Outline of Statistics
January 2, 2012

Contents

Note: To look up references, see the Consciousness Bibliography, listing 10,000 books and articles, with full journal and author names, available in text and PDF file formats at
http://www.outline-of-knowledge.info/Consciousness_Bibliography/index.html.

MATH>Statistics

statistics
    Mathematics {statistics} can be about probability, populations, distributions, and testing.

analysis of proximities
    After rank ordering, proximities can show approximate distances {analysis of proximities}.

order statistics
    Statistics {order statistics} can be about sequences.

MATH>Statistics>Correlation

correlation in statistics
    Two population-sample measurements can have relation {correlation, statistics}| {regression coefficient}. Sample members are measurement pairs. Both measurement values can increase {positive correlation}, decrease {negative correlation}, or have no correlation.

correlation coefficient
    For paired measurements, such as individual weight and height, sum of products of z scores, divided by number of individuals, makes a number {correlation coefficient}| between -1 and +1. Correlation coefficient R measures relations between two variables: $R^2 = SSR/SST = 1 - SSE/SST$, where SSR is residual sum of squares, SST is term sum of squares, and SSE is error sum of squares.
z scores
    Sum from $i = 1$ to $i = N$ of $z1(i) * z2(i) / N$, where z is z score and N is population size.
means
    Sum from $i = 1$ to $i = N$ of $(n1(i) * n2(i) / N - x1 * x2) / (s1 * s2)$, where n is value, x is mean, s is standard deviation, and N is population size.

values

$(N * ($sum from $i = 1$ to $i = N$ of $n1(i) * n2(i)) - ($sum from $i = 1$ to $i = N$ of $n1(i)) * ($sum from $i = 1$ to $i = N$ of $n2(i)))$ / $((N * ($sum from $i = 1$ to $i = N$ of $(n1(i))^2) - ($sum from $i = 1$ to $i = N$ of $n1(i)) * ($sum from $i = 1$ to $i = N$ of $n1(i)))^{0.5} * ((N * ($sum from $i = 1$ to $i = N$ of $(n2(i))^2) - ($sum from $i = 1$ to $i = N$ of $n2(i)) * ($sum from $i = 1$ to $i = N$ of $n2(i)))^{0.5}$.

correlation coefficient test

If population distribution is not normal, testing correlation coefficient {correlation coefficient test} can show if two event sets relate. Hypothesize that correlation coefficient is zero. Choose significance level. Degrees of freedom are sample size minus two for the two events. Convert correlation coefficient to t value: $t = r * (N - 2)^{0.5} / (1 - r^2)^{0.5}$, where r is correlation coefficient and N is number of individuals. If t value is less than t-distribution value at that significance level and degrees of freedom, do not reject hypothesis.

MATH>Statistics>Correlation>Regression

regression line

Lines {regression line}| {regression curve} closest to all points can go through correlation graphs.

linear regression

Regression curve can be straight line {linear regression}|.

best fit

Regression lines pass as close as possible {best fit}| to all points and so minimize sum of distances from points to regression line. Lines can pass closest to points if sum of squares of differences from line to points is minimum. Best-fit lines must pass through all property means: $x2 = m * x1 + b$, where x1 and x2 are means, m is slope, and b is y-intercept. $m = (N * ($sum from $i = 1$ to $i = N$ of $n1(i) * n2(i)) - ($sum from $i = 1$ to $i = N$ of $n1(i)) * ($sum from $i = 1$ to $i = N$ of $n2(i))) / (N * ($sum from $i = 1$ to $i = N$ of $n1(i)^2) - ($sum from $i = 1$ to $i = N$ of $n1(i)^2))$. $b = x2 - m * x1$.

prediction

Regression curves can predict {prediction} second-variable amount from first-variable amount. Second-variable value y equals regression-curve slope m times first variable x plus intercept b: $y = m*x + b$.

