes a property of a population. Greek letters. μ and σ

A **statistic** describes a property of a sample. Use Roman letters $\bar{X}$ and s

Test statistics before being identified are denoted by capital letters.

So we get say let $X = x_i$ when defining a sample value drawn from the entire population.

Regression Analysis

In statistical modelling, **regression analysis** is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome' or 'response' variable, or a 'label' in machine learning parlance) and one or more independent variables (often called 'predictors', 'covariates', 'explanatory variables' or 'features'). The most common form of regression analysis is linear regression, in which one finds the line (or a more complex linear combination) that most closely fits the data according to a specific mathematical criterion. For example, the method of ordinary least squares computes the unique line (or hyperplane) that minimizes the sum of squared differences between the true data and that line (or hyperplane). For specific mathematical reasons (see linear regression), this allows the researcher to estimate the conditional expectation (or population average value) of the dependent variable when the independent variables take on a given set of values.

$Y = a + rX$ where r is the regression coefficient.

Reliability

Reliability or reproducibility in statistics refers to consistency in measurement: the capacity to produce the same result for two identical states; or, more operationally, the closeness of the initial estimated value to the subsequent estimated value.

Sample Space

A sample space is a collection or a set of possible outcomes of a random experiment. The sample space is represented using the symbol, “S”. The subset of possible outcomes of an experiment is called events.

Scales

Categorical - Nominal scales are labels for categories but there is no implied order. They can be analysed for frequency and mode only. eg Crisp flavours. Even though they may use “numbers” as labels do not treat the variables as numeric. nb. Dice throws depend on the application.

Categorical - Ordinal scales introduce the idea of rank but lack a consistent interval. They can be analysed using measures of central tendency such as mode and median but not mean. eg “How satisfied are you on scale 1-5 – “5” being very happy”.

Numeric - Interval scales introduce the concept of equal intervals but there is no true zero. eg time and temperature. They can be analysed by mode, median and mean.

Numeric - Ratio scales introduce the idea of absolute zero. eg height and weight

Unbiased

An unbiased estimator is when a statistic does not overestimate or underestimate a population parameter. In other words, a value is unbiased when it is the same as the actual value of a particular parameter. There are different variables that can influence whether or not a particular sample's statistic is biased

Validity

Validity is the main extent to which a concept, conclusion, or measurement is well-founded and likely corresponds accurately to the real world.

Variables

Properties of members of the population that change are called **variables**.

Random variables are those that change due to an element of chance.

Qualitative variables are variables that are categorical (non-numeric).

Quantitative variables are numeric and can be **discrete** or **continuous**.

Discrete are technically “countably infinite”.

Continuous variables are “uncountable infinite”.

Explanatory variables are the variables that can be altered or manipulated in research (eg *a change in dosage*) while **Response** variable is the result of manipulation done to the variables (eg *the time it took for the reaction to occur*).