Statistics

Glossary of Terms

Brendan Edwards :11759222

Contents

A	Term	Definition
	Alternative hypothesis (H_1 or H_A) (H_0 or H_1)	the statement of change, which also indicates the direction of the change. Plays a role therefore in determining the rejection region
	Analysis of variance (ANOVA)	tests whether populations have the same mean • compares how far apart the sample means are with • how much variation there is within the samples

B		
	Bayes rule probability	***Definition required*** ***Discuss difference / similarity to conditional probability***
	Binomial Distribution Normal Approximation	a process where the shape of the binomial distribution is estimated using the normal distribution curve. The use of normal approximation makes calculating the probabilities of the binomial distribution easier (Chegg, 2022)

C		
	Categorical Data	Labels
	Chi-squared (X^2) statistic	a measure of how far the observed counts in a one-way or two-way table are from the expected counts.
	Coefficient of variation	measures the spread of a data set as a proportion of its mean: (σ/μ).
	Confidence Coefficient	the probability that a will contain the true value of the population parameter. For example, if the confidence coefficient is 0.95, 95 per cent of the confidence intervals so calculated for a large number of random samples would contain the parameter.
	Confounding	(also **confounding variable**, **confounding factor**, **extraneous determinant** or **lurking variable**) is a variable that influences both the dependent variable and independent variable, causing a spurious association. Confounding is a causal concept, and as such, cannot be described in terms of correlations or associations (Wikipedia, 2022b)
	Conditional probability	A measure of the probability of an event occurring (A), given that another event (by assumption, presumption, assertion or evidence) has already occurred (B). (Wikipedia, 2022a) P(A|B)
	Confidence Interval	point estimate of parameter $\pm$ margin of error
	Continuous data	data where the observations may take on any value in one or more intervals. For example: height, weight and temperature.
	Correlation	measures the strength and direction of the linear relationship between two quantitative variables
	Correlation coefficient	a numerical value that indicates the degree and direction of relationship between two variables; the coefficients range in value from +1.00 (perfect positive relationship) to 0.00 (no relationship) to −1.00 (perfect negative or inverse relationship).
	Cumulative distribution function	gives the probability of the random variable taking any value up to and including the specified value $P(X < x)$.

D		
	Descriptive statistics	deals with methods of organising, summarising and presenting numerical data in a convenient form.
	Degrees of freedom	describes the number of values in the final calculation of a statistic that are free to vary.
	Discrete data	observations belonging to it are distinct and separate; i.e. they can be counted (1,2,3, : : :). Examples might include the number of kittens in a litter or the number of patients in a doctor's surgery
	Disjoint Event	Probability Independent Events. The fact that A occurs, means that B cannot occur
	Distribution	of a variable tells us what values it takes and with what probabilities it takes these values

E		
	Error: Type 1	The rejection of the true Null hypothesis. Also known as the false positive error (Institute, 2022a)
	Error: Type 2	A hypothesis test that fails to reject the Null hypothesis that is false. Causes the user to ***erroneously not reject the false null hypothesis,*** because the test lacks the statistical power to detect sufficient evidence in favour of the Alternative hypothesis also known as a false negative (Institute, 2022b)
	Error sum of squares (SSE):	represents the amount of variability in the response variable which remains unexplained after a model has been fitted. (Also known as `Residuals SS')
	Event (E):	a group of one or more outcomes chosen from the sample space
	Event. Disjoint	Probability Independent Events. The fact that A occurs, means that B cannot occur, no outcomes in common Cannot depict by Venn Diagram
	Event. Independent	Probability Independent Events. If A occurs, B is multiplied. P(A) x P(B) Can depict by Venn Diagram
	Experiment	any process which results in the collection of data where one or more variables are deliberately manipulated
	Experimental Design Hypothesis Testing	The process of carrying out research in an objective and controlled fashion so that precision is maximized and specific conclusions can be drawn regarding a hypothesis statement. Generally, the purpose is to establish the effect that a factor or independent variable has on a dependent variable. (Bell, 2009) There are five key steps in designing an experiment: 1. Consider your **variables** and how they are related 2. Write a specific, testable **hypothesis** 3. Design experimental treatments to manipulate your **independent variable** 4. Assign subjects to groups, either **between-subjects** or **within-subjects** 5. Plan how you will measure your **dependent variable** (Bevins, 2022)

