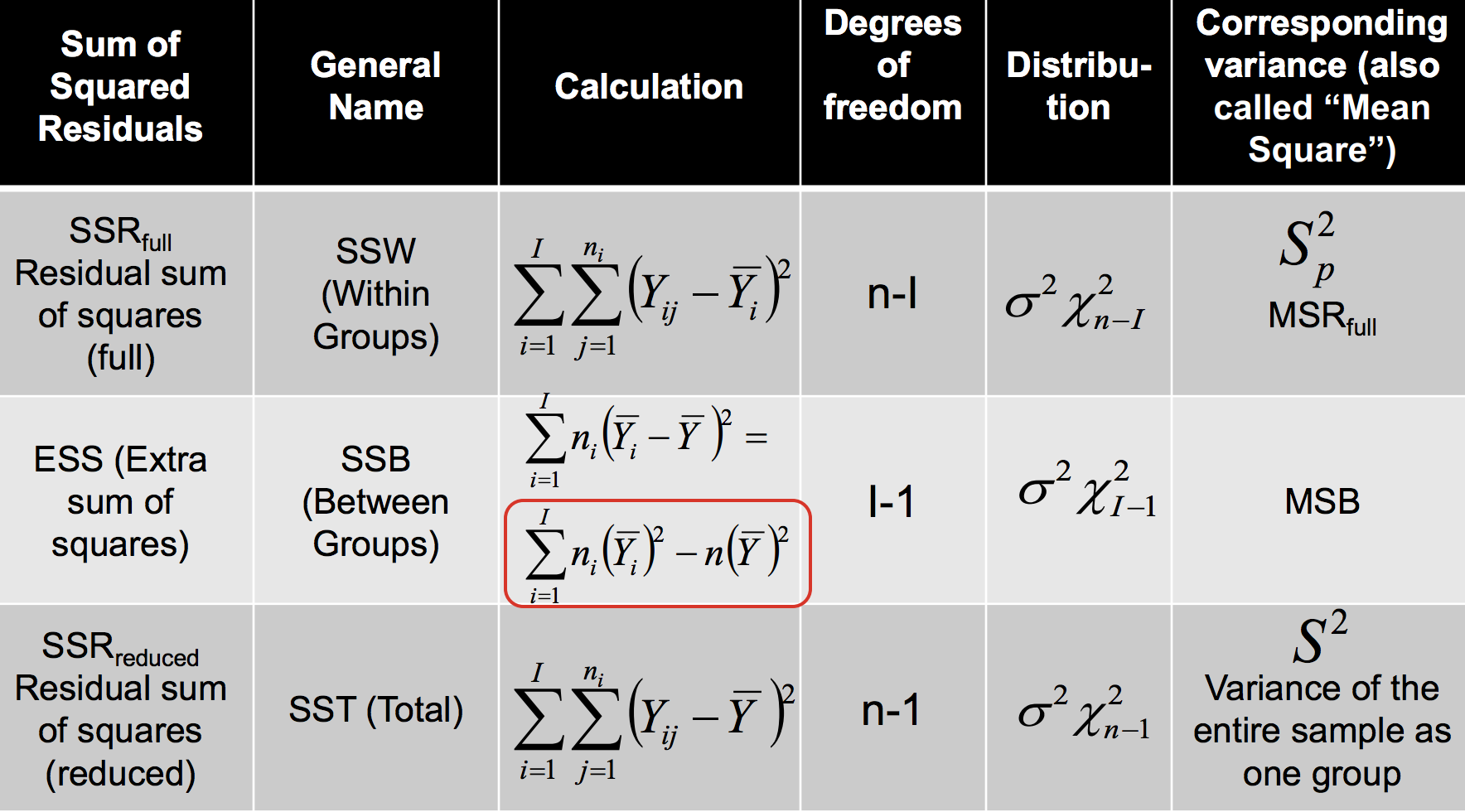
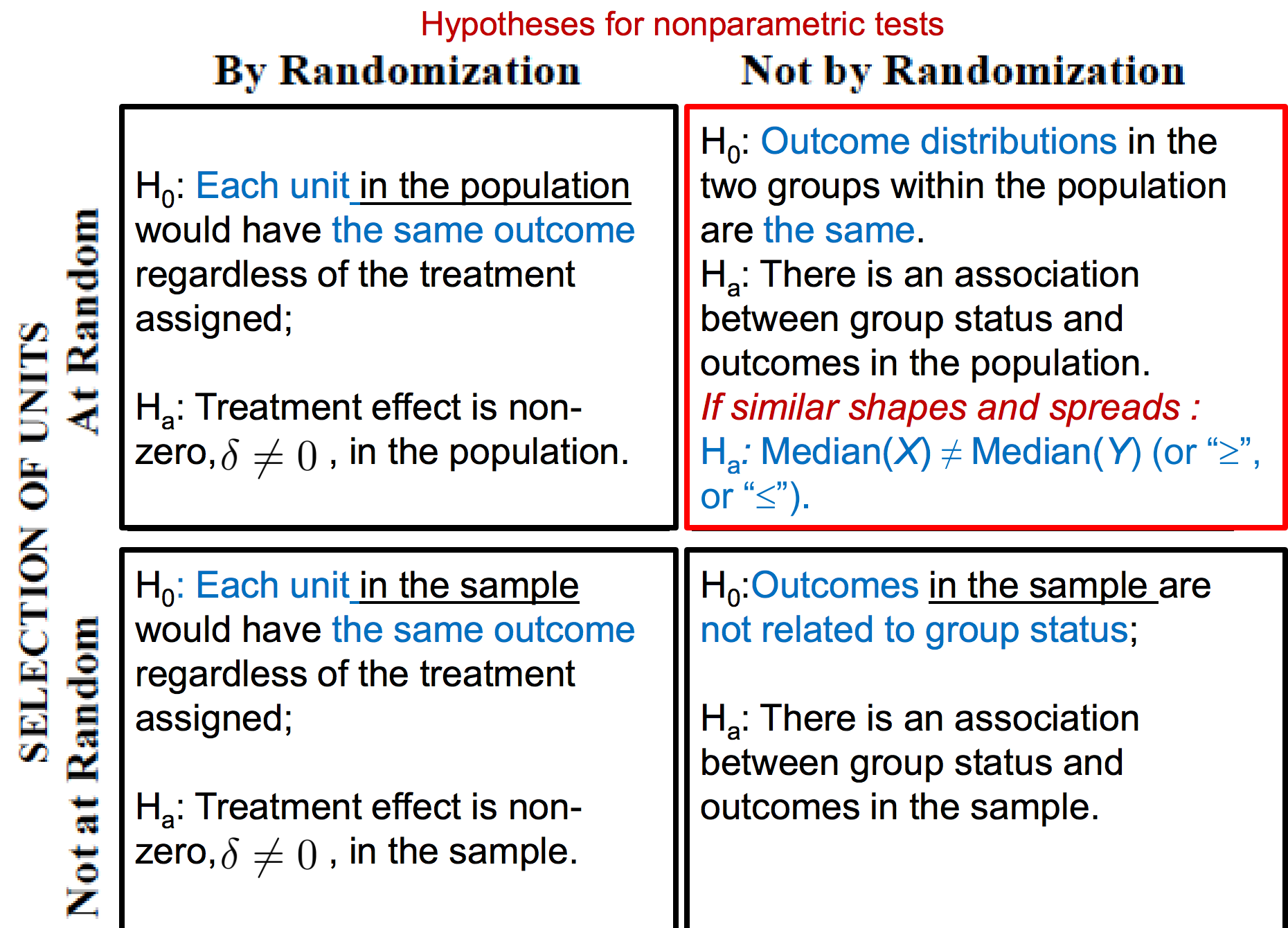
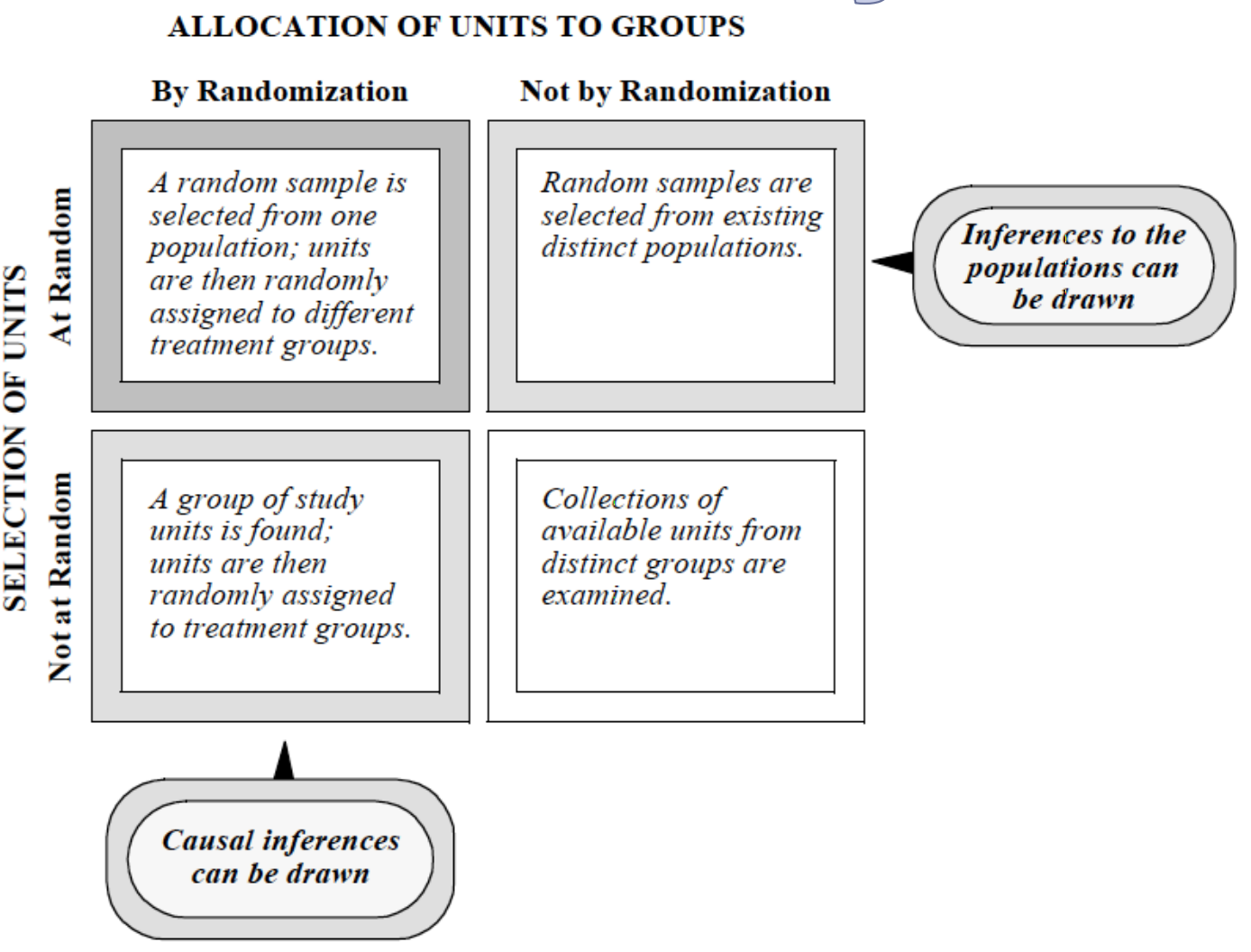
|  |  |
| --- | --- |
| **Study (experimental) unit (subject)** - one member of a set of entities being studied.  **Parameter (also, estimand)** -proportion of childless households in the population.  **Estimate** - proportion of childless households in the sample.  **Parameter** is a population characteristic.  **Estimand** is a parameter that is being estimated (**μ**).  **Statistic** is a function of the data (therefore, it is random).  **Statistic:** A function of the data, **μ** **μ** (y ). Test Statistic: Statistic used to weigh evidence  supporting and contradicting the null hypothesis.  **Reference Distribution**: Probability distribution of the test statistic, assuming that the null hypothesis istrue, f(**μ** (Y)|H0).  **Estimator** is a statistic used as a guess for the value of the estimand ( ).  **Estimate** is a quantity, which is a particular realization of the estimator (4/12 = 0.33).  **Target Population:** A collection of units a researcher is interested in; a group about which the researcher wishes to draw conclusions.  **Sampling frame:** Collection of units that are potential members of the sample.  **Overcoverage:** selecting units that are outside your sampling frame.  **Undercoverage**: excluding units that are within your sampling frame.  **Sample:** A [randomly selected] subset of a sampling frame  **Simple Random Sampling (SRS)** – every subset of n units has equal chance to be selected  **Stratified Sampling** – split the population into homogeneous subpopulations and use SRS (or another method) within a sampling frame of each subpopulation.  **Systematic Random Sampling** - select every k’th unit from the ordered sampling frame, starting randomly from of the first k positions.  **Variable probability sampling** – allow units to have nonequal probabilities of being sampled.  **p-value** - probability that the test statistic would be at least as extreme as observed, under the null hypothesis. (or “strength of evidence against the H0”)  **Significance level** (α) is the criterion compared against the p-value. The null hypothesis is rejected if p-value is lower than α.  α reflects the probability of rejecting the null hypothesis given that it is true (**Type I error**).  **Scope of inference**  **Internal validity**: are assumptions of the test satisfied?  **External validity**: possible to make inference to a broader population?  **Sampling distribution** of a statistic is a (reference) distribution that arises from a chance mechanism used to select a random sample from a population.  **Type I Error** (or significance level α): Probability of rejecting the null hypothesis, when the null hypothesis is true.  **Type II Error** (β): Probability of failing to reject the null hypothesis, when the null hypothesis is false.  **Power (1-β):** Probability of rejecting the null hypothesis, when a particular alternative hypothesis is true.  If alpha = 0, beta = 1; if alpha = 1, beta = 0; if alpha increases, beta decreases (power increases).  Increase power by: increase alpha; do one-tailed test; increase effect size; increase sample size  Most to least power, if assumption of t-tools is met  For small samples:  1. T-tools  2. Permutation test  3. Rank-sum test (or signed-rank test for paired data)  4. (Sign test for paired data)  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 1.58.19 PM.pngLarge samples rejection rates are basically equivalent | **Individual confidence level** is a probability that a single confidence  interval covers the true value.  **Familywise confidence level** is a probability that all confidence  intervals cover the corresponding true values.  **Common distributions and calculations:**  Population Variance: Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 10.14.46 AM.png  Sample Variance:  Sample variance is chi-sq distributed: Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 10.17.15 AM.png  Let Z0~N(0,1), independent of variance X^2~X2n,, thenMacintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 10.19.56 AM.png  Chi-sq distribution with n df: Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 10.54.02 AM.png  Pooled sample variance: Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 10.59.25 AM.png, where  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 11.00.13 AM.png  Sampling distribution of pooled sample variance: Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 11.02.44 AM.png  F-distribution: if Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 11.18.46 AM.png, independent of each other, then  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 11.17.47 AM.png(F-distribution with nx and ny d.f) |

