STA 104: Applied Statistical Methods: Nonparametric Statistics

Course Material Summary

University of California at Davis

Last Edit Date: 02/21/2022

STA 104 Summary

Topic	Content
Parametric vs non-parametric	Non-parametric statistics uses techniques that do not require typical assumptions of traditional techniques. In a traditional test, we assume: 1) Random sample was taken, or equivalently the X_i values are independent. 2) The sample size $n \geq 30$ (CLT), and the population is normal When we do not have these assumptions above, the distributions based on CLT cannot be used, which means we need to assume a named distribution. This is what non-parametric do. It is often called "distribution free". When assumptions are NOT violated, the parametric tests have more power. When assumptions are violated, the non-parametric tests have more power. * $Power = 1 - P(Type\ II\ error) = P\ (Reject\ H_0 \mid H_0\ False)$
Test for single median	
Binomial test	We use median because we do not have these assumptions in parametric tests. Median is actually a proportion, which is 50%. θ_m denotes all possible medians. θ_m^* is the hypothesized median. n is the sample size. **Assumption:** A random sample was taken from the population (equivalently, observations are independent). **Step 1: State the null and alternative hypothesis** $H_0: \theta_m = \theta_m^* \ \ v.s.\ \ H_1: \theta_m \neq \theta_m^*\ (H_0: p = 0.5\ \ H_1: p \neq 0.5)\ (two-sided)$ $H_0: \theta_m \leq \theta_m^* \ \ v.s.\ \ H_1: \theta_m > \theta_m^*\ (H_0: p \leq 0.5\ \ H_1: p > 0.5)\ (right\ tail,\ one-sided)$ $H_0: \theta_m \geq \theta_m^* \ \ v.s.\ \ H_1: \theta_m < \theta_m^*\ (H_0: p \geq 0.5\ \ H_1: p < 0.5)\ (left\ tail,\ one-sided)$ **Step 2: Calculate the test statistic** $B^+ = \#\ of\ X_i > \ \theta_m^*, B^+ \sim Binomial(n, \frac{1}{2})$ **Step 3: Calculate p-value** $H_1: \theta_m \neq \theta_m^*\ (H_0: p = 0.5\ \ H_1: p \neq 0.5)\ \ => \ \ p-value = \ 2(\min\{\mathrm{P}(\mathrm{X} \geq \mathrm{B}^+),\ \mathrm{P}(\mathrm{X} \leq \mathrm{B}^+)\})$ $H_1: \theta_m > \theta_m^*\ (H_0: p \leq 0.5\ \ H_1: p > 0.5)\ \ => \ \ p-value = \ \mathrm{P}(\mathrm{X} \geq \mathrm{B}^+)$ $H_1: \theta_m < \theta_m^*\ (H_0: p \geq 0.5\ \ H_1: p < 0.5)\ \ => \ \ p-value = \ \mathrm{P}(\mathrm{X} \leq \mathrm{B}^+)$ Interpretation: If the true median equals to θ_m^*, we can observe our data or more extreme with probability p-value. **Step 4: Reject H_0 if p-value $< \alpha$**
Normal Approximation to Binomial test	When we have a reasonable sample size, we may assume $B^+ = N(np, \sqrt{np(1-p)})$ based on CLT, where $p = 0.5$ under H_0. We now can use a z distribution.