MATH>Statistics>Distribution

distribution statistics

Measurement values can have frequencies {distribution, population}| {frequency distribution} {theoretical distribution}, such as frequencies at ages.

binomial distribution

Distributions {binomial distribution} can reflect probability of event series that have two possible outcomes. Terms equal $(n! / (r! * (n - r)!)) * p^r * (1 - p)^{(n - r)}$, where n is event number, r is favorable-outcome number, and p is favorable-outcome probability. Mean equals $n*p$. Variance equals $n * p * (1 - p)$.
ratio

Small p with caret (^) {p-hat} denotes ratio of subset X to sample N: $= X/N$. p-hat standard deviation $= (p * (1 - p) / N)^{0.5}$ or $(N * p * (1 - p))^{0.5} / N$, if $p = X/N$ and $1 - p = (N - x) / N$.

bivariate distribution

Distributions {bivariate distribution} can have two random variables. Average value of $(x1 - u1) * (x2 - u2)$, where x is value and u is mean, measures covariance.

hypergeometric distribution

N things can have x of one kind and N - x of another kind {hypergeometric distribution}. Mean equals $R*p$, where p is favorable-outcome probability and R is favorable-outcome number. Variance equals $R * p * (1 - p) * ((N - R) / (N - 1))$. Terms equal $(x! / (x1! * (x - x1)!)) * ((N - x)! / ((N - x1)! * (N - R - x + x1)!)) / (N! / (R! * (N - R)!))$, where x is thing or event and x1 is number of things that are the same.
comparisons

Hypergeometric distributions approximate binomial distributions, if probability is less than 0.1 and number of things is large. Hypergeometric distributions approximate Poisson distributions, if number of things is large and favorable-event number divided by thing number is less than 0.1. Hypergeometric distributions approximate normal distributions, if mean is greater than or equal to four.

normal distribution

Symmetrical distributions {normal distribution} {normal curve} {Gaussian curve} {Gaussian distribution} over continuous domain can have highest frequencies near mean and lowest frequencies farthest from mean. $y = (1 / (s * (2 * pi)^{0.5})) * (e^{(-(x - u)^2 / (2 * s^2)))}$, where x is domain value, y is frequency, u is mean, and s is standard deviation.

median

In normal distributions, mean equals mode equals median.

approximations

Non-normal distributions can transform to normal distributions using square root of x or logarithm of x.

purposes

Normal distribution models random errors {error curve}. Normal distributions result from measurements that have many factors or random errors. For example, height results from genetics, diet, exercise, and so on, and has normal distribution.

Passing inputs through different-threshold sigmoidal functions, and then finding differences, results in Gaussian distributions.

theorem

If many random same-size samples come from a large population with normal distribution, sums of samples make a normal distribution {central limit theorem, normal distribution}, as do sample means.

mean

Sample-mean mean is an unbiased population-mean estimate.

Pascal triangle

If binomial coefficients are in a triangle {Pascal's triangle} {Pascal triangle}, so nth row has coefficients for n, coefficients are sums of two nearest preceding-row coefficients. Pascal's triangle is 1 / 1 1 / 1 2 1 / 1 3 3 1 / 1 4 6 4 1 / ...

Poisson distribution

Asymmetrical distributions {Poisson distribution, statistics} can have most numbers near mean. Domain starts at zero and is continuous. Mean and variance equal n*p. Terms equal $(u^r / r!) * e^{-u}$, where r is favorable-outcome number, from zero to infinity, u equals n*p, p is favorable-outcome probability, and n is event number. Poisson distribution is binomial-distribution limit, as p goes to zero and N goes to infinity.

standard score

z scores can convert to whole numbers {T score} {standard score} between 0 and 100: T = 50 + z*10. T score has mean 50 and standard deviation 10.

z distribution

Normal-distribution variable can change from x to z = (x - u) / s {standard measure} {z distribution} {z score}, where u equals mean and s equals standard deviation. New mean equals zero, and new standard deviation equals one. z score measures number of standard deviations from mean.

MATH>Statistics>Distribution>Markov Process

Markov process

Outcome or state probability can depend on outcome or state sequence {Markov process}. Events can depend on previous-event order. Transition graphs show event orders needed to reach all events.

Markov chain Monte Carlo method

Taking samples can find transition matrices, or using transition matrices can generate data points {Markov chain Monte Carlo method} (MCMC). MCMC includes jump MCMC, reversible-jump MCMC, and birth-death sampling.

Markov-modulated Poisson process

Hidden Markov models {Markov-modulated Poisson process} can have continuous time.

autocorrelation function

Functions {autocorrelation function} (ACF) can measure if process description needs extra dimensions or parameters.

autoregressive integrated moving average

An average {autoregressive integrated moving average} (ARIMA) can vary around means set by hidden Markov chains.

Bayesian information criterion

Minus two times Schwarz criterion {Bayesian information criterion} (BIC) approximates hidden-Markov-model number of states.

Bayesian model

Models {Bayesian model} can estimate finite-state Markov chains.

direct Gibbs sampler

Methods {direct Gibbs sampler, sampling} can sample states in hidden Markov chains.