|  |  |  |
| --- | --- | --- |
| **(Fisher’s) Randomization Test**: is a distribution-free test for treatment effect in **randomized experiments.**  **H0** : Zero treatment effect for all units, δ=0. Each unit’s outcome is the same, regardless of the treatment assigned. Consequently, the distribution of outcomes is identical in two groups  **Ha** : Non-zero treatment effect for ALL units, δ≠0 .  Assumptions:   1. Random assignment to groups. 2. Under the H0, independence of study units.   **Test statistic:** Difference between average outcomes in the two groups\*.Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 9.51.45 AM.png  \*In fact, any summary of the data can represent a test statistic.  **Randomization distribution** is a reference distribution of a test statistics in a randomization test, where variation is due to random assignment of the treatment.  **Permutation test** is a distribution-free nonparametric test for association between group status and outcome in **observational studies.**  **H0 :** Outcomes not related to group status.  **Ha :** Outcomes are related to (associated with) group status.  Assumptions: Independence of study units.  **Test statistic:** Difference in average observed outcomes  between the two groups (or, any other statistic),  Permutation distribution is a reference distribution of a test statistics in a permutation test.  *NOTE: Permutation tests are actually evaluating whether the distributions are equivalent, not just the equality of the means:* | **One Sample z-Test**  H0: **μ = μ0**  HA: **μ != μ0**  **Test Statistic:** Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 10.09.10 AM.png (make sure to double p-value if doing 2-sided test, of ½ alpha)  Exact sampling distribution of Z under H0: Z~N(0,1)  **Assumptions:**  1. Independence  2. Known population variance, σ^2  3. Normality (CLT allows for deviations!)  4. Random sampling from population  **One Sample t-Test**  H0: **μ = μ0**  HA: **μ != μ0**  Test Statistic: Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 10.21.43 AM.png  Exact sampling distribution of T under H0: T ~tn-1 (t-distribution with df = n-1)  **Assumptions:**  1. Independence  2. Normality (CLT allows for deviations!)  3. Random sampling from population | Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 10.33.14 AM.png**Two Sample z-Test**  H0: **μx - μy = μ0**  **HA: μx - μy != μ0**  Test Statistic:  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 10.37.39 AM.png**Pooled Two-Sample t-Test**  H0: μx - μy = μ0  HA: μx - μy != μ0  Test Statistic:  Exact sampling distribution of T under H0: T ~tnx+ny-2 (t-distribution with df = nx + ny -2)  Note:s Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 10.40.24 AM.png,  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 10.41.22 AM.pngis an unbiased estimator of Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 10.41.29 AM.png  Assumptions: same as Welch’s, but  If using pooled variance – assume equal variances between populations.  **Unpooled (Welch) Two-Sample t-Test**  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 10.43.26 AM.pngH0: μx - μy = μ0  HA: μx - μy != μ0  Test Statistic:  Approximate Sampling distribution:  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 10.44.34 AM.png  Assumptions:  1. Independence between units: (within each population (Yi with Yj and Xi with Xj)) And )between populations (Yi with Xj)  2. Homogeneity of units within each population:(Equal means within each population; Equal variances within each population;  If using pooled variance – assume equal variances between populations.  3. Population distributions are Normal:CLT allows for deviations.  4. Random sampling from populations.  **Paired t-Test**  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 10.55.48 AM.png  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 10.56.45 AM.pngH0: μD = μ0  HA: μD != μ0  Test Statistic:  Exact sampling distribution of T under H0: T~tn-1 (where n is # of pairs)  Assumptions same as one-sample t-test |

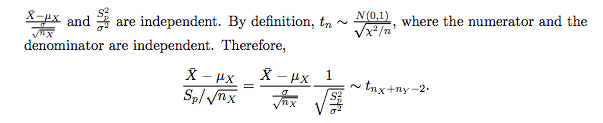
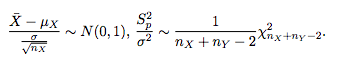
|  |  |  |
| --- | --- | --- |