	Assumption: A random sample was taken from the population (equivalently, observations are independent), and there are at least 5 observations above and below the hypothesized median. **Step 1: State the null and alternative hypothesis** $H_0: \theta_m = \theta_m^* \;\; v.s. \;\; H_1: \theta_m \neq \theta_m^* \; (H_0: p = 0.5 \;\; H_1: p \neq 0.5) \; (two-sided)$ $H_0: \theta_m \leq \theta_m^* \;\; v.s. \;\; H_1: \theta_m > \theta_m^* \; (H_0: p \leq 0.5 \;\; H_1: p > 0.5) \; (right\ tail,\ one-sided)$ $H_0: \theta_m \geq \theta_m^* \;\; v.s. \;\; H_1: \theta_m < \theta_m^* \; (H_0: p \geq 0.5 \;\; H_1: p < 0.5) \; (left\ tail,\ one-sided)$ **Step 2: Calculate the test statistic** $Z_S = \frac{S-n(0.5)}{\sqrt{n(0.25)}},\ where\ S = B^+ = \#\ of\ X_i > \ \theta_m^*, B^+ \sim Binomial(n, \frac{1}{2})$ **Step 3: Calculate p-value** $H_1: \theta_m \neq \theta_m^* \; (H_0: p = 0.5 \;\; H_1: p \neq 0.5) \;\; => \;\; p-value = \; 2P(Z > \lvert Z_S \rvert)$ $H_1: \theta_m > \theta_m^* \; (H_0: p \leq 0.5 \;\; H_1: p > 0.5) \;\; => \;\; p-value = \; P(Z > Z_S)$ $H_1: \theta_m < \theta_m^* \; (H_0: p \geq 0.5 \;\; H_1: p < 0.5) \;\; => \;\; p-value = \; P(Z < Z_S)$ Interpretation: If the true median equals to θ_m^*, we can observe our data or more extreme with probability p-value. **Step 4: Reject H_0 if p-value $< \alpha$**
Confidence Interval for median	Find a $(1-\alpha)100\%$ confidence interval for the median, using the normal approximation to binomial. **Step 1: Get the location** Lower bound location = $-z_{1-\frac{\alpha}{2}} * \left(\sqrt{0.25n}\right) + 0.5n$ Upper bound location = $+z_{1-\frac{\alpha}{2}} * \left(\sqrt{0.25n}\right) + 0.5n + 1$ **Step 2: Find the number in the rounded location to get the confidence interval** $(X_{lower\ bound\ location},\ X_{upper\ bound\ location})$
Estimation for Percentile and CDF	Find a $(1-\alpha)100\%$ confidence interval for the CDF at x. **Step 1: Get the proportion** $\hat{F}(x) = \hat{p} = \frac{\#\ of\ X_i \leq x}{n} \sim N(p, \sqrt{(p(1-p))/n})$ Step 2: **Get the lower and upper bound** Lower bound: $\hat{p} - z_{1-\frac{\alpha}{2}} * \sqrt{(p(1-p))/n}$ Upper bound: $\hat{p} + z_{1-\frac{\alpha}{2}} * \sqrt{(p(1-p))/n}$

	Step 3: Convert the lower and upper bound to percentile (lower bound percentile, upper bound percentile) $\times$ 100%
Confidence Intervals for percentiles	Find a $(1-\alpha)100\%$ confidence interval for the $(p^*)100^{th}$ percentile. **Step 1: Get the location** Lower bound location = $n(p^*) - z_{1-\frac{\alpha}{2}} * \sqrt{p^*(1-p^*)n}$ Upper bound location = $n(p^*) + 1 + z_{1-\frac{\alpha}{2}} * \sqrt{p^*(1-p^*)n}$ **Step 2: Find the number in the rounded location and get the confidence interval** $(X_{lower\ bound\ location}, X_{upper\ bound\ location})$ # When we get a location equals to 0 or n+1, we should use 1 or n as our location.
Tests for two groups	
Comparing two means	The goal is to determine whether two means are statistically different. Assumptions for parametric test are: 1) Random sample from both groups 2) Groups are independent 3) $\bar{X}_1$ and $\bar{X}_2$ are normal
Permutation test for two groups	Let $F_1(x)$ = CDF for group 1, $F_2(x)$ = CDF for group 2. If the distributions for the groups are equal, $F_1(x) = F_2(x)$. Both groups are from the same population. **Assumption:** A random sample was taken from each group, groups independent. **Step 1: state the null and alternative hypothesis** $H_0: F_1(x) = F_2(x)$ $v.s$ $H_1: F_1(x) \geq F_2(x)\ or\ F_1(x) \leq F_2(x)\ (two-sided)$ # distributions are different $H_1: F_1(x) \leq F_2(x)\ (right\ tail,\ one-sided)$ # group 1 tends to be larger than group 2 $H_1: F_1(x) \geq F_2(x)\ (left\ tail,\ one-sided)$ # group 2 tends to be larger than group 1 **Step 2: Calculate the observed statistic and all permutations** $D^{OBS} = \bar{X}_1 - \bar{X}_2$ or $D^{OBS} = total_1 - total_2$ or $D^{OBS} = median_1 - median_2$ Permutations = $\binom{m+n}{m} = \binom{m+n}{n} = \frac{(m+n)!}{m!n!}$ **Step 3: Calculate the permutation p-value** $H_1: F_1(x) \geq F_2(x)\ or\ F_1(x) \leq F_2(x)\ => \frac{\#\ of\ \lvert D_i \rvert \geq \lvert D^{OBS} \rvert}{permutations}$ $H_1: F_1(x) \leq F_2(x)\ => \frac{\#\ of\ D_i \geq D^{OBS}}{permutations}$