EM algorithm

Algorithms {EM algorithm} can have E and M steps.

finite mixture model

Hidden Markov models {finite mixture model} can have equal transition-matrix rows.

forward-backward Gibbs sampler

Stochastic forward recursions and stochastic and non-stochastic backward recursions {forward-backward Gibbs sampler, sampling} can sample distribution.

forward-backward recursion

Recursion {forward-backward recursion, sampling} can adjust distribution sampling. It is similar to Kalman-filter prediction and smoothing steps in state-space model.

hidden Markov model

Finite-state Markov chain {hidden Markov model} (HMM) samples distribution.
model

Hidden Markov models are graphical models. Bayesian models are finite-state Markov chains.
purposes

Hidden Markov chains model signal processing, biology, genetics, ecology, image analysis, economics, and network security. Situations compare error-free or non-criminal distribution to error or criminal distribution.
transition

Hidden-distribution Markov chain has initial distribution and time-constant transition matrix.
calculations

Calculations include estimating parameters by recursion {forward-backward recursion, Markov} {forward-backward Gibbs sampler, Markov} {direct Gibbs sampler, Markov}, filling missing data, finding state-space size, preventing switching, assessing validity, and testing convergence by likelihood recursion.

hidden semi-Markov model

Models {hidden semi-Markov model} can maintain distribution states over times.

Langevin diffusion

In Monte Carlo simulations, if smaller particles surround particle, where will particle be at later time? Answer uses white noise {Langevin equation} {Langevin diffusion}: 6 * k * T * (lambda / m) * D(t), where lambda/m is friction coefficient, T is absolute temperature, k is elasticity, and D(t) is distance at time t. At much later time, distance = (k * T) / (m * lambda).

likelihood recursion
Recursion {likelihood recursion} can calculate over joint likelihood averaged over time, typically taking logarithm.

marginal estimation
Methods {marginal estimation} can estimate hidden Markov distribution and probability.

maximum a posteriori
Methods {maximum a posteriori estimation} (MAP) can estimate hidden Markov distribution and probability.

Metropolis-Hastings
Methods {Metropolis-Hastings sampler} can sample from multivariate normal distribution centered on current data point, depending on Hastings probability.

predictive distribution
Distributions {predictive distribution} can measure how well models predict actual data.

Schwarz criterion
Asymptotic approximations {Schwarz criterion} to logarithm of state-number probability can depend on likelihood maximized over transition matrix, sample size, and free-parameter number.

state-space model
Models {state-space model} can use Gaussian distributions for hidden Markov models.

Viterbi algorithm
Algorithms {Viterbi algorithm} can find most likely Bayesian trajectory, using forward-backward recursion with maximization, rather than averaging, and can estimate hidden Markov distribution.

MATH>Statistics>Events

combination in statistics
Systems can permute things in sequences of positions. Systems can have things in sequences of positions but sequence does not matter or some things are the same, so some permutations are the same {combination, mathematics}. Combinations are number of different permutations. Number of combinations is less than or equal to number of permutations.
permutations
N positions and N things have number of permutations $N! = N * (N - 1) * (N - 2) * ... * 1$. N things and R positions, with $R < N$, have number of permutations $N! / (N - R)! = N * (N - 1) * (N - 2) * ... * (N - (R - 1))$. N things and R positions, with $N <= R$, have number of permutations $N! = N * (N - 1) * (N - 2) * ... * 1$.
combinations
Number of combinations C equals number of permutations P divided by factorial of number of positions R: $C = P / R!$. Number of combinations of n things taken r at a time is $n! / r!$.

Number of combinations of $n + 1$ things taken r at a time equals number of combinations of n things taken $r - 1$ at a time, plus number of combinations of n things taken r at a time: $(n + 1)! / r! = n! / (r - 1)! + n! / r!$.

Number of combinations of n things taken $n - r$ at a time equals number of combinations of n things taken r at a time: $n! / (n - r)! = n! / r!$.

Number of combinations of n things taken $r + 1$ at a time equals number of combinations of n things taken r at a time, times quantity $(n - r)$ divided by $(r + 1)$: $n! / (r + 1)! = (n! / r!) * (n - r) / (r + 1)$.
binomial
Systems can have things that have states. Two things, each with possible states a and b, have four permutations and three combinations: 1 a*a, 2 a*b, and 1 b*b. Three things with states a and b have eight permutations and four combinations: 1 a*a*a, 3 a*a*b, 3 a*b*b, or 1 b*b*b, and so on.