| **F-test for equal variance** (could also use levene’s test): (caution: will reject for large samples when ratio is close to 1)  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 11.16.19 AM.png  Same assumptions as t-test:  In testing the equality of population variances, two assumptions are required: independent samples and normally distributed populations.  2-sided p-value is just 2\*the one-sided value.  **Test Normality:**  Shapiro-wilk test, Anderson-Darling test, Kolmogorov–Smirnov Test | **CLT:**      **How does CLT apply?**  **# checking assumptions:**  **Independence:** plot subgroups, and see if relationship still holds (cluster effect)  Plot data over time (serial effects)  Plot data vs space (spatial effect)  **Normality:** Histograms should look normal – overlay with normal curve and kernel density plot  Boxplot or Q-Q plot will show outliers  **Equal Variance:** divide variances to see if ratio is <0.5 or > 2  **T-test is robust:**  -if populations are symmetrically skewed  -if sample sizes are equal (and large)  -when sample sizes are equal, pooled-ttest is robust to unequal variances  **Sensitive:** When the sample sizes are not equal, t-tests are more sensitive to skewedness and long-tailedness.  For small samples, t-tests are somewhat sensitive to markedly different skewedness in two groups.  Watch out for outliers.  **Robustness of F-tests:**  Not resistant to outliers. Fairly robust to non-normality. Independence w/I and b/w groups is necessary. Equal variance is necessary.  **F-Test Assumptions:**  Equal variances is crucial  Not resistant to outliers  Normality is not critical. | **How does CLT apply?**  **# checking assumptions:**  **Independence:** plot subgroups, and see if relationship still holds (cluster effect)  Plot data over time (serial effects)  Plot data vs space (spatial effect)  **Normality:** Histograms should look normal – overlay with normal curve and kernel density plot  Boxplot or Q-Q plot will show outliers  **Equal Variance:** divide variances to see if ratio is <0.5 or > 2  **T-test is robust:**  -if populations are symmetrically skewed  -if sample sizes are equal (and large)  -when sample sizes are equal, pooled-ttest is robust to unequal variances  **Sensitive:** When the sample sizes are not equal, t-tests are more sensitive to skewedness and long-tailedness.  For small samples, t-tests are somewhat sensitive to markedly different skewedness in two groups.  Watch out for outliers.  **Robustness of F-tests:**  Not resistant to outliers. Fairly robust to non-normality. Independence w/I and b/w groups is necessary. Equal variance is necessary.  **F-Test Assumptions:**  Equal variances is crucial  Not resistant to outliers  Normality is not critical. |

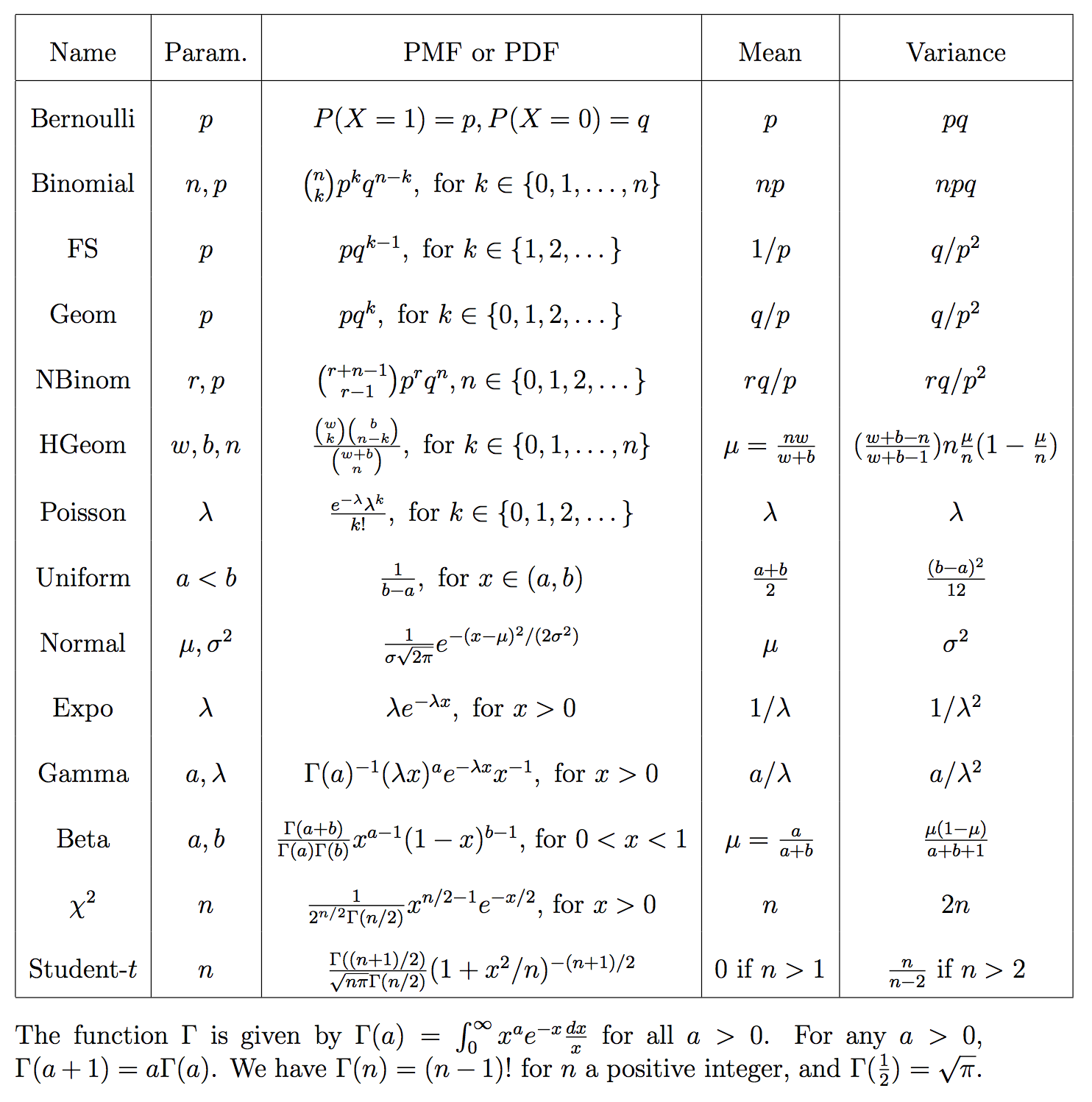
|  |  |
| --- | --- |
| **Confidence intervals**: Yi ~iid N(mu, sigma^2) : If we could repeat the experiment with randomizations, 95% of our confidence intervals could contain the estimand.  One Sample t-test:Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 11.31.48 AM.png  Two Sample t-test: Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 11.33.41 AM.png  **Two-Sample t-test for I samples (I > 2) Conf int**  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 3.06.24 PM.png | **Transformations:**  Log Transformation:   1. **Randomized Experiments** (T=Treatment, C = Control)   Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 11.36.30 AM.png Responses: Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 11.37.33 AM.png  Interpretation: Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 11.38.30 AM.png -- The response of an experimental unit to the treatment will be exp(ZbarT – ZbarC) times as large as its response to the control   1. Observational studies with **random sampling**:   Interpretation: Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 11.40.25 AM.png -- The median of the second population is exp(Zbar2 – Zbar1) times as large as the median of the first population.   1. In paired t-test: Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 11.45.17 AM.png   **Randomized Experiments**:Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 11.43.35 AM.png  **Observation studies w/ rand. Sampling**: Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 11.44.48 AM.png  Square Root Transformation sqrt(Y): Good for moderately skewed data, but hard to interpret results  Reciprocal Transformation 1/Y: good for severely skewed data, can be used with negative data  Logit log(Y/(1-Y)) or arcsin(2Y-1): good for proportions |

|  |  |
| --- | --- |
