**Statistics Test Summary Table**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Test** | **Independent Variable Type** | **Dependent Variable Type** | **Purpose & Assumptions** | **Test Statistic Equation** | **DOF** | **SAS Code** |
| **One-sample**  **t test** | Continuous | Continuous | **Purpose:** compares sample mean to population mean with unknown sigma  **Assumptions:**   * Population is normal or M&M guidelines are met * Unknown population variance (sigma) | https://lh6.googleusercontent.com/iwraU4t7O3-w9ii5hZD8tRXllD-oAEVz9oHZJNcgqB7zAuxBWyXStCk0bxtBLYmV0zFv5pE3_hntySQG94l38p201irq-EDwXE8RvLDZTizdiNlsBWWuFb4lQhszWKBotMxpfnds | n-1 | https://lh5.googleusercontent.com/OshVlZ1xHYzTagW8yRrKPM5nZggoRjy2mdVERWb2xxW_LH-RNmG8vimjUHqYTtXEC0sGEK-H8cmRGbsQQq7lKadT7w0-s4JMhfDYzaV8EmK3jJcqfmFzFwiDRBlrIfR32F5yVdWl |
| **Paired t-test** | Continuous | Continuous | **Purpose:** Compares two population means with two samples in which observations in one sample can be paired with observations in the other sample; Use a one-sample t-test to test whether the true average difference between paired observations is 0; -Reduces confounding with a more precise estimate than a 2 sample t-test  **Assumptions:**   * Normal distribution or M&M guidelines are met | https://lh4.googleusercontent.com/773qsIRQW0LThM4YRMWAFu8gu7TYlcxsmM1yPJrmxdRV0U7abDIDjkpRNlSuvilWVrsm8hLazHF8R4zn8yszTajvCEfq2r9PgMeweFcjLUj0vLR9U2U8GbMV3x--xqQdITQXP5Ga | n-1 | https://lh6.googleusercontent.com/5JIuDoqoRxKOsEzyG_A1stqntkdoVEiJhN3V0VH5mBHBh1YZL225qsgTOAcn-CKA0-DpW9RodVhygHDNKiIE3aHPIOVB_CAhi7bn8a0xrcOZ0j6PNwYvhiYOCUkFdjB-LKrmQLMm  proc ttest data = "data";  paired write\*read;  run; |
| **One sample Z-test** | Continuous | Continuous | **Purpose:** compares sample mean to population mean where sigma is known  **Assumptions:**   * Assumptions: population is normal or M&M guidelines are met * Known population variance (sigma), which is highly unlikely | One Sample Z Test | n-1 | (unlikely to have scenario where σ is known but μ, the population mean, is not known) |
| **Wilcoxon Signed rank test** | Comparing population means | N/A | **Purpose:** compares two population medians  **Assumptions:**   * Non-parametric alternative to one sample t-test comparing medians of populations * Used if population distribution is non-normal or unknown, if population variances are unknown and M&M’s guidelines are not met * Population is symmetric (not necessarily normal); if the population is not symmetric the test can still be used but may be less powerful |  | n-1 | https://lh4.googleusercontent.com/3Wtuvc6FrBvzmgg9cFZc2XyVt_n0KZT84c6HA3RCxnWGiSAAgTclZ8MGLhvQMYCgOgGj5WWs72wERm_oBzvHci2-WTNYYMcNrPjo-FC952wo8HlOs93TCKvzAzL1vetQBEwowfbX  proc univariate data = hsb2a;  var diff;  run; |
| **Pooled two sample t-test** | Categorical | Continuous | **Purpose:** Comparing The Means of Two Normal Populations When The Population Variances Are Assumed Equal  **Assumptions:**   * Only one tailed tests can be performed * Population variances are assumed to be equal * Normal distribution or M&M: total sample size meets this if non-normal | https://lh6.googleusercontent.com/P_NuIovqVo92QIHFGpvMsYuorBbsY8-GAlPhxhimpU1o97hkbAV4lsBtEq1DyKmsMhy7_pAUPNTFJtagw80HVSgxLS4rEK1lNFmAu5YmMCxDdPb3eFPqdZsQKAHsyNBPDhvEw_Tl  https://lh6.googleusercontent.com/Z8g0mDt0ZDJ9xKug61T8rk7IwGEqkCjP4645ZRyGZv-YaYistwIHuq2fbUyLz6IghkvRK3VH6Ry5pgikcBKxBsMLu63ZnyFRuVVPNowHfBi6bA7C2oFgVVhm423yAH_Au94BZ8pk  CI:  https://lh6.googleusercontent.com/t07kBdOoe0sgtp_2hzjnq0LUXtK_iuDTr2j21RHANv1MgLgE2rHdkUEk_iX9oAiWQR5jIac16wleSwovUP2fLhg1s-GiOuNz57oCW329yMI7g7iRMBWpMFILQPFE4L13qnZ4PJGM | https://lh3.googleusercontent.com/9RGyXeYJ3wKOWpxqhiUE0B7g5B5tCiZl73PMLO148q7MGwBhlT96SLwSvleQ2fYeijWCbNd_sa0xGHUW32eLB3scE8lJ2254u-qIirY1VWOvqGl39lT-60C0mAArwYbHRUaIW48L | proc ttest data=read sides=2 alpha=0.05 h0=0;  title "Two sample t-test example";  class method;  var grade;  run; |