Combination-term coefficients derive from binomial powers. $(a + b)^2 = a^2 + 2*a*b + b^2$, $(a + b)^3 = a^3 + 3*a^2*b + 3*a*b^2 + b^3$, $(a + b)^4 = a^4 + 4*a^3*b + 6*a^2*b^2 + 4*a*b^3 + b^4$, and so on.

independent event

First-event outcome can have no affect on second-event outcome {independent event}. For independent events, to find probability of outcome series, multiply event probabilities: P = p1 * p2 * p3 * ... * pN, where pi are probabilities, and N is event number.

mutually exclusive events

Event outcome can mean that other outcomes cannot happen {disjoint events} {mutually exclusive events}|. For equally probable outcomes, outcome probability is 1/N, where N is number of mutually exclusive outcomes. To find probability of at least one of mutually exclusive outcomes, add all probabilities: P = p1 + p2 + p3 + ... + pN, where pi are probabilities, and N is outcome number.

permutation in statistics

Systems can have things that can be at sequence positions, or a thing can have a succession of states {permutation}| {arrangement}.

states

Number P of permutations for succession of states is product of number N of states raised to power of number R of events: N^R.

things

For sequences, number of permutations is product of number N of things times number of things minus one, and so on, until number R of sequence positions: N * (N - 1) * (N - 2) * ... * (N - R + 1). Using an outcome makes it unavailable for succeeding events.

stochastic process

Event series {stochastic process}| can have different results with different probabilities. Events can be independent {Bernoulli sequence}. Events can depend on preceding events in a Markov process.

transition graph

Graphs {transition graph} can show event orders needed to reach all events.

MATH>Statistics>Graph

graph statistics

Two-dimensional displays {graph} can show distributions.

bar graph

Values or value ranges can have bars whose height or length indicates frequency {bar graph}. For example, yellow has three items, orange has four items, and green has two items. Straight lines can connect bar-top midpoints {frequency polygon}.

histogram

Bar graphs {histogram}| can display frequency versus property value over property-value range. For example, 0 to 5 has 15 items, 5 to 10 has 20 items, and 10 to 15 has 12 items.

scattergram

Coordinate systems {scatter diagram} {scattergram}| can display data dispersion.

MATH>Statistics>Probability

probability statistics

Event-outcome chance {probability} {risk} is between zero and one. To find probability, count how many times outcome happens compared to how many times event repeats: p = outcomes / events. Probability can be from theory {a priori probability} or from experiment {empirical probability}.

conditional probability

First-event outcome can influence second-event outcome {conditional probability}| {prior probability}. For conditional probability, to find probability that outcome happens in first event and outcome happens in second event,

multiply first-outcome probability times modified second-outcome probability {conditional probability law} {law of conditional probability}: P = p1 * p2(p1).

Kolmogorov probability
Systematic probability {Kolmogorov probability} {Kolmogorov axioms} can use three axioms. Outcome probability is zero or positive real number. Probability that event has some outcome is one. For disjoint subsets, probability of union of subsets is sum of subset probabilities {sigma-additivity} {additivity}.

large numbers law
The more times event repeats, the closer to actual probability outcome-probability becomes {large numbers law} {law of large numbers}.

Monte Carlo fallacy
People can believe that an improbable situation that has not happened recently is more likely to happen now, or that the past affects next outcome of random process {Monte Carlo fallacy} {gambler's fallacy}.

risk in probability
Expected outcome divided by outcome value measures risk {risk, outcome}. Expected outcome value is worth or gain multiplied by probability.

weak law of large numbers
After many independent events, relative frequency approaches outcome probability {weak law of large numbers}.

MATH>Statistics>Probability>Events

birthday probability
Probability is 0.5 that at least two people out of 25 have same birthday {birthday, probability}.

Buffon problem
What is probability that needle falls on parallel lines {Buffon's problem} {Buffon problem}? Probability that needle touches line is 2 * l / (pi * d), where l is needle length and d is distance between lines.

Monty Hall problem
First, contestant chooses one of three outcomes. Then contestant learns that one of other two outcomes is incorrect. Do people switch to another outcome or keep current choice {Monty Hall problem}? Switch, because first guess was only 1/3 correct, but second guess must be 2/3 correct, because contestant now knows that probability of one outcome is zero.

MATH>Statistics>Randomness

middle-square method
To generate pseudorandom number sequence, square number and use result without first and last digits, then repeat {middle-square method}.

Monte Carlo method
Sampling from tables of random digits {Monte Carlo method} can determine empirical probability.

MATH>Statistics>Statistical Error

systematic error
Data-collection methods can cause errors {systematic error}.