| **Nonparametric alternatives to t-tools: Rank Sum test:**  Hypotheses:  H0: (if treatments are randomized): Same as fisher’s randomization test  H0: (if treatments are not randomized): same as permutation test  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 11.55.31 AM.png\* If shapes and spreads of the two populations are similar: HA: There is a difference in medians between the two populations.  Transform data:  Ties are averaged  Test statistic: Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 11.56.12 AM.png  Assumptions: Independence  Normal approximation to rank Sum test (if no ties, and samples greater than 10):Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 1.27.17 PM.png  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 1.28.53 PM.pngWith ties, Rbar is sample mean (mean of all ranks) and Sr^2 is sample variance for combined set of nx + ny ranks: Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 1.28.13 PM.png  We can use this Z-statistic: | **Sign test for paired data** (lower power than Wilcoxon signed-rank test)  **General formuation: Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 1.33.50 PM.png**  **If differences are continuous rv’s: Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 1.33.59 PM.png**  **If differences are symmetric: Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 1.34.03 PM.png**  **Test Statistic: K = number of positive differences.**  **Exact distribution of K: K ~Binom(m, 0.5), where m is final number of pairs with nonzero Di.**  **Normal approximation (for large samples): Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 1.38.53 PM.png**  **Assumptions:**   1. The differences *Di* are assumed to be independent. 2. Each *Di* comes from the same continuous population. 3. The values *X*i and *Y*i represent are ordered (at least the [ordinal scale](http://en.wikipedia.org/wiki/Ordinal_scale)), so the comparisons "greater than", "less than", and "equal to" are meaningful. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Wilcoxon Signed-Rank Test for Paired Data:**  (more assumptions, but higher power than sign test)  **Hypotheses: General formulation:**  **H0:**The rank of the magnitude of within-pair difference is unrelated to the sign of the difference.  **Ha:** The rank of the magnitude of within-pair  difference is related to the sign of the difference.  **If differences are continuous r.v.’s:**  **H0**: Median(Di)=0  **Ha**: Median(Di)≠0 (or >0, or <0)  **If the distribution of differences is symmetric :**  **H0**: E(Di)=0  **Ha:** E(Di)≠0 (or >0, or <0)  1. Calculate differences (+ or -).  2. Rank the absolute values of differences |Di|  3. Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 1.43.03 PM.png  Test Statistic: (sum of ranks for positive differences)  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 1.43.08 PM.png  **Exact sampling distribution**: do permutation test, and switch group status for each pair.  **Normal approximation** (for large sample sizes, m is >= 20, where m is the number of non-zero differences):  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 1.49.05 PM.png  Assumptions:   1. Data are paired and come from the same population. 2. Each pair is chosen randomly and independently. 3. The data are measured at least on an [ordinal scale](http://en.wikipedia.org/wiki/Ordinal_scale), but need not be normal.   **Kruskal-Wallis Test** (Nonparametric ANOVA)  H0: Median(Y1) = Median(Y2) = …=Median(YI)  HA: at least one median is different from the others  Test Statistic – convert to ranks, and do ANOVA on ranks:  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 3.36.59 PM.png  SR^2 is the sample variance of **all** ranks (not pooled)  Assumptions:  1. Independence.  2. Random Sampling | | **Two-Sample t-Test for I samples**  H0: μm - μk = μ0  HA: μm - μk != μ0  Test Statistic: Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 2.20.19 PM.png  Exact sampling distribution of T under H0 is:  T~tn1 + n2 + ….+nI – I, where I is number of groups.  **Assumptions:**  Independence of units within each group;  Independence of units between groups;  Normality of all I populations;  Equal variances for all I populations;  Random sampling.  **Anova F-test for Equality of All means in I samples**  H0: μ1 = μ2 = …=μi (reduced model – equal means model)  Ha: at least one mean is different from the others (full model – separate means model)  **SSR**full = SSW (within groups)  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 3.15.31 PM.pngMacintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 3.15.55 PM.png  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-15 at 7.59.50 PM.png  **SSR**reduced = SST (**total error**)  S^2 is the sample variance of the entire sample, taken as one group:  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 3.17.21 PM.png  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 3.18.01 PM.png  **ESS** = SSB (between groups)  ESS = SSRreduced - SSRfull  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 3.20.17 PM.png  Test Statistic:  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 3.20.50 PM.png  MSB = Mean square between  MSRfull=Mean square residuals (within groups)  Easy way to calculate: MSRfull == Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 3.25.15 PM.png  Exact sampling distribution of R under H0:  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-12 at 3.23.21 PM.png  **Assumptions (check w/ Q-Q or residual plot):**  1. Normality of populations;  2. Equal population variances for all groups;  3. Independence within each group;  4. Independence between groups;  5. Homogeneity within each group;  6. Random sampling | **F-test for Equality in a subset of groups:**  **“Are groups 2-I all the same?”**  **Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-15 at 7.39.10 PM.png**  **Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-15 at 7.41.36 PM.pngTest statistic:**  **Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-15 at 7.40.54 PM.png**  **Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-15 at 7.42.02 PM.png**  **ANOTHER VERSION: F-test for Equality in a subset of groups:**  **“Is group 1 different from all other groups?”**  **Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-15 at 7.44.25 PM.png**  **Note: SSRfull now is equal to SSRreduced from above**  **Hypothesis test for linear combinations of means:**  **“Is one group mean equal to the average of a combination of other group means?”**  A linear combination of group means is : Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-15 at 8.07.06 PM.png  , where Ci’s are fixed coefficients chosen by the researcher  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-15 at 8.08.14 PM.png  If all coefficients sum to 0, then the linear combination is called **a contrast**.  Estimate of linear combination:  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-15 at 8.09.35 PM.png  Sampling distribution of G under H0:  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-15 at 8.09.40 PM.png  **T-test for Linear combination of means:**  **Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-15 at 8.11.15 PM.png**  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-15 at 7.59.50 PM.png **(same as F-test s-pooled)**  **Approximate sampling distribution:Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-15 at 8.14.44 PM.png**  **Assumption: Same as ANOVA F-test** | |
| **Mulitple Comparisons:**  **Familywise Type-I error**  **Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-15 at 8.20.07 PM.png**  **N perfectly dependent tests <= probability of at least one type-I error among N indep. Tests <= max familywise error**  **Bonferroni: (When all else fails)**  Set individual significance levels at alpha/N, where N is the # of tests  For pairwise mean comparisons, half width of CI = Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-15 at 8.25.59 PM.png | **Tukey HSD (Honest Significant Difference): (when interested in all pairs of groups)**  - Considers the largest difference between any two sample means for I groups  - ensures familywise conf leve = 1-alpha  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-15 at 8.30.04 PM.png, nbar is # in each sample), I is # of groups  Half –width for CI:  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-15 at 8.31.42 PM.png  Assumptions:  Normality, equal variances, equal sample sizes  **Tukey-Kramer Procedure (like Tukey HSD, but no need for equal sample sizes):**  **Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-15 at 8.34.46 PM.png** | | | Formulas:  N Choose k:  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-15 at 11.00.59 PM.png |

|  |  |  |
| --- | --- | --- |
| **Log-transformed, observational t-test**  **Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-15 at 11.05.44 PM.png** | **When does pooled t-test rejection rate differ:**  If the variances are actually different, the pooled t-test may be invalid. If the larger variance comes from the group with smaller sample size, the t- test tends to reject the null too often. (One point was taken off if you didn’t explain that the rejection rate tends to be too high when the larger variance is in the group with the smaller sample size; the rejection rate tends to be too low when the larger variance is in the group with the larger sample size). | **Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-10-16 at 11.06.30 AM.png** |