| **Unpooled two-sample t-test** | Categorical | Continuous | **Purpose:** Comparing The Means of Two Normal Populations When The Population Variances Unequal  **Assumptions:**   * Population variances are unequal * Large enough sample size (total sample size meets M&M’s guidelines) or normally distributed data * Between the pooled and unpooled if population variances are unknown → use the largest p-value or the most conservative estimate | https://lh4.googleusercontent.com/W4jwBd1d8uOBpNnNpw6HMvGXAr1K4-FGbooDS36Uu6SeaT8nUQoljG5QbktRVcW0wf5vRoaC3fXXjiak5pQ4iII6NFiVpuKp6U48tdMciZ8U5Jt0ahQbBQLr2P6JjlK296Z6eGmB  CI:  https://lh6.googleusercontent.com/LitKPFp4p37MTBQf42VozIpDMUpXaQ7B1olGANH4NiqGOBdzC7SfFTSLVngKFAgEYFsDzyUod5RvUvh15UUUcrLPaiBBQuleOfbLEXFr9_CwLPFEmKjNC_jGmUk6LW65qE2X5v5B | Satterthwaite’s Formula  df = min( n1-1, n2-1) | proc ttest data=read sides=2 alpha=0.05 h0=0;  title "Two sample t-test example";  class method;  var grade;  run; |
| **Wilcoxon Rank sum test (Mann-Whitney U)** | Categorical | Continuous | **Purpose:** Comparing two independent continuous samples when assumptions of normality / sample size are not met for the independent samples t-test.  **Assumptions:**   * Use if population is not normal or sample size is small enough and cannot assume normality via M&M guidelines | http://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_nonparametric/ada-reference.gifhttp://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_nonparametric/lessonimages/equation_image37.gif  http://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_nonparametric/ada-reference.gifhttp://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_nonparametric/lessonimages/equation_image38.gif  where R1 = sum of the ranks for group 1 and R2 = sum of the ranks for group 2 | NA | Proc npar1way data = "data" wilcoxon;    class female;    var write;  Run; |
| **Chi-square test of homogeneity** | Categorical | Categorical | **Purpose:** compare two independent random samples proportions of categorical variables  **Assumptions:**   * At least 80% of all cell values must be >=5 |  | (#Rows-1)\*(#columns-1) | proc freq data = "c:/mydata/hsb2";    tables schtyp\*female / chisq;  run; |
| **Chi-square test of independence** | Categorical | Categorical | **Purpose:** the null hypothesis is that the row variable and column variable are independent, the alternative is that these variables are associated (i.e. not independent). Only one sample is collected in this test; once this sample is drawn from the population, each object/person is classified according to the two variables under study.  **Assumptions:**   * One sample * R&C table: one sample, multiple categorical variables * 80% values >/= 5 * Fishers Exact Test: 2 x 2 contingency table with counts < 5 |  | (R-1)(C-1) | proc freq data = "c:/mydata/hsb2";    tables schtyp\*female / chisq;  run; |