Gauss-Laplace law
Normal distribution indicates error frequencies {Gauss-Laplace law}.

MATH>Statistics>Statistical Population

statistical population
    Statistics find information about group {population, statistics} {statistical population}. Population has number of people or objects, which have measurable properties {statistic, data} {descriptive statistic} {datum}. Experimenters can check population sample or collect information from all population members {census, population}.

central limit theorem
    If many random same-size samples come from a population with normally distributed measurements, sample sums and means are both normal distributions {central limit theorem, population}.

degrees of freedom in statistics
    Samples have numbers {degrees of freedom, statistics}| of independent values, free to change. Degrees of freedom equal value number minus one, because last value is total minus other values and so is dependent. If situation has factors, factor degrees of freedom are factor number minus one, because last factor depends on the others.

sampling from population
    Taking {sampling}| one value {sample, population} from a population can be random {random sample}. Sample sets {aggregate, population} have number of samples {sample size}.
attributes
    Samples have properties {parameter, population} {sampling statistic}, such as weight. Samples have property values {attribute data} {measurement data}, such as high weight. Sampling can infer {statistical inference} population mean, variance, mode, and median.
types
    Random samples can have same subgroup proportions as population {stratified random sample}, such as same age distribution.
    Sampling can return sampled member to population {sampling with replacement} or not {sampling without replacement}.
subsets
    Sample subsets {sample class} share an attribute, such as high weight.

trend statistics
    Data can have linear changes over time {trend}| {trend line}.

weighting
    Numbers can have coefficients {weight, multiplier} based on frequency or importance {weighting}|. Averages {weighted average} can use weights.

MATH>Statistics>Statistical Population>Estimation

estimation
    Means can have an estimate {estimate} {estimation}|.

confidence interval
    Intervals {confidence interval}| around estimated means have a confidence level that population mean is in the interval. Unbiased estimate can have 95% confidence if low estimate equals $x - (1.96 * s) / N^{0.5}$ and high estimate equals $x + (1.96 * s) / N^{0.5}$, where x is mean, s is standard deviation, and N is sample size.

expected value
    On average, situations have sets of expected outcomes {expectation, statistics} {expected value}|. Averaging outcome values can find expected value: sum from $i = 1$ to $i = N$ of $(w(i) * p(i)) / N$, where N is number of values, $w(i)$ is outcome worth, and $p(i)$ is outcome probability.

Stein paradox

Average is not the best estimate {Stein's paradox} {Stein paradox}. Best estimate is average plus factor times difference between average and grand average: e = u * f * (u - U). Factor depends on standard deviation, as shown by Bayes.

unbiased estimate
Estimates have population, which sets correct confidence interval {unbiased interval}. Sample means are unbiased population-mean estimates. Sample variances are unbiased population-variance estimates.

MATH>Statistics>Statistical Population>Percentile

percentile
Percent of numbers that are less than a number plus half of percent of numbers equal to the number is a statistic {percentile rank} {percentile}|. For example, if n is at 50th percentile {second quartile, percentile} {fifth decile}, half of all values are less than or equal to n.

inter-quartile range
Dispersion {inter-quartile range} can be between lower quartile 25% and upper quartile 75%. Median splits inter-quartile range.

quantile
Value has cumulative probability {quantile} {inverse density function} {percent point function} for that value and all lower values. Quantile function is integral of probability function from minimum to value.

quartile
For distributions, one quarter {quartile} goes from 0th to 25th percentile {first quartile}. One quarter goes from 25th to 50th percentile {second quartile, distribution}. One quarter goes from 50th to 75th percentile {third quartile}.

quartile deviation
Deviation {quartile deviation} (Q) can be half the difference between first and third quartiles.

MATH>Statistics>Statistical Population>Statistic

statistic
Population measurement, such as weight, has calculated numbers {statistic, population}, such as median, mode, mean, and range.

coefficient of variability
Standard deviation divided by mean {coefficient of variation} {coefficient of variability} {variation coefficient} can measure population variation.

mean
Sum from i = 1 to i = N of n(i) /N, where N equals number of numbers, and n(i) equals number, is a statistic {mean, population}| {average}. For example, the numbers 1, 2, 2, 3, 4, 5, and 6 have mean equal to (1 + 2 + 2 + 3 + 4 + 5 + 6) / 7. Average is number-group balance point, because sum of differences between numbers and mean equals zero.

median statistic
If numbers are in sequence, middle number of odd number of numbers, or average of two middle numbers of even number of numbers, is a statistic {median, population}|. For example, the numbers 1, 2, 2, 3, 4, 5, and 6 have median equal 3.

mode statistic
Number with greatest frequency is a statistic {mode, population}|. For example, the numbers 1, 2, 2, 4, 5, and 6 have mode equal 2.

range statistic

Difference between lowest and highest number is a statistic {range, number}. For example, the numbers 1, 2, 2, 3, 4, 5, and 6 have range equal 5.

relative error

Mean can divide into quotient error {relative error, statistics}.

