|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Nguyen Ngoc Linh Chi – Machine Learning Cheat-sheet**  *METRIC: unit of measurement providing way to objectively quantify performance*  *Measurement: act of obtain data associated vs metric*  *MEASURES: numerical values associated with metric.*  **Data classification by Measurement scales:**  CATEGORICAL (NOMINAL) DATA - sorted into categories according to specified characteristics.  ORDINAL DATA - can be ordered or ranked according to some relationship to 1 another.  INTERVAL DATA - ordinal but have constant differences between observations and have arbitrary zero points.  RATIO DATA - continuous and have natural zero.  **DATA RELIABILITY AND VALIDITY:**   |  |  |  | | --- | --- | --- | | (1st) RELIABILITY: Data is accurate and consistent  (2nd) VALIDITY: Data correctly measures what it is supposed to measure. | | | | A tire pressure gage that consistently reads several pounds of pressure below true value | NO | YES | | Number of calls to customer service desk (counted correctly) used to assess customer dissatisfaction | YES | NO | | Customer rating on food quality is used to assess customer satisfaction | NO | NO |   **PERCENTILES:** kth percentile is value at or below which at least k percent of observations lie.  COMPUTING PERCENTILES:  • Find kth percentile for variable in sample size n  • Rank of  **QUARTILES:** BREAK DATA INTO 4 PARTS  25th percentile, Q1; 50th percentile, Q2;  75th percentile, Q3; and 100th percentile, Q4.  **VARIANCE** ~ average of squared deviations from mean. If sample data is also population data, then n = N to compute population variance  **STANDARD DEVIATION** ~ square root of variance (popular measure of risk)  **CHEBYSHEV’S THEOREM**: For any data set, proportion of values that lie within standard deviations of mean are at least  Why is this useful? Able to use mean and standard deviation to find percentage of total observations that fall within given interval about mean  **EMPIRICAL RULES:** For normally distributed data set, proportion of values that lie within k (k > 1) standard deviations of mean follow empirical rules:  Application of Empirical Rule - Process Capability Index is measure of how well manufacturing process can achieve specifications  Using sample of output, measure dimension of interest, and compute total variation using third  empirical rule.  **STANDARDIZED VALUES, Z-SCORE**, provides relative measure of distance observation is from mean (independent of units of measurement)  **COEFFICIENT OF VARIATION** *provides relative measure of dispersion in data relative to mean:*  Provides relative measure of risk to return  Useful when comparing variability of two or more data sets with different scales  Smaller CV 🡪 smaller risk  Reciprocal of CV 🡪 return to risk  **MEASURES OF DISPERSION:** Dispersion refers to degree of variation (numerical spread/ compactness)  RANGE data values  INTERQUARTILE RANGE IQR difference between first and third quartiles: (use 50% data)  **MEASURES OF SHAPE: SKEWNESS:**  Skewness describes lack of symmetry of data.  COEFFICIENT OF SKEWNESS (CS):  left-skewed data; right-skewed  suggests high degree of skewness.  suggests moderate skewness.  suggests relative symmetry  **SHAPE AND MEASURES OF LOCATION**  • *Symmetrical*, *unimodal*,  • *Negatively* *skewed (left skewed, tails off toward the right)*,  • *Positive* *skewed (right skewed, tails off toward the left)*,  **KURTOSIS** refers to (i.e., high, narrow) or flatness (i.e., short, flat-topped) of histogram.  COEFFICIENT OF KURTOSIS (CK): measures degree of kurtosis of population  CK < 3 🡪 data is flat + wide degree of dispersion.  CK > 3 🡪 data is peaked with less dispersion | **DESCRIPTIVE STATISTICS FOR CATEGORICAL DATA:**  PROPORTION *is fraction of data that have certain characteristic, are key descriptive statistics for categorical data, i.e. defects or errors in quality control applications or consumer preferences in market research.*  **COVARIANCE** is measure of linear association between two variables, X and Y.  POSITIVE covariance 🡪 direct relationship  NEGATIVE covariance 🡪 inverse relationship  Magnitude 🡪 degree of association  **CORRELATION** is measure of linear association between two variables, X and Y (not dependent on units of measurement)  RANGE: (Strong negative) and 1 (Strong positive linear relationship); 0 indicates no linear relationship;  Also known as: Pearson product moment correlation or Pearson's correlation coefficient  **OUTLIERS:** Mean and Range are sensitive to outliers  HOW DO WE IDENTIFY POTENTIAL OUTLIERS?  Extreme outliers are to left or right Q3  Mild outliers are between to left of Q1 or right of Q3  WHAT DO YOU DO WITH OUTLIERS?  Leave them in data if it is important  Remove them if they are different from rest  Correct error in data entry  **PROB. DIST.:** characterization of possible values that random variable may assume along with prob. of assuming these values  3 PERSPECTIVES FOR DEVELOPING  • theoretical arguments  • empirical data – empirical prob. dist.  • using subjective values and expert judgement  WHY DO WE NEED TO KNOW ABOUT DISTRIBUTION?  Helps you to understand underlying process that generates sample data. Useful in building decision models with theoretical dist. of data. Helps to compute probabilities of occurrence of outcomes to assess risk and make decisions  **NORMAL DIST.:**  is bell-shaped curve  Properties: **1**. Symmetric; **2**. Mean = Median = Mode;  **3**. Range of X is unbounded;  **4.** Empirical rules apply (i.e., area under density function within ± 2 standard deviation is 95.4%, within ±3 standard deviation is 99.7%)  STANDARD NORMAL DIST. is normal dist. with mean of 0 and standard deviation of 1, called Z. Scale along z-axis represents number of standard deviations from mean of  zero  **Data Modelling and Dist. Fitting**: Sample data limits our ability to predict uncertain events; potential values outside range of sample data are not included; better to identify underlying prob. dist. from which sample data come by “fitting” theoretical dist. to data and verifying goodness of fit statistically  **Goodness of Fit:** fitting data to prob. dist.  CHI-SQUARE (need at least 50 data points)  KOLMOGOROV-SMIRNOV (works well for small samples and only for non-parametric data)  ANDERSON-DARLING (puts more weight on differences between tails of dist.)  SHAPIRO’S WILKINS NORMALITY TEST (test data against normal dist.) 🡪 P-value > 0.05 implies that dist. of data is not significantly different from normal dist. In other words, we can assume data is normal.  **Statistical Sampling**  SAMPLING: foundation of statistical analysis.  ESTIMATORS: measures used to estimate unknown population parameters  POINT ESTIMATE: single number derived from sample that is used to estimate value of population parameters  **Sampling Error:** SAMPLING (STATISTICAL) ERROR: samples are only subset of total population  SAMPLING ERROR depends on size of sample relative to population.  NON-SAMPLING ERROR: sample does not adequately represent target population, results from poor sample design or choosing wrong population frame. (e.g., convenience sample)  **A Sampling Experience:** Sample size increases, average of sample means is all still close to expected value;  Standard deviation of sample means becomes smaller, meaning that means of samples are clustered closer together around true expected value.  **Sampling Dist.:** Sampling dist. of mean is dist. of means of all possible samples of fixed size n from some population. Standard deviation of sampling dist. of mean is called standard error of mean  🟑As n increases, standard error, sampling error . | **Estimating sampling error empirical rule**  Using empirical rule for 3 standard deviations away from mean, ~99.7% of sample means should be between:    ;  🟑 n increases, standard error, sampling error .  **CENTRAL LIMIT THEOREM**  If sample size is large enough, then sampling dist. of mean ~normally distributed regardless of dist. of population ~ sample mean = population mean  If population ~ normally distributed, sampling dist. ~ normal distr. for any sample size.  **INTERVAL ESTIMATES:**  prob. interval is any interval [A, B] such that prob. of falling between and B is . prob. intervals are centred on mean/ median.  **CONFIDENCE INTERVALS** is range of values between which value of population parameter is believed to be, along with prob. that interval correctly estimates true (unknown) population parameter.  is called level of confidence  *For 95% confidence interval, if we chose 100 different samples, leading to 100 different interval estimates, we would expect that 95% of them would contain true population mean.*  **C.I. for Mean with KNOWN Pop. Std Deviation**  Example: *Standard deviation in filling volumes is constant at 15 . sample of 25 bottles revealed mean filling volume of 796 . 95% confidence interval estimate of mean filling volume for*  *population is*  **C.I. for Mean UNKNOWN Pop. Std Deviation**  Find 95% confidence interval estimate of mean revolving balance of homeowner applicants. Sample mean = $12,630.37; s = $5393.38; standard error = $1037.96;  **Confidence Interval for Proportion**  (sample proportion), where x is number in sample having desired characteristic and n is sample size.  Computing Confidence Interval for Proportion  **USING C.I. FOR DECISION MAKING:**  **1**. Required volume for bottle-filling process is 800 and sample mean is 796 mls. We obtained confidence interval for population mean of [790.12, 801.88]. Should machine adjustments be made? Although sample mean is less than 800, sample does not provide sufficient evidence to draw that conclusion that population mean is less than 800 because 800 is contained within confidence interval.  **2**. 1,300 voters found that 692 voted for particular candidate in two-person race. This represents proportion of 53.23% of sample.  Could we conclude that candidate will likely win election? 95% confidence interval for proportion is [0.505, 0.559] This suggests that population proportion of voters who favour this candidate is highly likely to exceed 50%, so it is safe to predict winner.  **3.** What if sample proportion is 0.515, and confidence interval for population proportion is [0.488, 0.543]? Even though sample proportion is larger than 50%, sampling error is large, and confidence interval suggests that it is reasonably likely that true population proportion could be less than 50%, so you cannot predict winner.  **PREDICTION INTERVALS** is 1 that provides range for predicting value of new observation from same population.  While confidence interval is associated with sampling dist. of statistic, but prediction interval is associated with dist. of random variable itself.  Compute 95% prediction interval for revolving balances of customers.  Sample mean = $12,630.37; s = $5393.38; standard error = $1037.96; = 2.056. | **CONFIDENCE INTERVALS and SAMPLE SIZE**  Determine appropriate sample size needed to estimate population parameter within specified level of precision .  Sampling error was ± 5.88 mls, sd=15. What sample size is needed to reduce margin of error to at most 3 mls?    Suppose that we wish to determine number of voters to poll to ensure sampling error of at most ± 2%. With no information, use = 0.5 (proportion who poll):  Use sample proportion from preliminary sample as estimate of or set = 0.5 for conservative estimate to guarantee required precision  **HYPOTHESIS TESTING**  Null hypothesis: What you do not want to see  Alternative hypothesis: what you want to see  HYPOTHESIS TESTING PROCEDURE:  Step 1. Identify population parameter and formulate hypotheses to test.  Step 2. Select level of significance (risk of drawing incorrect conclusion).  Step 3. Determine decision rule on which to base conclusion.  Step 4. Collect data and calculate test statistic.  Step 5. Apply decision rule, draw conclusion.  *Equality part of hypotheses is always in null hypothesis 1 sample hypothesis test*  **One-Sample Hypothesis Tests**      • α is typically set to 0.01 (strong), 0.10 (weak) or 0.05 (commonly accepted) 🡪 value of α can be controlled.  • value of β cannot be specified in advance and depends on (unknown) population parameter.  • Generally, as α decreases, β increases.   |  |  |  | | --- | --- | --- | |  | is actually | | |  | FALSE | TRUE | | Reject | Correct | TYPE I error  () (1) | | Accept | Type II error ( (2) | Correct |   *(1): Charge person guilty by mistake*  *(2): Person is guilty but lack of evidence to proof.*  **IMPROVING POWER OF TEST**  Power of test  • prob. of not committing type II error  • should be high to make valid conclusion  How to ensure sufficient power?  🟑Power of test is sensitive to sample size  🟑small sample sizes 🡪 low power  🟑large sample required for small ⍺  **Selecting Test Statistic:**  *if data is population, if data is sample*  Decision to reject or fail to reject null hypothesis is based on computing test statistic from sample data.        **REJECTION REGION and P-VALUE for**  **VALUE:** If , do not reject , else reject  **NORMAL DISTRIBUTION:**   |  |  |  | | --- | --- | --- | |  | *Rejection region* | *p-value* | |  |  |  | |  |  |  | |  | ; |  |   **DISTRIBUTION:**   |  |  |  | | --- | --- | --- | |  | *Rejection region* | *p-value* | |  |  |  | |  |  |  | |  | ; |  | | **T-TEST:** Paired two-sample for means  **TEST FOR EQUALITY OF VARIANCES** between two samples using new type of test, F-test.  • To use this test, we must assume that both samples are drawn from normal populations.  •  • F-test statistic:  **F-DIST.** has two degrees of freedom, 1 associated with numerator of F-statistic, , and 1 associated with denominator .  *Population with larger variance will be assigned numerator*  **ANALYSIS OF VARIANCE (ANOVA):** Used to compare means of two or more population groups; fairly robust to departures from normality  at least 1 mean is different from others  • ANOVA measures variation between groups relative to variation within groups.  • Test statistic has F-dist. so if F-statistic is large enough based on level of significance chosen and exceeds critical value, we would reject .  **ANOVA Assumptions:** Independence, Normality, and homogeneity of variances:  **1.** Randomly and independently obtained (validated if random samples are chosen)  **2.** Normally distributed;  **3.** Have equal variances  *If sample sizes are equal, violation of third assumption does not have serious effects, but with unequal sample sizes, it can.*  *Comparing sample means of two populations, use t-test rather than ANOVA*  **TREND LINES and REGRESSION ANALYSIS:**  **REGRESSION ANALYSIS** is tool for building mathematical and statistical models that characterize relationships between dependent (ratio) variable and 1 or more independent, or explanatory variables (ratio or categorical), all of which are numerical.  *Simple linear regression involves single independent variable* >< *multiple linear regression*  **RESIDUALS** are observed errors associated with estimating value of dependent variable using regression line:  Errors associated with individual observation  **REGRESSION STATISTIC: Multiple** , where r is sample correlation coefficient. r varies from -1 to +1 (r is negative if slope is negative).  **(R-squared)** is measure of “fit” of line to data. value of will be between 0% and 100%. *A value of 1.0 indicates perfect fit and all data points would lie on line; larger value of better fit.*  value , order of polynomial ;  **ADJUSTED R SQUARE** - adjusts for sample size and number of X variables.  **STANDARD ERROR** - variability between observed and predicted Y values. This is formally called standard error of estimate, .  **Why use adjusted R Square?** R-squared has additional problems that adjusted R-squared is designed to address. Problem 1: Add predictor to model, R-squared increases, even if due to chance alone. It never decreases 🡪 model with more terms appear to have better fit simply because it has more terms. Problem 2: If model has too many predictors and higher order polynomials, it begins to model random noise in data. This is known as overfitting model and it produces misleadingly high R-squared values and lessened ability to make predictions. adjusted R-squared increases only if new term improves model more than would be expected by chance. It decreases when predictor improves model by less than expected by chance. adjusted R-squared can be negative, but it’s usually not, always  **Regression as analysis of variance:** ANOVA conducts F-test to determine whether variation in Y is due to varying levels of X.  ANOVA test for significance of regression:  : population slope coefficient = 0  : population slope coefficient ≠ 0  **Notice p-value (Significance F):** When p-value is less than threshold (significance level), justifies rejection of null hypothesis.  Null hypothesis is rejected when p < 0.05 and not rejected when p > 0 .05. Rejecting indicates X explains variation in Y  **Residual analysis and Regression assumption**  Residual = Actual Y value - Predicted Y value  Standard residual = residual/standard deviation  Rule of thumb: Standard residuals outside of ±2 or ±3 are potential outliers. | **Four Major Assumptions of Simple / Multiple Linear Regression:**  1. Linearity (of relationship between Y & Xs): Residual vs. fitted - *Find straight horizontal line*  2. Normality of Errors = Errors (e; residuals) are normally distributed:  Normal Q-Q plot - Look for linear relationship  3. Homoscedasticity = Constant / Equal variance of errors (e) for all values of X = Impact of X on Y is same for all X values:  Residual vs. fitted; Scale-location - Look for straight horizontal line  4. Independence of errors = There is no correlation between errors (e) calculated from regression model - Need additional plot/test  \* Residual time series plot  \* Durbin-Watson test  • For cross-sectional data, this is usually not major issue  • Panel/time-series data need to check  • Issues 2, 3, and 4 are often interrelated  • Cross-sectional data – data is collected only once, from different individuals/entities  • Panel/time-series data – data is collected multiple times from each individual/entity  **Multicollinearity**: occurs when there are strong correlations among independent variables, and they can predict each other better than dependent variable. Becomes difficult to isolate effect of 1 independent variable on dependent variable, signs of coefficients may be opposite of what they should be, making it difficult to interpret regression coefficients, and p-values can be inflated. Correlations exceeding may indicate multicollinearity  **Overfitting**: fitting model too closely to sample data at risk of not fitting it well to population in which we are interested.  –value will increase if we fit higher order polynomial functions to data 🡪 make it difficult to explain phenomena rationally.  In multiple regression, if we add too many terms to model, then model may not adequately predict other values from population.  Overfitting can be mitigated by using good logic, intuition, theory, and parsimony  **Principle of parsimony:** Good models are as simple as possible  **Interactions:** occurs when effect of 1 variable is dependent on another variable. We can test for interactions by defining new variable as product of two variables, , and testing whether this variable is significant, leading to alternative model.  Difference between correlation and interaction: Whether two variables are associated says nothing about whether they interact in their effect on third variable. interaction between two variables means effect of 1 of those variables on third variable is not constant— effect differs at different values of other.  **FORECASTING TECHNIQUES:**  **Qualitative and Judgmental techniques** rely on experience and intuition.  **Historical analogy approach** obtains forecast through comparative analysis with prior situations; **Delphi method** questions anonymous panel of experts 2-3 times in order to reach convergence of opinion on forecasted variable; **Indicators** are measures that are believed to influence behaviour of variable we wish to forecast. Indicators are often combined quantitatively into **index**, single measure that weights multiple indicators, thus providing measure of overall expectation; **Leading indicators:** series of measure change before variable change; **Lagging indicators:** series of measures that follow change of variable.  **STATICALLY FORECASTING MODELS:**  ***Time Serie***–*stream of historical data, daily*  Have components such as:  **1.** random behaviour; **2.** trend: is gradual upward or downward movement of time series; **3.** seasonal effects: is 1 that repeats at fixed intervals of time, typically year, month, …; **4.** cyclical effects: describe ups and downs over much longer time frame, i.e. several years  Stationary time series have only random behaviour. | **MOVING AVERAGE MODEL:** smoothing method based on idea of averaging random fluctuations in time series to identify underlying direction in which time series is changing. Simple moving average forecast for next period is computed as average of most recent k observations. Larger values of k result in smoother forecast models since extreme values have less impact  **EXPONENTIAL SMOOTHING MODEL:**  **Simple**  is called smoothing factor/ coefficient/ constant.  Value of dictates how much weight is given to most recent observed value versus last expected value;  regulates importance of most recent observations with respect to smoothed mean of previous values;  : *assign an almost constant weight to all past observations;*  : *assign an almost constant weight to all recent observations.*  **Double** *Rewrite simple exponential smoothing:*          Predicted value is a function of last estimates of level linear trend  modulates importance of most recent value of trend; with respect to smoothed trend of previous periods.  : *assign almost weight to trends in past*  : *most recently exhibited trend is pre-dominant.*  **Regression-based forecasting for Time Series with Linear trend:** Simple linear regression can be applied to forecasting using time as independent variable.  **AUTOCORRELATION:** When autocorrelation is present, successive observations are correlated with 1 another; for example, large observations tend to follow other large observations, and small observations also tend to follow 1 another.  In such cases, other approaches, called AUTOREGRESSIVE MODELS, are more appropriate.  **Forecasting time series with Seasonality:** When time series exhibit seasonality, different techniques provide better forecasts than ones we have described:  Multiple regression models with categorical variables for seasonal components;  *HOLT-WINTER MODEL*, similar to exponential smoothing models in that smoothing constants are used to smooth out variations in level and seasonal patterns over time.  **Holt-winter model for forecasting time series Seasonality and Trend:**  Holt-Winters additive model applies to time series with relatively stable seasonality:  is smoothed estimate of level at time  is smoothed estimate of change in trend value at time  is smoothed estimate of appropriate seasonal component at  Holt-Winters multiplicative model applies to time series whose amplitude increases or decreases over time and is  **Regression forecasting with Causal variable:** In many forecasting applications, other independent variables besides time, i.e. economic indexes or demographic factors, may influence time series. Explanatory/causal models, often called econometric models, seek to identify factors that explain statistically patterns observed in variable being forecast, usually with regression analysis  **Practice of forecasting:** Judgmental and qualitative methods are used for forecasting sales of product lines and broad company and industry forecasts. Simple time-series models are used for short and medium-range forecasts. Regression methods are typically used for long term forecasts. | **DATA MINING:**  PARALLEL CO-ORDINATES CHART consists of set of vertical axes, 1 for each variable selected. For each observation, line is drawn connecting vertical axes. point at which line crosses axis represents value for that variable.  *SCATTERPLOT matrix* *combines several scatter charts into 1 panel, allowing user to visualize pairwise relationships between variables.*  A VARIABLE PLOT plots matrix of histograms for variables selected.  **DIRTY DATA:** Real data sets that have missing values or errors, are called “dirty” and need to be “cleaned” before analysing them.  🟑*Approaches for handling missing data.*  *🟑Eliminate records that contain missing data*  *🟑Estimate reasonable values for missing observations, i.e. mean or median value*  *🟑Use data mining procedure to deal with them. XLMiner has capability to deal with missing data in Transform menu in Data Analysis group.*  *🟑Try to understand whether missing data are simply random events or there is logical reason.*  *🟑Eliminating sample data indiscriminately could result in misleading information and conclusions about data*  **CLUSTER ANALYSIS**, also called **DATA SEGMENTATION,** is collection of techniques that seek to group or segment collection of objects (observations or records) into subsets or clusters, such that those within each cluster are more closely related to 1 another than objects assigned to different clusters.  🟑Objects within clusters should exhibit high amount of similarity, whereas those in different clusters will be dissimilar.  **CLUSTER ANALYSIS METHODS:**  **Hierarchical clustering**, data are *not* *partitioned* into *particular* *cluster* in single step. Instead, series of partitions takes place, which may run from single cluster containing all objects to n clusters, each containing single object. *Hierarchical clustering may be represented by two-dimensional diagram known as dendrogram, which illustrates fusions or divisions made at each successive stage of analysis.*  **Agglomerative clustering methods** proceed by series of fusions of n objects into groups.  **Divisive clustering methods** separate n objects successively into finer groupings.  **Agglomerative clustering methods:**  SINGLE LINKAGE CLUSTERING (NEAREST-NEIGHBOR): Distance between groups is defined as distance between closest pair of objects, where only pairs consisting of 1 object from each group are considered. At each stage, closest 2 clusters are merged  COMPLETE LINKAGE CLUSTERING: distance between groups is distance between most distant pair of objects, 1 from each group  AVERAGE LINKAGE CLUSTERING: Uses mean values for each variable to compute distance between clusters  WARD’S HIERARCHICAL CLUSTERING: Uses sum of squares criterion  **CLASSIFICATION METHODS** seek to classify categorical outcome into 1 of two or more categories based on various data attributes.  🟑For each record in database, we have categorical variable of interest and number of additional predictor variables.  🟑*For given set of predictor variables, we would like to assign best value of categorical variable.*  **MEASURING CLASSIFICATION:** Find prob. of making misclassification error and summarize results in classification matrix, which shows number of cases that were classified either correctly or incorrectly.  **USING TRAINING AND VALIDATION DATA:**  Data mining projects typically involve large volumes of data. data can be partitioned into:  ▪ training data set – has known outcomes and is used to “teach” data-mining algorithm  ▪ validation data set – used to fine-tune model  ▪ test data set – tests accuracy of model  **CLASSIFYING NEW DATA:** after classification scheme is chose and best model is developed based on existing data, we use predictor variables as inputs to model to predict output. |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **CLASSIFICATION TECHNIQUES/MODELS:**  **k-NEAREST NEIGHBOURS (K-NN) ALGORITHM**  🟑Finds records in database that have similar numerical values of set of predictor variables  🟑Measure Euclidean distance between records in training data set. nearest neighbour to record in training data set is 1 that that has smallest distance from it.  🟑If , then rule classifies record in same category as its nearest neighbour.  rule finds k-Nearest Neighbours in training data set to each record we want to classify and then assigns classification as classification of majority of k nearest neighbours.  🟑Typically, various values of k are used and then results inspected to determine which is best.  HOW TO CHOOSE VALUE K?  Selecting value of K in K-nearest neighbour is most critical problem. Small value of K means that noise will have higher influence on result i.e., prob. of overfitting is very high. Large value of K makes it computationally expensive and defeats basic idea behind KNN (that points that are near might have similar classes).  Simple approach to select k is  To optimize results, we can use CROSS VALIDATION. We can test KNN algorithm with different values of K. Model which gives good accuracy can be considered to be optimal choice.  **DISCRIMINANT ANALYSIS** *is technique for classifying set of observations into predefined classes.*  Uses predefined classes based on set of linear discriminant functions of predictor variables  Based on training data set, technique constructs set of linear functions of predictors, known as discriminant functions:  are discriminant coefficients (weights), are input variables (predictors), c is constant (intercept)  MAXIMUM NUMBER OF FUNCTIONS = number of groups, or number of variables in analysis, whichever is smaller.  For new observation, each of discriminant functions is evaluated, and observation is assigned to class if discriminant function has highest value.  **LOGISTIC REGRESSION** is variation of linear regression in which dependent variable is categorical. Estimates prob. of belonging to category using regression on predictor variables  Seeks to predict prob. that output variable will fall into category based on values of independent (predictor) variables.  This prob. is used to classify observation into category. Generally used when dependent variable is binary—takes on two values, 0 or 1  **Classification using logistic regression:**  Estimate prob. p that observation belongs to category 1, , and, consequently, prob. that it belongs to category 0, .  Then use , typically 0.5, with which to compare p and classify observation into 1 of two categories.  Dependent variable is called logit, which is natural logarithm of – called odds of belonging to category 1.  form of logistic regression model is  logit function can be solved for p:  **ASSOCIATION RULE MINING**, often called affinity analysis, seeks to uncover associations and/or correlation relationships in large data sets.  Association rules identify attributes that occur together frequently in given data set.  Market basket analysis, for example, is used determine groups of items consumers tend to purchase together.  *Association rules provide information in form of if then (antecedent consequent) statements*  **MEASURING STRENGTH OF ASSOCIATIONS:**  **SUPPORT** for (association) rule is percentage (or number) of transactions that include all items both antecedent and consequent.  **CONFIDENCE** of (association) rule is ratio of number of transactions that include all items in consequent as well as antecedent (namely, support) to number of transactions that include all items in antecedent | |  | | **No Seasonality** | | **Seasonality** | | | | | **LINEAR OPTIMISATION MODELS:**  **Building linear optimization models:**  Step 1. Identify decision variables – unknown values that model seeks to determine.  Step 2. Identify objective function – quantity we seek to minimize or maximize.  Step 3. Identify all appropriate constraints – limitations, requirements, or other restrictions that are imposed on any solution, either from practical or technological considerations or by management policy.  Step 4. Write objective function and constraints as math expressions  **Linear optimization model** (often called **linear program**/**LP**) has 2 basic properties.  1. objective function and all constraints are linear functions of decision variables: This means that each function is simply sum of terms, each of which is some constant multiplied by decision variable.  2. All variables are continuous: This means that they may assume any real value (typically, nonnegative).  **How simplex method works?** simplex method evaluates impact of constraints in terms of their contribution to objective function for each variable. For simple case of only 1 constraint, optimal (maximum) solution is found by simply choosing variable with highest ratio of objective coefficient to constraint coefficient.  **Example 3:** Crebo Manufacturing produces 4 types of structural support fittings. Machining centres have capacity of 280,000 minutes per year. Gross margin/unit and machining:   |  |  |  |  |  | | --- | --- | --- | --- | --- | | Product | Plugs | Rails | Rivets | Clips | | Gross margin/unit | 0.3 | 1.3 | 0.75 | 1.2 | | Minute/unit | 1 | 2.5 | 1.5 | 2 |   How many units of each product type should produce to maximize gross profit margin?  Objective: Maximize gross profit margin  Constraints:    Clips have highest marginal profit per unit of resource consumed.  Maximum possible production of clips  = 280,000 minutes ÷ minutes/unit  = 280,000 ÷ 2 = 140,000  Profit for maximum production of clips  = gross margin/unit \* max possible production  = $1.20 \* 140,000 = $168,000  **Outcomes:**  Unique optimal solution: there is exactly 1 solution that will result in maximum (or minimum) objective.  Alternative (multiple) optimal solution: objective is maximized (or minimized) by more than 1 combination of decision variables, all of which have same objective function value.  Unbounded solution: objective can be increased or decreased without bound (i.e., to infinity for maximization problem or negative infinity for minimization problem)  Infeasibility: no feasible solution exists  **Sensitivity analysis for Decision Variable:**  Sensitivity Analysis allows us to understand how optimal objective value and optimal decision variables are affected by changes in objective function coefficients, impact of forced changes in certain decision variables, or impact of changes in constraint resource limitations or requirements.  Sensitivity Analysis applies to changes in only 1 of model parameters at time; all others are assumed to remain at their original values  Reduced Cost: How much objective function coefficient needs to be reduced for nonnegative variable that is zero in optimal solution to become positive.  If variable is positive in optimal solution, its reduced cost is zero. If objective coefficient of any 1 variable that has positive value in current solution changes but stays within range specified by Allowable Increase and Allowable Decrease, optimal decision variables will stay same; however, objective function value will change.  **Sensitivity analysis for Constraints:**  **SHADOW PRICE** - how much objective function will change as right-hand side of constraint is increased by 1. Whenever constraint has positive slack, shadow price is zero.  When constraint involves limited resource, shadow price represents economic value of having additional unit of that resource. | | **Using sensitivity analysis:**  If change in objective function coefficient remains within Allowable Increase and Allowable Decrease ranges, then optimal values of decision variables will not change. However, you must recalculate value of objective function using new value of coefficient.  If change in objective function coefficient exceeds Allowable Increase or Allowable Decrease limits, then you must re-solve model to find new optimal values.  If change in right-hand side of constraint remains within Allowable Increase and Allowable Decrease ranges, then shadow price allows you to predict how objective function value will change 🡪 Multiply change in right-hand side (positive if increase, negative if decrease) by value of shadow price. However, you must re-solve model to find new values of decision variables.  If change in right-hand side of constraint exceeds Allowable Increase or Allowable Decrease limits, then you cannot predict how objective function value will change using shadow price 🡪 You must re-solve problem to find new solution.  **INTEGER OPTIMIZATION:**  **Solving models vs. General Integer Variable:**  Decision variables that we force to be integers are called general integer variables.  Algorithms for integer optimization models first solve LP relaxation (no integer restrictions imposed) and gradually enforce integer restrictions using systematic searches.  **Sensitivity analysis for Integer Optimization:**  Because integer models are discontinuous by their very nature, sensitivity information cannot be generated in same manner as for linear models  To investigate changes in model parameters, it is necessary to re-solve model.  **Example 1:** A company makes 110-inch wide rolls of thin sheet metal and slices them in smaller rolls of 12, 15, and 30 inches.  A cutting pattern is configuration of number of smaller rolls of each type that are cut from raw stock. Six different cutting patterns are used.   |  |  |  |  |  | | --- | --- | --- | --- | --- | | Size of End Item | | | | | | Pattern | 12in. | 15in. | 30in. | Scrap | | 1 | 0 | 7 | 0 | 5in. | | 2 | 0 | 1 | 3 | 5in. | | 3 | 1 | 0 | 3 | 8in. | | 4 | 9 | 0 | 0 | 2in. | | 5 | 2 | 1 | 2 | 11in. | | 6 | 7 | 1 | 0 | 11in. |   Demands for coming week are 500 12-inch rolls, 715 15-inch rolls, and 630 30- inch rolls.  Problem is to develop model that will determine how many 110-inch rolls to cut into each of six patterns to meet demand and minimize scrap.  Model development: Let be number of 110-inch rolls to cut using pattern. needs to be whole number (general integer variable) because each roll that is cut generated different number of end items. only constraints are end-item demand, non-negativity, and integer restriction    **Workforce scheduling model** is practical, yet highly complex, problem in many businesses i.e. food service, hospitals, and airlines.  Typically, huge number of possible schedules exist, and customer demand varies by day of week and time of day, further complicating problem of assigning workers to time slots.  **HOW TO FIND MODEL BEST FIT WITH DATA?**  **Method 1**. USE R2 CHANGE (FOR NESTED MODELS, LINEAR REGRESSION): Find most parsimonious model by using R2; Rationale - Parsimonious model is preferred if it fits data (at least) equal to more complex model; Two models are considered as “nested” if one is constrained version of other      🡪 (2) nested in (1) because they are same if b3=0  🡪 (2) is more “parsimonious” than (1); estimate less no. of coefficients (= parameters)  We prefer Model (2) over Model (1) if R2 change between Models (2) and (1) are not statistically significant (= simpler but equally well fit data) | HOW TO **COMPARE NESTED MODELS** IN R?  **1)** Fit each Model (1) and Model (2) 🡪 (model1) / (model2)  **2)** Use F-test to test R2 change is statistically different 🡪 (model 1, model 2)  **3)** If test cannot reject H0 (= No R2 difference) choose (2) ; otherwise stay with (1)  **Method 2**. USE INFORMATION CRITERION (IC; FOR NON-NESTED MODELS): *Find best-performing model by using information criterion; Rationale - Best-performing model is preferred, considering its complexity and fit to data; Commonly used information criterion measures for model selection*  - Akaike Information Criterion (AIC) or its adjusted version (AICc)  - Baysian Information Criterion (BIC) 🡪 AIC/BIC are transformed values of a function of residuals; smaller is better  How to **SELECT A MODEL** by using information criterion in R?  1. Fit candidate models -> e.g., lm(model1) / nls(model2) / …  2. For each model, calculate AIC or BIC -> AIC(model1) / AIC(model2) / AIC(…) (or use BIC) 3. Choose model having smallest value of AIC or BIC  IMPORTANT!!!  - ICs are statistical measures; assume one candidate model is (close to) “TRUE” model  - “True model” - a model that represents true, exact relationship between Y and X(s)  - In practice, you CANNOT check this assumption; usually ok for multiple regression/time series  - You need to choose model fit measures most suitable for your model(s) and data!!!  **INTERPRET REGRESSION ANALYSIS RESULT? (e.g., Y = b0 + b1\*X + e)**  (y) Intercept (also called constant) = Mean value of Y if value of all Xs = 0  Coefficient of X = impact of X on Y;  t-value = statistical test on each coefficient and constant (0 or not)  F-statistic = statistical test on all coefficients and constant (all 0 or not)  R2 - Measure of fit for linear regression model. proportion of variance in dependent variable explained by exploratory variable(s) = 1 – (Sum of squares of residuals;  SSR / Sum of squares total; SST = Pearson Correlation Coefficient R (Y, Predicted Y from model) 2 = “R” squared!  SSR = ∑(Y – predicted Y)2 = Unexplained variance of Y  SST = ∑(Y – mean Y) 2 = Total variance of Y Adjusted R2 (for linear regression): R2 tend to increase as number of predictors in model increases. Adjusted R2 calculates “accurate R2” by penalizing R2 with number of predictors and sample size = 1 – [(1 - R2) \* (n – 1) / (n – p – 1)];  n = sample size, p = no. of predictors  Pseudo R2 (for nonlinear regression; also can be “adjusted”). Model fit in context of nonlinear models is usually defined in two ways  These are mathematically different from R2 in linear models; therefore “Pseudo” R2  1) Degrees of improvement from intercept-only model - McFadden’s Pseudo R2  2) Use same idea of linear R2; variance of Y explained by X - Efron’s Pseudo R2 - Efron’s Pseudo R2 works ok for simple regression where linearity assumption is not severely violated  **Measures of Location** (Mean, Median, Mode)  **Measures of Dispersion** (Range, Variance, Standard deviation, Chebyshev’s Theorem, Coefficient of Variation)  **Measures of Shape** (Skewness, Kurtosis)  **Measures of Association** (Covariance and Correlation) | | | |  |  | | --- | --- | | 🟑Which of the following are the common types of decision that can be enhanced by using analytics?  ⎯Pricing: setting prices for customer, industrial goods, government contracts, …  ⎯Customer segmentation: identifying and targeting the customers groups in retail insurance and credit card and industries.  ⎯merchandising: determine brands to buy, qualities and allocations  ⎯location: finding best locations for bank branches and ATMs  🟑Complete linkage clustering of two groups compute the distance between pair of objects  🟑Predictive analysis:  ⎯Seeks to predict the future by examining the historical data, detecting patterns or relationships in these data and then extrapolating the relationships forward in time  ⎯Predictive analysis can predict the risks and find relationships in data not readily apparent with traditional analysis  ⎯Using advantaged techniques, predictive analysis can help detect hidden pattern in large quantities of data to segment and group data into coherent sets to predict behaviour and detect trends. | 🟑Descriptive analysis:  ⎯characterises, consolidate and classify data to convert it into useful information for the purposes of understanding and analysing business performance.  🟑Overfitting can be prevented by adding more data  🟑The weights of determining the discriminant functions are computed by maximizing the variance between groups relative to the variance within groups.  🟑A study to predict the number of home runs scored in a softball league with 32 teams of 9 players, based on the different material used to make the bat ( alloy, composite, aluminium, hybrid) and the player’s experience playing in the league. 🟑What is the appropriate number of independent variables for the regression model?  materials + year experience variables.  🟑Residuals help detect outliers that bias the regressions analysis.  🟑 Explain the difference as the level of confidence decreases from 95% to 90%.  When the level of confidence decreases from 95% to 90%, the range of CI increase 🡪 the rejection area decreases. | | | | | | |
| **No trend** | | *Simple moving average or simple exponential smoothing* | | *Holt-winter no-trend smoothing model or multiple regression* | | | | |
| **Trend** | | *Double exponential smoothing* | | *Holt-winter addictive or Holt-winter multiplicative model.* | | | | |
|  | | **Population of size N** | | | | **Sample of n observations** | | |
| **Mean** | |  | | | |  | | |
| **Variance** | |  | | | |  | | |
| **Co-var.** | |  | | | |  | | |
| **Co-**  **relation** | |  | | | |  | | |
| **EXPECTED** confidence is number of transactions that include consequent divided by total number of transactions.  **LIFT** is ratio of confidence to expected confidence.  Higher lift ratio, stronger association rule; value greater than 1.0 is usually good minimum. | | | **Errors metrics and Forecast Accuracy:** | | | | | |
| Mean absolute deviation: focus on mean value of errors | | | | Mean square error/deviation: focus on variance of errors | |
| For all metrics, smaller values 🡪 better data | | | | | |
| Root mean square error focus on standard deviation of errors | | | | Mean absolute percentage error: cannot be used if a time series contains 0 (division by 0) | |
| Example: supermarket database has 100,000 point-of-sale transactions; 2000 include both and B items; 5000 include C; and 800 include A, B, and C  Association rule: “If and B are purchased, then C is also purchased.”  Support = 800/100,000 = 0.008  Confidence = 800/2000 = 0.40  Expected confidence = 5000/100000 = 0.05  Lift = 0.40/0.05 = 8  lift ratio indicates how much more likely we are to encounter event and B are purchased, as compared to entire population of transactions.  **Cause and Effect modelling:**  Correlation analysis can help us develop cause-and effect models that relate lagging and leading measures.  Lagging measures tell us what often external business results such has happened and are as profit, market share, or customer satisfaction.  Leading measures predict what will happen and are usually internal metrics i.e. employee satisfaction, productivity, and turnover.  **MONTE CARLO:**  **Monte Carlo simulation:** isprocess of generating random values for uncertain inputs in model, computing output variables of interest, and repeating this process for many trials to understand dist. of output.  **Perform following steps:**  1. Develop visual model  2. Determine prob. dist. that describes uncertain inputs in model  3. Identify output variables you wish to predict  4. Set number of trials or repetitions for simulations  5. Run simulation  6. Interpret results  **Market basket analysis**, for example, is used determine groups of items consumers tend to purchase together.  Association rules provide information in form of if then (antecedent consequent) statements**.