F		
	Family-Wise Error (Hypothesis Testing) (Statology, 2020) (Wikipedia, 2021a)	When conducting multiple hypothesis tests at once, the probability of getting a false positive increases. **Family-wise error rate = $1 - (1-\alpha)^n$** where: • **α:** The significance level for a single hypothesis test • **n:** The total number of tests

H		
	Hypothesis (Alternative) H_1 OR H_A	The statement of change Also indicates the direction of change. Used to determine the rejection region
	Hypothesis (Null) H_0	The statement of `no change' Statement of no effect, no difference, no association
	Hypothesis (Family-Wise Error) (Statology, 2020)	when conducting multiple hypothesis tests at once, the probability of getting a false positive increases. **Family-wise error rate = $1 - (1-\alpha)^n$** where: • **α:** The significance level for a single hypothesis test • **n:** The total number of tests
	Hypothesis Test Statistic	a quantity calculated from the sample of data. Used to decide if the Null Hypothesis should be rejected
	Hypothesis Single Sided	μ is on one specific side of parameter $\mu >$ *parameter* and *parameter* $< \mu$
	Hypothesis Two Sided	μ may lie on either side of parameter $\mu \neq$ *parameter*

I		
	Interval data	observations can be measured but the zero point is arbitrary, and forming ratios is not meaningful. For example, when measuring temperature, 0°C does not represent the absence of temperature, so temperatures (in °C) form interval data.
	Independent Event ***probability***	Probability Independent Events. ***Better definition required*** Event Probabilities are multiplied. P(A) x P(B)

L		
	Lurking Variable	A variable that is not included in a statistical analysis, yet impacts the relationship between two variables within the analysis.

M		
	Margin of Error	maximum difference suggested by the confidence interval between • the point estimate and • the true value
	Mean	Average of the Observations. Numerically, it equals the sum of the observations divided by the number of observations.
	Median	the observation that falls in the middle of a set of measurements when the measurements are arranged in order of magnitude (lowest to highest)
	Mode	the observation that occurs most frequently in a set of discrete (Real / Whole Numbers) data
	Mutually Exclusive Event probability	***Definition required*** ***Discuss difference between independent event***

N		
	Nominal data	Categorical Observations converted to Numbers Observations can be counted, but not ordered or measured For example, in a data set 1 = males 0 = females.
	Normal Approximation Binomial Distribution	a process where the shape of the binomial distribution is estimated using the normal distribution curve. The use of normal approximation makes calculating the probabilities of the binomial distribution easier (Chegg, 2022)
	Null Hypothesis (H_0)	is the statement under the belief of `no change' from what has historically been the case; i.e. a statement of no effect, no difference, no association

O		
	Ordinal data	observations can be ranked, put in order. You can count and order but not measure nominal data. For example, rating a television show on a scale of 1 to 4, representing `strongly like', `like', `dislike' and `strongly dislike'
	Outlier	an observation in a data set which is far removed in value from the other observations 1.5 x Inter Quartile Range

P		
	Parameter	A value, usually unknown, used to represent a certain characteristic of a population has to estimated also called a 'statistic'
	Pearson's Coefficient	a measure of the strength of the association between two continuous variables (Kenton, 2021)
	Population	the entire set of observations or measurements under study
	Probability	is the proportion of times the event occurs in many repeated trials
	Proportion	is the probability of times the event occurs in many repeated trials
	p-value	the probability computed, assuming H_0 *is true*, that the observed outcome would take a value as extreme or more extreme than that actually observed

Q		
	Qualitative data	nominal or ordinal observations usually, categories or names
	Quantitative variable	numerical values arithmetic operations such as adding and averaging make sense

R		
	Range	a measure of the spread of some data. the difference between the largest and smallest observed value
	Ratio data	observations can be measured there exists a true zero point. Forming a ratio of two observations is meaningful. Example: the measurement of length or volume
	Replication	replication is repetition of an experiment or observation in the same or similar conditions. Replication is important because it adds information about the reliability of the conclusions or estimates to be drawn from the data (statistics.com, 2013)
	Robust	to be as resistant as possible to the effects of extreme data values or outliers