MATH>Statistics>Statistical Population>Statistic>Moment

dispersion in statistics

Number sets have variance spread {dispersion, statistics}|. Dispersion is torques of numbers around balance point: sum from $i = 1$ to $i = N$ of $(n(i) - x)^2 / N$, or sum from $i = 1$ to $i = N$ of $(n(i))^2 / N - x^2$, where $n(i)$ are numbers, and x equals mean.

kurtosis

Fourth moment {kurtosis, distribution} measures distribution fatness or slimness.

root mean square

Square root of mean of squares of differences between numbers and mean {root mean square} (RMS) can equal standard deviation.

skewness

Third moment {skewness, distribution} measures distribution asymmetry, whether it is more to right or left of mean. Skew distribution is not symmetric. To find skewness, calculate median and compare to mean.

standard error of the mean

Sample-mean distribution standard deviation {standard error of the mean}| is smaller than population standard deviation: $s / N^{0.5}$, where s is population standard deviation, and N is sample size.

standard deviation

Variance has a square root {standard deviation}|.

variance in statistics

Torques of numbers around balance point measure dispersion {variance, distribution}|: sum from $i = 1$ to $i = N$ of $(n(i) - x)^2 / N$, or sum from $i = 1$ to $i = N$ of $(n(i))^2 / N - x^2$, where $n(i)$ are numbers, and x equals mean.

MATH>Statistics>Tests

statistical tests

Using a parameter {statistic, test} can test {test, statistics} {statistical tests} whether two sample groups from population are similar.

statistic

Statistic can be mean or variance.

hypothesis

Hypothesis is false if it leads to contradiction, and then hypothesis opposite is true. Assume groups have no difference for statistic {null hypothesis, statistical test}. Try to reject null hypothesis.

significance

Choose allowable probability {significance level, test}, usually 5%, that rejected hypothesis is actually true {type-one error, test} or that accepted hypothesis is actually false {type-two error, test}.

calculation

Calculate probability that sample group is from population.

comparison

Reject hypothesis if probability is less than 5%. The statistic {two-tailed test, error} can be too high and/or too low, the usual case. The statistic {one-tailed test} can be too low or too high. It is hard to prove hypotheses true, but hypotheses are false if they lead to contradiction, thus proving hypothesis opposite.

hypothesis testing

Statistics can test statement {hypothesis, statement} about population {hypothesis testing}.

To test, take sample. Sample can be from hypothesized population or not. Sample can be represent hypothesized population or not. If sample is from hypothesized population and represents it, hypothesis is true. If sample is not from hypothesized population and does not represent it, hypothesis is false.

errors

If sample is from hypothesized population but does not represent it, hypothesis is true but seems false. Errors {Type I error} can happen, with probability {alpha risk}.

If sample is not from hypothesized population and represents it, hypothesis is false but seems true. Errors {Type II error} can happen, with probability {beta risk}.

independence

Perhaps, events, such as political party, job type, sex, or age category, do not affect each other {independence, statistics}. Tests can see if they do affect each other {dependence, statistics}.

analysis of variance

Testing variable found in mutually exclusive groups from same population, normally distributed or not, can show if groups are equivalent {analysis of variance} {variance analysis} (ANOVA).

measurements

In ANOVA, measurements are sums of four parts: mean, class or treatment effect, sampling or measurement effect, and normally distributed random error. ANOVA checks if sampling error or class error is great enough compared to random error to make samples or classes actually different.

process

Assume groups are equivalent and so have same mean. Set significance level to 5%.

For two groups, calculate variance ratio {variance ratio} {F value, ANOVA}. Group degrees of freedom are group number minus one: $C - 1$. Sample degrees of freedom df are sum of group degrees of freedom N_i, minus group number C. $df = N_1 + N_2 + ... - C$.