	$H_1: F_1(x) \geq F_2(x)$ => $\frac{\#\, of\, D_i \leq D^{OBS}}{permutations}$ **Step 4: If p-value < α, reject H_0.** # When we have asymmetric distributions, we use the median to compare outliers. Otherwise, we use total or mean. ## If the sample sizes of each group are the same, the test results from total and mean are the same. Otherwise, the results will be different.
Approximate Permutation Test	Steps for an approximate permutation test (for coding): 1) Record D^{OBS} 2) Create one vector of all observations 3) Randomly shuffle the (m + n) observations, and assign first m to group 1, last n to group 2 4) Compute D_i = observed difference (in means/medians/totals) 5) Repeat step 3 and 4, R > 2000 times 6) Based on these R random values of D_i, we have an approximate p-values are: $H_1: F_1(x) \geq F_2(x)\ or\ F_1(x) \leq F_2(x)$ => (# of $\lvert D_i \rvert \geq \lvert D^{OBS} \rvert$) / R $H_1: F_1(x) \leq F_2(x)$ => (# of $D_i \geq D^{OBS}$) / R $H_1: F_1(x) \geq F_2(x)$ => (# of $D_i \leq D^{OBS}$) / R 7) If p-value < α, reject H_0
Confidence Interval for p-value	A $(1-\alpha)100\%$ CI for a p-value p^* is: $p^* \pm z_{1-\frac{\alpha}{2}} \sqrt{p^*(1-p^*)/R}$
Normal Approximation to permutation	We use the overall mean $\bar{x}^*$ and overall standard deviation S^* in our test statistics. For this test, we need $n + m \geq 30$. **Assumption:** A random sample was taken from each group, groups independent. **Step 1: state the null and alternative hypothesis** $H_0: F_1(x) = F_2(x)$ $v.s$ $H_1: F_1(x) \geq F_2(x)\ or\ F_1(x) \leq F_2(x)\ (two-sided)$ # distributions are different $H_1: F_1(x) \leq F_2(x)\ (right\ tail,\ one-sided)$ # group 1 tends to be larger than group 2 $H_1: F_1(x) \geq F_2(x)\ (left\ tail,\ one-sided)$ # group 2 tends to be larger than group 1 **Step 2: Calculate the test statistic** $Z_S = \frac{\bar{x}_1 - \bar{x}^*}{S^*/\sqrt{m}}$ or $Z_S = \frac{\bar{x}_2 - \bar{x}^*}{S^*/\sqrt{n}}$ where $\bar{x}^* = \frac{1}{m+n}\sum x_i$, $S^* = \sqrt{\frac{1}{m+n-1}\sum(x_i - \bar{x}^*)^2}$ **Step 3: Calculate the permutation p-value**