**Derive test stat for paired t-test**

****

**Derive distribution for T^2**

|  |  |  |
| --- | --- | --- |
| **Linear Regression:**  Line of best fit is found by minimizing sum of squared residuals (SSR)  SSRes =  =  “residual standard error” == (sigma-hat) == sqrt(residual variance)  Estimate of residual variance:    Sampling distribution of the sample variance:    Simple Linear Regression: describes the mean of Yi as a linear function of Xi    = residual variance        Exact Sampling Distributions:        T-tests for Least Squares estimates    OR in R: t-stat = (estimate)/ SE  T-test for mean response at a particular X-value:      Confidence Intervals for Least Squares estimates:    Confidence interval for mean response at a particular X-value (X0)    Confidence Band for Mean Response (Scheffe’s Method)    Confidence interval for PREDICTED future response:    Confidence Band for PREDICTED future response (Scheffe’s)    As the sample size goes to infinity, the width of the confidence interval for mu{Y|X=X0} goes to zero and the width of the prediction interval for Pred{Y|X=X0} goes to 2z0.975σ (.975 quantile of Z dist)  Inverse Prediction (called Calibration, or predicting X from Y)  1. Can refit the regression, switching Y and X  2. Can draw confidence bands and estimate Y-interval graphically (works best closer to the mean of X) | **Linear regression with binary X’s will give same t-stat and p-value of a t-test with equal variances**  Alternative interpretation of Beta(1)  , where  Assumptions:  1. **Linearity** – Check: look at scatterplot for outliers/nonlinearity  If violated: 0 and1 , and predictions may be biased  Strategies: transformations(log(x), 1/x, x^2, log(y), 1/y); add interactions; use nonlinear functions of X – spline or polynomial regression, generalized additive models.  2**. Independence of errors** – observations of Yi and Xi don’t travel together  Check: design of study. Plot residuals vs. time/distance.  See if units interact in some way.  See if there is spatial or temporal proximity  Is there a common data-generating source.  Are units clustered (or related).  If violated: Doesn’t change B1 or B0, but it does change SE and CI’s and results of t-tests. If errors are positively correlated, we have less information thatn we presume, resulting in inflated type I error.  Dealing with violation: Add more predictors (model the dependencies), multilevel models, repeated measures ANOVA, model dependencies.  3**. Equal variance of errors** Var(i) = σ2  -check: Scatterplot, residual plot (y = residuals, x = predicted values)  -if violated: SEs and CIs are wrong. B0 and B1 estimates are not different.  - Prediction interval is very sensitive, compared to confidence interval of mean  -strategies: Transformations, weighted regression.  4. Normality of errors (least important): i ~N(0, σ2)  -check w/ Q-Q plot  -if violated: B0 and B1 not effected, and due to CLT, SE not affected much (unless it’s really long-tailed or there are outliers)  - Prediction is more sensitive  -Strategies: ignore,, use robust regression  5. Random Sample  Interpreting Q-Q plots:  Long-Tailed (hist is high in middle  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-12-17 at 10.42.55 AM.png  Short-Tailed:  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-12-17 at 10.46.24 AM.png  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-12-17 at 10.49.54 AM.png  More Q-Q plots    Log transformation for linear regression:  Transform X:  Interpretation of : Multiplicative effect. Doubling X will result in a B1\*log(2) change in Y  Transform Y: If log(Y) is symmetric, then : Increase of X by 1 unit is associated with a multiplicative change in median{Y|X} by exp(B1).  -Changing X by C units will be associated with a change in the median ratio of e^(C\*B1)  Transform both X and Y:        Alternatively, after log-log transformation (when -20<B1<20), a one percent increase in X yields a change in the median of Y by B1 percent.  Doubling X is associated with a multiplicative change in Median{Y|X} by 2^(B1)  R^2: The proportion of variation in the response, Y, explained by the model for the means.    R^2 = r^2 (correlation coef for simple linear regression) | **Pearson Correlation:**    **Estimation of Pearson correlation:**          **Connection b/w sample correlation and least squares estimates (note: Sx and Sy are sample variances of X and Y)**    **Regression Line for Standardized Variables (slope is corr):**  **-> A one unit change in tilde(X) == a one-sd change in Xi**  **Regression effect (regression toward the mean)** implies that if you take new measurements of something that can vary, the mean of the group that measured lower than population mean will go up, and the mean of the group that measured higher than population mean will go down.  