Numerator is sample-mean-difference variance: $((\text{sum from } i = 1 \text{ to } i = N_1 \text{ of } (n_1(i))^2) / N_1 - (\text{sum from } i = 1 \text{ to } i = N_1 + N_2 + ... \text{ of } (n(i))^2) / (N_1 + N_2 + ...)) / (C - 1) + ((\text{sum from } i = 1 \text{ to } i = N_2 \text{ of } (n_2(i))^2) / N_2 - (\text{sum from } i = 1 \text{ to } i = N_1 + N_2 + ... \text{ of } (n(i))^2) / (N_1 + N_2 + ...)) / (C - 1) + ...$ Denominator is population variance: $((\text{sum from } i = 1 \text{ to } i = N_1 + N_2 + ... \text{ of } (n(i))^2) - ((\text{sum from } i = 1 \text{ to } i = N_1 \text{ of } (n_1(i))^2) / N_1 + (\text{sum from } i = 1 \text{ to } i = N_2 \text{ of } (n_2(i))^2) / N_2 + ...)) / (N_1 + N_2 + ... - C)$, where n is frequency, N is sample size, and C is sample number.

F values form distributions {F distribution, ANOVA} that vary with degrees of freedom and significance.

If calculated F value is less than F value in F distribution for same degrees of freedom and significance level, accept that samples have same mean. If calculated F value is more, reject hypothesis, so at least one sample is not random, or samples are from different populations.

mean square

For samples or treatments, degrees of freedom can divide into sum of squares of differences between values and mean {mean square}. Mean square estimates population variance. F value is sample or treatment mean square divided by error mean square.

missing data

Least-squares method can estimate missing data.

types

Replications are like classes {randomized blocks plan}.

Testing interactions between treatments can be at same time as testing interactions between samples {two-way classification}.

comparison to t test

With two samples, F test and t test are similar.

chi square

In populations, tests {chi square} can show whether one property, such as treatments, affects another property, such as recovery level. Properties have mutually exclusive outcomes, such as cured or not.

process

Hypothesize that events are independent. Select significance level, typically 5%. Make contingency table.

Calculate degrees of freedom: $(R - 1) * (C - 1)$, where R is number of table rows and C is number of table columns.

Calculate chi square value: $X = \text{sum from } i = 1 \text{ to } i = R \text{ and from } j = 1 \text{ to } j = C \text{ of } (x(i,j) - f(i,j))^2 / f(i,j)$, where $x(i,j)$ is observed frequency and $f(i,j)$ is expected frequency. $f(i,j) = (\text{sum from } i = 1 \text{ to } i = R \text{ of } x(i,j)) * (\text{sum from } j = 1 \text{ to } j = C \text{ of } x(i,j)) / (\text{sum from } i = 1 \text{ to } i = R \text{ and from } j = 1 \text{ to } j = C \text{ of } x(i,j))$, where $x(i,j)$ is observed frequency.

result

If calculated chi square value is less than actual chi square value for degrees of freedom at significance level, do not reject hypothesis.

contingency table

Table rows can be possible outcomes of one variable, and table columns can be possible outcomes of other variable {contingency table}. Table cells are numbers {observed frequency} having both outcomes. For example, table rows can be Wide and Narrow people, and table columns can be Tall and Short people, so table 0 1 / 2 3 has 0 of Wide and Tall and 3 of Narrow and Short.

F test

Testing two same-size samples, each from different and not necessarily normally distributed populations, can show if populations are the same {F test}. Hypothesize that samples are the same. Set significance level to 5%. Sample degrees of freedom are sample size minus one. Calculate variance ratios between two samples {F distribution, test} {F value, test}: v1 / v2, where v is sample variance. If calculated F value is less than actual F value for degrees of freedom at significance level, do not reject hypothesis. F distribution measures variance distribution.

factorial experiment

Experiments {factorial experiment} can use factors. Several variables {factor, variable} can affect dependent variable. Set up ANOVA. Calculate factor effects, by finding average of differences between variable measurements, while holding other factors at their constant means. Calculate factor interactions, by finding differences between variable-measurement changes, at varying other-variable levels. Small differences mean little interaction. For no interactions, factor effects can add.

hierarchical classification

ANOVA {nested classification} {hierarchical classification} can have sample treatments and classes.

Latin square

Randomizing treatments over replications {Latin square} can control two variation sources.

mixed model ANOVA

ANOVA {mixed model ANOVA} can have no treatments or classes, with sample subsamples.

Model I ANOVA

ANOVA {fixed effect model} {Model I analysis of variance} can have no treatments or classes and only replicate samples.