	$H_1: F_1(x) \geq F_2(x)$ or $F_1(x) \leq F_2(x)$ $=>$ $p - value = 2P(Z > \|Z_S\|)$ $H_1: F_1(x) \leq F_2(x)$ $=>$ $p - value = P(Z > Z_S)$ $H_1: F_1(x) \geq F_2(x)$ $=>$ $p - value = P(Z < Z_S)$ **Step 4: If p-value < α, reject H_0.**
Wilcoxon Rank Sum (WRS) test	**Assumption:** A random sample was taken from each group, groups independent. **Step 1: State the null and alternative hypothesis** $H_0: F_1(x) = F_2(x)$ $v.s$ $H_1: F_1(x) \geq F_2(x)$ or $F_1(x) \leq F_2(x)$ $(two - sided)$ # distributions are different $H_1: F_1(x) \leq F_2(x)$ $(right\ tail,\ one - sided)$ # group 1 tends to be larger than group 2 $H_1: F_1(x) \geq F_2(x)$ $(left\ tail,\ one - sided)$ # group 2 tends to be larger than group 1 **Step 2: Calculate test statistic** 1) Combine the m + n values into one group 2) Calculate the rank for each data point: $R(x_i) = \#\ of\ data \leq x_i,\ i = 1, \dots, m + n$ Note: If there are ties, average the ranks of the tied observations, and assign the tied values as their ranks 3) Calculate the total rank in group 1 (arbitrary choice of groups). This is our test statistic, $W_{OBS} = \sum_{group\ 1} R(x_i)$. **Step 3: Calculate the exact p-value** Permutations = $\binom{m+n}{n}$, W_i = sum of rank in group 1 $H_1: F_1(x) \geq F_2(x)$ or $F_1(x) \leq F_2(x)$ $=>$ $2 * \min\left(\frac{\#\ of\ W_i \geq W_{OBS}}{\text{permutations}}, \frac{\#\ of\ W_i \leq W_{OBS}}{\text{permutations}}\right)$ $H_1: F_1(x) \leq F_2(x)$ $=>$ $\frac{\#\ of\ W_i \geq W_{OBS}}{\text{permutations}}$ $H_1: F_1(x) \geq F_2(x)$ $=>$ $\frac{\#\ of\ W_i \leq W_{OBS}}{\text{permutations}}$ **Step 4: If p-value < α, reject H_0.** # WRS tends to have higher power when the distribution is skewed, outliers are present, since assigning ranks essentially removes all influence of both issues. ## Permutation tests tend to have higher power when the distribution is thought to be symmetric, and when using the mean.
Large Sample Approximation to WRS	Let N = m + n, and $R(x_1), \dots, R(x_N)$ be the corresponding combined ranks of the two groups. Let S_1 = sum of ranks in group 1.

	Under the assumption that the distributions are equal, every $R(x_i)$ should have been equally likely to come from both groups. **Assumption:** A random sample was taken from each group, independent groups, combined sample size at least 30. **Step 1: State the null and alternative hypothesis** $H_0: F_1(x) = F_2(x) \quad v.s \quad H_1: F_1(x) \geq F_2(x)\ or\ F_1(x) \leq F_2(x)\ (two-sided)$ # distributions are different $H_1: F_1(x) \leq F_2(x)\ (right\ tail,\ one-sided)$ # group 1 tends to be larger than group 2 $H_1: F_1(x) \geq F_2(x)\ (left\ tail,\ one-sided)$ # group 2 tends to be larger than group 1 **Step 2: Our test-statistic is** $Z_s = \frac{W_{OBS} - E[S_1]}{\sqrt{\sigma_S^2}}$ where $W_{OBS} = \sum_{group\ 1} R(x_i),\ E[S_1] = m\mu_R,\ \sigma_S^2 = \frac{mn\sigma_R^2}{N-1}$ where $\mu_R = \frac{1}{N}\sum R(x_i)$, $\sigma_R^2 = \frac{1}{N}\sum(R(x_i) - \bar{x}_R)^2$ If $N \geq 30$, we have $S_1 \sim N(m\mu_R, \frac{mn\sigma_R^2}{N-1})$ # If there are no ties, then $E[S_1] = \frac{m(N+1)}{2}$, $\sigma_R^2 = \frac{mn(N+1)}{12}$ **Step 3: Get the p-value** $H_1: F_1(x) \geq F_2(x)\ or\ F_1(x) \leq F_2(x) \implies P(Z > \|Z_S\|)$ $H_1: F_1(x) \leq F_2(x) \implies P(Z > Z_S)$ $H_1: F_1(x) \geq F_2(x) \implies P(Z < Z_S)$ **Step 4: If p-value $< \alpha$, reject H_0.**
Mann-Whitney Test (alternative to WRS)	Let $X_1, \ldots, X_m$ be our sample from group 1. Let $Y_1, \ldots, Y_n$ be our sample from group 2. **Assumption**: A random sample was taken from each group, groups independent. **Step 1: State the null and alternative hypothesis** $H_0: F_1(x) = F_2(x) \quad v.s \quad H_1: F_1(x) \geq F_2(x)\ or\ F_1(x) \leq F_2(x)\ (two-sided)$ # distributions are different $H_1: F_1(x) \leq F_2(x)\ (right\ tail,\ one-sided)$ # group 1 tends to be larger than group 2 $H_1: F_1(x) \geq F_2(x)\ (left\ tail,\ one-sided)$ # group 2 tends to be larger than group 1 **Step 2: Calculate test statistic**