S		
	Sample	a set of observations selected from the population. By studying the sample, it is hoped to draw valid conclusions about the population
	Sample space (S)	a list of all the possible outcomes of an experiment
	Significance level (α)	gives the probability of making a Type I error | that is, falsely rejecting H_0, the null hypothesis
	Simpson's Paradox	an instance in which the total data set shows one trend while subsets of the data set show the opposite trends or none at all. Simpson's paradox goes by many names, among them the Yule-Simpson effect, reversal paradox, or amalgamation effect (Terms, 2019)
	Skewness	asymmetry in the distribution of the sample data values. if the long tail is on the right of the distribution The data are `skewed to the right' (`positively skewed')
	SRS Simple Random Sample	Simple Random Sample. a subset of individuals (a sample) chosen from a larger set (a population) in which a subset of individuals are chosen randomly (Wikipedia, 2021b)
	Standard deviation	a measure of to what degree the individual observations of a data set are dispersed or spread out around the mean
	Statistic	formula
	Statistical inference	the process of drawing conclusions about a population based upon information in a sample

T		
	Treatment (ANOVA)	Method of testing
	Test statistic	a quantity calculated from the sample of data. Its value is used to decide whether or not the null hypothesis should be rejected in the hypothesis test
	Treatment (ANOVA)	In an experiment, the factor (also called an independent variable) is an explanatory variable manipulated by the experimenter. Each factor has two or more levels, i.e., different values of the factor. Combinations of factor levels are called **treatments**. (Trek, 2022)
	Trial	

V		
	Variable	is any characteristic of an individual that actually varies. For example, if you deliberately include only males in a sample, then `sex' is not a variable in your data set
	Variance	a measure of the spread of a set of data. It is an `average' of the squares of the deviations of the observations from their mean. Its square root is the standard deviation

Bell, S. (2009). *Experimental Design*. Retrieved 2022.06.13 from https://www.sciencedirect.com/topics/earth-and-planetary-sciences/experimental-design

Bevins, R. (2022). *A Quick Guide to Experimental Design | 5 Steps & Examples*. Retrieved 2022.06.13 from https://www.scribbr.com/methodology/experimental-design/

Chegg. (2022). *Normal Approximation*. Retrieved 2022.05.15 from https://www.chegg.com/homework-help/definitions/normal-approximation-31

Institute, C. F. (2022a). *What is a Type I Error?* Retrieved 2022.06.13 from https://corporatefinanceinstitute.com/resources/knowledge/other/type-i-error/

Institute, C. F. (2022b). *What is a Type II Error?* Retrieved 2022.06.13 from https://corporatefinanceinstitute.com/resources/knowledge/other/type-ii-error/

Kenton, W. (2021). *Pearson Coefficient*. Investopedia. Retrieved 2022.06.13 from https://www.investopedia.com/terms/p/pearsoncoefficient.asp

statistics.com. (2013). *Replication*. Retrieved 2022.06.13 from https://www.statistics.com/glossary/replication/

Statology. (2020). *What is the Family-wise Error Rate?* Retrieved 2022.05.20 from https://www.statology.org/family-wise-error-rate/

Terms, S. (2019). *Simpson's Paradox Definition*. Retrieved 2022.06.13 from https://scienceterms.net/psychology/simpsons-paradox/

Trek, S. (2022). *Treatment*. Stat Trek. Retrieved 2022.05.15 from https://stattrek.com/statistics/dictionary.aspx?definition=treatment

Wikipedia. (2021a). *Family-wise error rate*. Retrieved 2022.05.20 from https://en.wikipedia.org/wiki/Family-wise_error_rate

Wikipedia. (2021b). *Simple random sample (SRS)*. Retrieved 2022.06.07 from https://en.wikipedia.org/wiki/Simple_random_sample

Wikipedia. (2022a). *Conditional probability*. https://en.wikipedia.org/wiki/Conditional_probability

Wikipedia. (2022b). *Confounding*. Retrieved 2022.05.31 from https://en.wikipedia.org/wiki/Confounding