Sum of Squares Decomposition for Linear Regresssion:  **SSR for full model:**      **SSR for single-mean model:**    S^2 is the Var(Y variable)  **SSR between the linear model and single mean model:**  **, where H0 is mu = B(0)**    **R^2: the proportion of variation in Y explained by the model.**    **R^2 = [Sy^2 (n-1) – sigmaHat^2\*(n-1)]/[Sy^2 (n-1)]**  **Lack of Fit F-Test (when you have replicates at several levels of x)**      **Equal means model vs. Simple Linear Regression (testing if B1 = 0):**    **SSReg = SSR(lm) – SSR(equal Means)**  **SSR(full) = Simple Linear Regression**  **\* Output is F-statistic from linear regression summary in R.**  **Simple Linear Regression vs. Separate Means Model (like ANOVA): Lack of Fit F-Test**    **LR = Linear Regression (full)**  **SM = Separate Means (SSReg)**  **Exact Sampling Distribution under H0 is**  **Separate means vs. Equal means model:**  **ANOVA F-Test (see earlier stuff)** |

|  |  |  |
| --- | --- | --- |
| **Multiple Regression:**      **, where K = # of covariates (B1 through Bk)**  **Exact sampling distribution of residual variance:**    **Interpretation:** For a subpopulation of cars with Mileage=Mileage0, the difference in average prices of cars with and without cruise control is estimated to be $9,951(or Beta(2))  **Interaction:** Coefficients will be the same as fitting two models, but residual variance will be different.  Calculations for Multiple Regression:  **X is design matrix.**              **Assumptions: Exactly the same as linear regression.**  1. Linearity: check (pairwise scatterplots, trellis graphs to look at subpopulations)  2. Independence and 5. Random sampling: check == study design.  3. Equal variance and independence: check residual plot  4. Normality: check with Q-Q plot  \* Multiple regression on Categorical variables gives same p-value and F-stat as aov()  Multiple regression with multiple continuous variables and no interaction:    With interaction:    **\* When x’s are centered, the slopes are interpreted with reference to the average values of covariates.**  **\* Centering doesn’t affect the overall model fit (F-statistic)**  T-tests and CIs about Individual Coef’s for multiple regression:  == estimate / std. error    Standard errors for coefs == (it’s the sqrt of diagonal of the var/cov matrix) | Tests for Linear Combinations of Coefficients:  C is a column-vector of numbers    Example    Confidence interval for mean of Y at a specific values of X’s:    Confidence band with Scheffe’s method    Prediction interval for Y at specific X’s:    Prediction band with Scheffe’s correction:    Extra Sum of Squares Test for hierarchical models:  H0: Simpler (reduced) model is better  H1: Full model is better  In R: anova(simple\_model, full\_model)  \*\* If you only drop one term, then the F-stat will be the t-stat^2 for that term.  \*\* Note: t-stat == estimate / standard error    Test statistic for Extra Sum of Squares (note, Normality assumption must hold!):    Under H0, R ~  Adjusted R^2 (sigma-hat(regression) = sigma-hat full):  Where SSR(reduced) = , which is from the single (equal) mean model.  **If you’re deciding which variables to include by hand:**  1. Keep variable is sign makes sense (significant or not)  2. Keep if significant, but sign is unexpected (could indicate additional interactions, unobserved confounder, or ecological fallacy).  3. Remove if it’s insignificant and sign in unexpected.  4. Consider interactions for inputs with large effects.  5. Interpretation of coefficients is very difficult if there is multicollinearity.  \*When explanatory variables are related, it may be impossible to “hold all other variables fixed” while changing one of them – it is “outside of the experience provided by the data”  **Techniques for variable selection (sensitive to outliers and influential points):**  1.Fixed set by design (treatment indicator + background variables);  2.Fit all possible subsets of models and find the one that fits the best according to some criterion: E.g., Adj-R2, Cp statistic, AIC, or BIC, F-stat greater than 4  3. Sequential: forward / backward / stepwise selection;  Total Number of p-parameter hierarchical models:      **Ecological Fallacy:**  Occurs when one makes conclusions about individuals based only on analyses of aggregated group data.  