| **Simple Linear Regression** | Continuous or Categorical | Continuous | **Purpose:** for estimating the best-fitting line between 1 independent variable (continuous or independent) and 1 continuous dependent variable; perform hypothesis tests to determine whether or not the relationship is statistically significant; make predictions about the average value of Y for values of X of interest  **Assumptions:**  -Both independent and dependent variables are normally distributed  -ρyx is significant | True model: Y =β0 + β1X+ E  -β0  = true (population) y-intercept  -β1 = true (population) slope  -Every 1 increase in X, Y increases by β1  -E = random error term (assumed to be equal to 0 on average)  Estimate of model:  Macintosh HD:Users:emilywang:Desktop:Screen Shot 2019-12-02 at 9.22.28 PM.pngTest of significance for true slope is H0: β1 =0 vs. HA: β1≠ 0  Macintosh HD:Users:emilywang:Desktop:Screen Shot 2019-12-02 at 9.22.37 PM.png | n-2 | PROC REG CORR DATA = H.cake;  model yum = flavor;  RUN; |
| **Multiple Linear Regression** | Continuous or Categorical | Continuous | **Purpose**: a single regression model that estimates the relationship between all relevant predictors and an outcome with greater precision than separate simple linear regressions  **Assumptions:**  -Linearity: the mean value of Y is a linear function of X1, X2, … Xk  -Independence: the sample must be a random sample with Y values being independent of each other  -Homoscedasticity: the variance of Y does not depend on X1, X2, … Xk  -Normality: Y is a normal random variable | Multiple Linear Regression Model relates the dependent variable Y to the set of independent variables X1, X2, … Xk:  Y = β0 + β1X+ β2X2 + … + βkXk + E  Y = dependent variable  Xj = independent variable j (j = 1, 2, … , k)  β0 = true y-intercept of straight line  β1 = true slope associated with Xi adjusted for all other independent variables  E = random error term, assumed ~N(0, σ2)  Macintosh HD:Users:emilywang:Desktop:Screen Shot 2019-12-03 at 12.10.56 AM.png-Each slope estimate conveys how much an outcome is expected to change for a one unit increase in the predictor after adjusting for the effects of all other predictors  Overall F-test for Multiple Regression  H0: β1 = β2 = … = βk = 0 (no linear association) vs. HA: At least one β≠ 0 (linear association)  Test statistic: F = MSR/MSE  Partial T-tests in Multiple Regression  H0: βi = 0 (all other βj ≠ 0 or controlling or adjusting for effects of the other independent variables) vs. HA: βi ≠ 0 (all other βj ≠ 0) | F-statistic: k, n-k-1  Partial T-test: n-k-1 | PROC REG DATA = H.cake;  model yum = flavor sugar fat frosting;  RUN; |
| **One-way fixed effects ANOVA** | Categorical dummy variables with 2 or more categories | Continuous | **Purpose:** to compare the means of two populations by levels of the independent variable  One-way = only one overall predictor (i.e. independent variable) is being investigated; predictor is also called a factor with groups that are being compared representing different levels of the factor (a.k.a. factor levels)  Fixed effects = factor levels were specifically chosen for the study as opposed to being randomly chosen from all possible groups  **Assumptions:**  -Dependent variable is normally distributed | Overall F-test for ANOVA  H0: population means are equal vs. HA: population means are not all equal  Multiple Comparisons in One-Way ANOVA  -In order to identify and describe how the means are different, the most common approach is to perform all possible pairwise (i.e. two-sample) tests comparing the means (i.e. perform two-tailed tests for each of the following H0: μi-μj = 0 for all i and j, where i=1, … k, j = 1, … k, i≠j and k=number of populations  -Instead of two ordinary two-sample t-tests, can use Tukey’s method and Bonferroni’s Method | k-1, n-k | PROC GLM DATA = H.cake;  CLASS flavor;  MODEL yum = flavor;  MEANS flavor;  RUN; |