Model II ANOVA

ANOVA {random effect model} {Model II analysis of variance} can have one random sample from any number of treatments or classes.

non-parametric test

Tests {distribution-free test} {non-parametric test}, such as sign tests, can be for unknown distributions. Sign and difference value can make a non-parametric test {sign rank test}.

null hypothesis

Assume groups have no difference for parameter {null hypothesis, test}|.

sequential analysis

Methods {sequential analysis} can test paired attribute data to decide between classes/treatments. For all pairs, check treatment differences. Count only significant differences. Observe until number of accepted counted differences exceeds number required by significance level and total number of observations, or until total number of observations exceeds threshold. If difference number is greater than threshold, accept that one treatment is better.

sign test

Testing two samples with matched pairs {sign test} can show if they are from same, not necessarily normally distributed, population.

matched pairs

Before and after comparison has matched pairs.

hypothesis

Hypothesize that first and second samples show no change.

significance

Set significance level to 5%.

degrees of freedom

Degrees of freedom are sample size minus one.

calculation

For all matched pairs, subtract before-value from after-value to get plus, minus, or no sign. Add positive signs.

Use probability of 1/2 for getting positive sign, because samples are the same by hypothesis. Calculate binomial-distribution z score: $(P - 0) / (N * 0.5 * 0.5)^{0.5}$, where P is positive-sign number, N is sample size, mean equals zero, and standard deviation equals $(N * p * (1 - p))^{0.5}$.

test

If z score is less than normal-distribution value with same degrees of freedom at significance level, do not reject hypothesis.

significance level

Choose allowable probability {significance level, statistics}|, usually 5%, that rejected hypothesis is actually true {type-one error, significance} or that accepted hypothesis is actually false {type-two error, significance}.

t test

Sample can test the hypothetical mean of normally distributed population {t test} {one-sample t test}. Hypothesize that sample and population means are equal. Set significance level to 5%. Sample size less one gives independent-value number {degrees of freedom, t test}. Calculate distribution of same-size-sample means with same degrees of freedom. Result is similar to normal distribution, except distribution includes degrees of freedom {t value} {t distribution}: $t = (x - u)/e$, where x is sample mean, u is hypothetical population mean, and e is sample-mean standard error. If calculated t value is less than actual t value for significance level and degrees of freedom, do not reject hypothesis.

two samples

Testing two independent samples from population can show if samples are from same population. Hypothesize that first and second sample means are equal. Set significance level to 5%. Degrees of freedom involve both sample sizes: $(N1 - 1) + (N2 - 1) = N1 + N2 - 2$. Calculate t value: $t = (x1 - x2)/e$, where x is sample mean. e is standard error of difference, which equals $(((v1 * (N1 - 1) + v2 * (N2 - 1)) / (N1 + N2 - 2))^{0.5}) * ((1 / N1 + 1 / N2)^{0.5})$, where v is sample variance and N is sample degrees of freedom. If t value is less than t-distribution value with same degrees of freedom at significance level, do not reject hypothesis.

paired samples

Testing two paired samples, or matched pair samples, can show if they are from same population. Hypothesize that first and second sample means are equal. Set significance level to 5%. Degrees of freedom are sample size minus one. Calculate t value: $t = $ sum from $i = 1$ to $i = N$ of $(n1 - n2)/e$, where N is sample size and n is sample value. e is standard error of difference, which equals $(N * ($ sum from $i = 1$ to $i = N$ of $(n1 - n2)^2) - ($ sum from $i = 1$ to $i = N$ of $(n1 - n2)^2)) / (N - 1)^{0.5}$. If t value is less than t-distribution value with same degrees of freedom at significance level, do not reject hypothesis.

two-tailed test

Statistic can be too high and/or too low {two-tailed test, statistics}, the usual case.

type-one error

Rejected hypothesis can be actually true {type-one error, statistics}.

type-two error

Accepted hypothesis can be actually false {type-two error, statistics}.

z test

Testing {z test} an outcome that recurs in successive events can show if it is a true outcome.

set up

Assume outcome is true outcome. Set significance level to 5%. Degrees of freedom are sample size minus one. Add number of times outcome happens. Use probability, usually 0.5, suggested by null hypothesis for outcome.

test

Calculate z score for binomial distribution: $(P - 0) / (N * p * (1 - p))^{0.5}$, where P is outcome number, p is probability, N is event number, mean equals zero, and standard deviation equals $(N * p * (1 - p))^{0.5}$. If z score is less than normal-distribution value with same degrees of freedom at significance level, do not reject hypothesis.

MATH>Statistics>Theories

Fisher theory

Theories {Fisher theory} can be about experiment design.

Neyman-Pearson theory

Theories {Neyman-Pearson theory} can be about hypothesis testing.