** In other situations, historical data are not available, and we can draw upon properties of common prob. dist. to help choose representative dist. that has shape that would most reasonably represent analyst’s understanding about uncertain variable.  Uniform or triangular dist. are often used in absence of data.  **Sampling methods:**  Monte Carlo sampling selects random variates independently over entire range of possible values of distribution. Monte Carlo sampling is more representative of reality and should be used if you are interested in evaluating model performance under various what-if scenarios.  **Confidence interval for Mean:**  Each time you run simulation, you will obtain slightly different results.  Confidence interval:  Because Monte Carlo simulation will generally have very large number of trials, we may use standard normal z value instead of t-dist. in confidence interval formula. | | | | | **Flaws of averages:** evaluation of model output using average value of input is not necessarily equal to average value of outputs when evaluated with each of input values.  In newsvendor example, quantity sold is limited to smaller of demand and purchase quantity, so even when demand exceeds purchase quantity, profit is limited.  Using average values in models can conceal risk.  **Monte Carlo using simulation using Fitted Distribution:**  Sampling from empirical data has some drawbacks.  Empirical data may not adequately represent true underlying population because of sampling error.  Using empirical dist. precludes sampling values outside range of actual data.  **Steps for “Fitting” theoretical dist. and computing goodness of fit:**  Choose suitable theoretical model:  For instance, normal or power law model. This task is informal; descriptive statistics like histogram and skewness indicator of observed data can be valuable hints;  Estimate model parameters:  Each theoretical model has parameters, for instance, mean and standard deviation for normal model. This task consists of estimating most likely model parameters for empirical dataset;  Determine significance level:  This tricky step establishes how good observed data match theoretical model with estimated parameters. If computed significance level is beyond pre-defined threshold, goodness-of-fit hypothesis is accepted, otherwise it is rejected  **Estimate model parameters:**  maximum likelihood estimation method (MLE) is most popular method to estimate dist. parameters from empirical sample.  It finds model parameters that maximize likelihood of observed data with respect to theoretical model.  **Determine significance level:**  Fit normal distribution, use Shapiro-Wilk test:  If p-value is lower than threshold (usually fixed to 0.05) then normality hypothesis is rejected.  Fit arbitrary distribution, use Kolmogorov-Smirnov test:  If p-value is lower than given threshold, goodness-of-fit hypothesis is rejected.  **Cash-budget model** is process of projecting and summarizing company’s cash inflows and outflows expected during planning horizon.  Most cash budgets are based on sales forecasts. Because of inherent uncertainty in sales forecasts, Monte Carlo simulation is appropriate tool for modelling cash budgets. | | | |
| **Nguyen Ngoc Linh Chi - Machine Learning Cheat-sheet**  **PERMUTATION:** A permutation of set of objects is ordering of objects in row.  of set of elements are:  REMARK:  **COMBINATION:** of set of n elements  **BINOMIAL COEFFICIENTS:** For any , we have:  **NUMBER OF ELEMENTS IN POWER SET:** , if set S has n elements, total # subsets of S is .  REMARK:  **NUMBER OF INTEGER SOLUTIONS: #** non-negative integer solutions of equation OR # r-combinations with repetition allowed that can be selected from a set of n objects is  **ARRANGING IN A CIRCLE:**  For n distinct objects arranged in a circle, there are  **CONDITIONAL PROBABILITY** of B given that A is  **GENERAL MULTIPLICATION RULE:**    **INVERSE PROBABILITY:**  **INDEPENDENT vs. MUTUALLY EXCLUSIVE:**  Two events A & B being *independent* and *mutually exclusive* are *NOT same thing*.      If A & B are *mutually exclusive & non-trivial (positive probability)* then A & B *cannot*be *independent*.  **INDEPENDENCE VS. MUTUALLY EXCLUSIVE**  Two events A & B being *independent* and *mutually exclusive* are *NOT same thing*.      If A & B are *mutually exclusive & non-trivial (positive probability)* then A & B *cannot*be *independent*.  **PAIRWISE INDEPENDENT EVENTS:**  A set of events are said to be pairwise independent  **MUTUALLY INDEPENDENT EVENTS:**  A set of events are said to be *mutually independent/ independent*  are mutually independent  different cases.  Mutually independence pair-wise independence  Pair-wise independence mutually independence  **PARTITION:** If are *mutually**exclusive* and *exhaustive*  a *partition* of S.  **RULE OF TOTAL PROB.:** If is *partition*  **BAYES’S THEOREM:** Let be partition of S.  **CHEBYSHEV’S INEQUALITY:** Don’t know how X behave  **RANDOM VECTORS & RANGE SPACE**  Let be an experiment and a sample space. a two-dimensional random vector. range space is  **INDEPENDENT RANDOM VARIABLE:**  X and Y are independent  X and Y are independent  are independent | | **COVARIANCE:**  *;*  **independent**  **CORRELATION COEFFICIENTS:**  If X and Y are **independent**, then .  On other hand, does not imply independence.  **STANDARD ERROR:**  **STANDARD NORMAL**: X is called as standard normalrandom variable when ; and Z, then      )  **:**  quantile of random variable X is :  **POISSON BINOMIAL:** Let . ; remains a constant as .  The approximation is good when and  OR and . If is close to 1, we can still use Poisson distribution to approximate binomial probabilities.  **NORMAL BINOMIAL**  Use when:  When is small and is not extremely close to 0 or 1, approximation is fairly good.  Use normal approximation only if  **Continuity correction:** Suppose X is a binomial random variable mean , variance .  **UNBIASED ESTIMATOR:** Let be estimator of (random var. based on sample). If , is unbiased estimator of  is an unbiased estimator of  An unbiased estimator of is  **SAMPLING DISTRIBUTION RELATED TO SAMPLE MEAN:**  Infinite population or from a finite population with replacement having mean and variance , sample distribution of sample mean has mean and variance is:   |  |  | | --- | --- | |  |  |   **LAW OF LARGE NUMBER LLN:** Let be a random sample of size with mean and variance . Then,  **CENTRAL LIMIT THEOREM:** Let be a random sample of size with mean and variance .  ***Normal distribution provides an excellent approximation to sampling distribution of mean if n ≥* 30*.*** If are (approximately) , then is (approximately) regardless of sample size .  **GAMMA FUNCTION:** ( is a complex number with positive real part). Gamma function is defined by    For integer values of | | | | | | | **DISTRIBUTION:** the chi-square or distribution with n degree of freedom is the distribution of a sum of the square of independent standard random variables .  Let , then  For large n, approximately.  If are independent chi-square random variables with degree of freedom has distribution with degrees of freedom.  **– TABLE:** table contains values of for various n:  **SAMPLING DISTRIBUTION RELATE TO SAMPLE VARIANCE:**  Let S2 be sample variance of a random sample of size n taken from a normal population with and  is degrees of freedom.  **STUDENT DISTRIBUTION:**    .  The t-table *shows*  In table degree of freedom  If random sample was selected from a normal population then  are independent, so are  **FISHERS F-DISTRIBUTION:** *(ratio between two estimate of var.)*  Random **samples** of size are selected from 2 **normal** **population** with variances  Table F-distribution gives value of such that  **CONFIDENCE INTERVAL:**  The interval computed is called confidence interval for . fraction is called confidence coefficient or degree of confidence  **CI for with KNOWN :**  is number with an upper-tail probability of for standard normal distribution Z.  **SAMPLE SIZE FOR ESTIMATING :** For margin of error ,  **CONFIDENCE INTERVALS and SAMPLE SIZE**  Determine appropriate sample size needed to estimate population parameter within specified level of precision .    *Suppose that we wish to determine number of voters to poll to ensure sampling error of at most ± 2%. With no information, use = 0.5 (proportion who poll):*  Use sample proportion from preliminary sample as estimate of or set = 0.5 for conservative estimate to guarantee required precision  **CONFIDENCE INTERVAL FOR PROPORTION**  (sample proportion), *where is number in sample having desired characteristic and sample size.* | | |  |  |  | | --- | --- | --- | | ***CI of ; KNOWN with***  ***NORMAL population or*** | |  | | ***CI of ; UNKNOWN***  ***NORMAL population &*** | |  | | ***CI of ; UNKNOWN with***  ***NORMAL population or*** | |  | | ***CI of ; KNOWN***  ***NORMAL population or*** | |  | | ***CI of ; UNKNOWN***  ***with*** | |  | | ***CI of ; UNKNOWN ; NORMAL population and :***Define  *is pooled sample variance, an estimator for*  *confidence interval for is* | | | | ***CI of ; KNOWN***  ***with*** |  | | | ***CI of with NORMAL population,*** |  | | | ***CI of with*** |  | | | | ***CI of ; KNOWN with***  ***NORMAL population*** |  | | | | ***CI of ; UNKNOWN with NORMAL population*** |  | | | | ***CI of ; UNKNOWN and : NORMAL population*** | | |   **HOW TO DO A HYPOTHESIS TEST**  Step 1: Null and Alternative hypothesis (what you want to test)  The equal part is always in null hypothesis  Step 2: Determine level of significance () and power ().   |  |  |  | | --- | --- | --- | |  |  |  | | Reject | Type I error | Correct decision  = | | Not reject | Correct decision | Type II error |   Step 3: Identify test statistic, distribution and rejection criteria.    Step 4: Compute test statistic value based on your data.  Step 5: Conclusion.  **VALUE:** If , do not reject , else reject  **PAIRWISE ASSOCIATION TEST:** On the left is used non-parametric test of pairwise correlation (robust to non-normal population)    **Reference for statistical analysis using SAS, Stata, SPSS, R**  [https://stats.idre.ucla.edu/other/mult-pkg/whatstat/#](https://stats.idre.ucla.edu/other/mult-pkg/whatstat/)  **HYPOTHESIS *on* +KNOWN :NORMAL population or :**  To test:  When is true, we have test statistic:  **HYPOTHESIS on + UNKNOWN NORMAL population**  To test:  When is true, we have test statistic: | | |  | **DISCRETE RANDOM VARIABLE** | | | | **CONTINUOUS RANDOM VARIABLE** | |
| ***Probability***  ***Mass/***  ***Density***  ***Function*** |  | | | |  | |
| ***Cumulative Distribution Function*** | If only *possible values are integers* and if *a and b are integers*, | | | |  | |
| ***Mean/***  ***Expectation/***  ***Expected values*** |  | | | |  | |
| *a; b constant, .* | | | |  | |
| ***Expectation/***  ***Mean***  ***of Function*** |  | | | |  | |
| is called moment of X | | | | | |
| ***Variance*** |  | | | |  | |
| ; *;* | | | |  | |
| ***Joint Prob. Mass/***  ***Density Function*** |  | | | |  | |
| ***Marginal Distribution*** |  | | | |  | |
| ***Conditional Probability***  ***Mass***  ***Function*** |  | | | |  | |
| The conditional probability mass/density function of X:  X and Y **independent** for some x and y | | | |  | |
| **TWO-SIDED TEST ~ CONFIDENCE INTERVAL:**  confidence interval contains , is not located within rejection region will NOT be rejected.  **HYPOTHESIS on with KNOWN NORMAL population or :** To test:  When is true, we have test statistic:  **HYPOTHESIS TEST ON :** To test:  We can use test statistic   |  |  | | --- | --- | |  | Rejection region | |  |  | |  |  | |  | ; |   **HYPOTHESIS on with UNKNOWN**    To test:  When is true, we have test statistic:  **HYPOTHESIS on with UNKNOWN NORMAL population;**  To test:  When is true, we have test statistic:  **HYPOTHESIS TEST ON PAIR SAMPLES:**  For paired sample, define:  To test:  When is true, we have test statistic: | | |  | **PROBABILITY MASS/DENSITY FUNCTION** | | **MEAN; VARIANCE** |
| ***Discrete***  ***uniform***  ***distribution*** |  | |  |
| ***Continuous***  ***uniform***  ***distribution*** |  | |  |
| ***Bernoulli***  ***trials*** | *Experiment with 2 outcomes (“success”; “failure”)* | |  |
| ***Binomial***  ***distribution*** | and | |  |
| ***Negative***  ***binomial***  ***distribution*** | *#trials before obtain k successes;* | |  |
| ***Geometric***  ***distribution*** | Memoryless property of Geometric: | | #*required trials until first success is achieved* |
| #*failures until first success is achieved* | |  |
| ***Poisson***  ***random***  ***variable*** | *# success in fixed interval/period/region* | |  |
| ***Exponential***  ***distribution*** | Memoryless property of Exponential distribution: | |  |
| ***Normal***  ***distribution*** |  | |  |
| ***distribution*** | *degree of freedom* *gamma func.*; | |  |
| ***Student’s***  ***distribution*** | *Z* *~* and U~ | |  |
| ***The F***  ***distribution*** |  | | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **HYPOTHESIS TEST ON :** To test:  We can use test statistic   |  |  | | --- | --- | |  | Rejection region | |  |  | |  |  | |  | or |   **SAMPLE TEST ON PROPORTION:**  **Calculator with Normal distribution:**      **REJECTION REGION and P-VALUE for**  **NORMAL distribution**:   |  |  |  | | --- | --- | --- | |  | *Rejection region* | *p-value* | |  |  |  | |  |  |  | |  | or |  |   **distribution:**   |  |  |  | | --- | --- | --- | |  | *Rejection region* | *p-value* | |  |  |  | |  |  |  | |  | or |  |   **PEARSON CORRELATION TEST:**  *Test association between 2 quantitative variables:*  The test calculates the Pearson correlation coefficient and the p-value for testing non-correlation  Let x and y be two quantitative variables, where n samples are observed. *The linear regression coefficient is*  Under , the test statistic follows the Student distribution with degree of freedom  **NON - parametric test of PAIRWISE association:** When to use it? Observe the data distribution: presence of outliers; the distribution of the residuals is not Gaussian.  **Spearman rank-order correlation (quantitative ~ quantitative):** measure of monotonicity of the relationship between two datasets  Like other correlation coefficients, this one varies between -1 and +1 with 0 implying no correlation.  Correlations of -1 or +1 imply an exact monotonic relationship.  Positive correlations imply that as  Negative correlations imply that as  **Wilcoxon signed-rank test (quantitative ~ )**  The Wilcoxon signed-rank test is a non-parametric statistical hypothesis test used when comparing two related samples, matched samples, or repeated measurements on a single sample to assess whether their population mean ranks differ (i.e. it is a paired difference test). It is equivalent to one-sample test of the difference of paired samples.  It can be used as an alternative to the paired Student’s t-test, t-test for matched pairs, or the t-test for dependent samples when the population cannot be assumed to be normally distributed.  It has lower sensitivity compared to t-test. May be problematic to use when the sample size is small  Null hypothesis difference between the pairs follows a symmetric distribution around zero.  **Mann- Whitney U test (quantitative ~ categorial 2 level):** also called Mann-Whitney-Wilcoxon/Wilcoxon rank-sum test/Wilcoxon-Mann-Whitney test is a nonparametric test of the null hypothesis that two samples come from the same population against an alternative hypothesis, especially that a particular population tends to have larger values than the other.  It can be applied on unknown distributions contrary to e.g. a t-test has to be applied only on normal distributions.  **Linear model:**  Image result for linear model  Given n random samples with |  |  |  |  |  |  |