Simpson’s paradox  Group-level correlations vs. individual correlations  Group average vs. individual likelihood | **Partialing Out:**      Bj(hat) measures the sample relationship between Y and Xj after all other predictors have been partialled out.        **Detecting Multicollinearity:**  Pair-wise correlations are > 0.5 (or > 0.4 as suggested by some sources).  Large changes in estimated regression coefficients when a new predictor variable is added or deleted.  Significant p-value for the extra-sum-of-squares F- test after dropping two or more insignificant regression coefficients.  R^2 may be nearly as large after dropping important predictors from the model.  Coefficient may be significant in SLR (simple linear regression) but insignificant in MLR (multiple linear regression) for the same predictor.  Note that polynomial terms, interactions and indicator variables for a multi-level factor, are likely to be highly collinear.  VIF:    High VIF (say, above 5 or 10) indicates a multicollinearity problem  Mean Square Ratio:      Dealing with multicollinearity:  In general, the problem of collinearity can not be solved.  \*Leave the model as is if only interested in predicting Y.  \*If interested in relative importance of predictors and  inference about their “effects”: \*Examine pairwise correlations and combine correlated variables, if possible;  \*Drop “redundant’’ explanatory variables;  \*Centering around sample averages helps to decrease  correlations among polynomial terms.  \*Get more data. |

|  |  |  |
| --- | --- | --- |
| Dealing with influential points:    Finding influential points:  Leverage flags cases with unusual explanatory variables (X).  Studentized residuals flag cases with unusual responses (Y).  Cook’s distance measures overall influence - the effect of omitting case i on regression coefficients and, therefore, on fitted values.  Leverage (hi) of Simple linear Regression:    Leverage for Multiple Linear Regression:    Var(Y-hat-i) can be found by centering the data at hi and squaring the SE of the intercept.  \*We can regard hi as a scale-free measure of a distance between (Xi1, Xi2, ..., XiK), the vector of predictors for subject i, and 1,2,...,K , the vector of sample averages in the dataset.  Studentized Residuals – should see about 5% outside + or - 2:    Cook’s Distance:    Measures the effect of omitting case i on regression coefficients and, therefore, fitted values.  Will be large if either leverage or studentized residual is large, or both.  Measures the influence on all the coefficients.  Alert for observations if:  Alert if  Leverage hi  2(K 1)/n (twice the expected value)  Cook’s distance > 1  Studentized residuals |stud - ri | 2  Weighted Regression:  Use if there is Heteroskedasticity  Least squares estimators of β are unbiased and consistent;  Usual estimators for the standard errors of least squares are biased, so the usual confidence intervals and test statistics are incorrect, and may lead to incorrect  conclusions.  When variances are unequal, weights known.    When to use weighted regression (weights = wi):  \*Responses are estimates, with variances known,  sigma^2 /wi = sigma(i)^2  \*Responses are averages, with known sample sizes, wini  \*Variance proportional to some X (or X2), wi  f(Xi)  \*Known survey weights.  \* If their magnitudes increase proportional to one of the Xj‘s, fit a WLS with weights 1/Xj.  \*\* Use Iteratively Reweighted Least Squares (IRLS), rls() in R.  Can also use regression with t-errors, where weights are considered random: | **Example Problems:**  **Derive least squares estimate for slope when B0 = 0.**    **Sample correlation becomes:**  **Alternatives:**      **Shortcut Equations:**    **Derive least squares estimator for B1:**  **1. take partial derivative of sum of squared residuals:**        **Derive intercept for least squares:**  **Take partial derivative w.r.t B0:**  **Set = 0 and solve for B0**  **means B0 = Ybar – B1\*Xbar** | When to transform:  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-12-18 at 8.59.46 AM.png try e^y~x  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-12-18 at 1.40.31 PM.png OR this:Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-12-18 at 9.02.34 AM.png  ^log-transform y (or sqrt)  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-12-18 at 9.05.45 AM.png transform to 1/y  Macintosh HD:Users:callinswitzer:Desktop:Screen Shot 2014-12-18 at 1.52.27 PM.pngtransform x (1/x or log(x)) – mean is curved, but SD is approximately constant. |