	$U_{MW} = (\# \, of \; pairs \; (X_i < Y_i)) + \frac{1}{2}(\# \, of \; pairs \; (X_i = Y_i))$ # If group 1 is lower than group 2, U_{MW} will be closed to the maximum # of pairs. ## If group 1 is larger than group 2, U_{MW} will be closed to 1. ### Number of possible pairs = m*n #### The test statistic is in U or Mann-Whitney Distribution. **Step 3: Calculate p-value** Let $U_{1-\frac{\alpha}{2}} = \left(1 - \frac{\alpha}{2}\right) 100^{th}$ percentile of U (upper) Let $U_{\frac{\alpha}{2}} = \left(\frac{\alpha}{2}\right) 100^{th}$ percentile of U (lower) $H_1: F_1(x) \geq F_2(x) \; or \; F_1(x) \leq F_2(x) \;\; => \;\; If \; U_{MW} > U_{1-\frac{\alpha}{2}} \; or \; U_{MW} < U_{\frac{\alpha}{2}} \;\; => \; < \alpha$ $H_1: F_1(x) \leq F_2(x) \;\; => \; If \; U_{MW} < U_{\frac{\alpha}{2}} \;\; => \;\; < \frac{\alpha}{2}$ $H_1: F_1(x) \geq F_2(x) \;\; => \;\; If \; U_{MW} > U_{1-\frac{\alpha}{2}} \;\; => \;\; < \frac{\alpha}{2}$ # Need to look at the table of Mann-Whitney Distribution. **Step 4: If p-value < $\boldsymbol{\alpha}$, reject $\boldsymbol{H_0}$.**
Kolmogorov Smirnov (KS) Test	Assumption: A random sample was taken from each group, groups independent. Distributions should be continuous. **Step 1: State the null and alternative hypothesis** $H_0: F_1(x) = F_2(x) \;\; v.s \;\; H_1: F_1(x) \geq F_2(x) \; or \; F_1(x) \leq F_2(x) \; (two - sided)$ # distributions are different **Step 2: Calculate test statistic** Let $\hat{F}_1(x)$ = empirical CDF of group 1 Let $\hat{F}_2(x)$ = empirical CDF of group 2 Then, 1) Combine the data from both groups 2) Calculate $\hat{F}_1(x)$ for both groups observations Calculate $\hat{F}_2(x)$ for both groups observations 3) Calculate the difference between $\hat{F}_1(x) - \hat{F}_2(x)$ for all observations 4) Our test-statistic is then $K_S = \max \left\lvert \hat{F}_1(x) - \hat{F}_2(x) \right\rvert$ **Step 3: Calculate p-value** The p-value is a permutation p-value: (# of $\left\lvert \hat{F}_1(x) - \hat{F}_2(x) \right\rvert \geq K_s$) / $\binom{m+n}{n}$ or divided by R if it's a random permutation test

<table>
<tr><td></td><td>Step 4: If p-value < α, reject H_0.</td></tr>
<tr><td>Confidence Interval for shift parameter</td><td>Step 1: Find all n*m pairwise differences, $X_i - Y_i$

Step 2: Order the pairwise differences, call them pwd(1), pwd(2), …, pwd(n*m)

Step 3: We want the locations, call them ka and kb
$P\big(pwd(ka) \leq \Delta \leq pwd(kb)\big) = 1 - \alpha \quad => \quad ka \leq U \leq kb - 1$
kb - 1 because of discrete data

Step 4: Confidence interval is $(ka = U_{\frac{\alpha}{2}} + 1,\ kb = U_{1-\frac{\alpha}{2}})$

If CI of Δ has both bounds > 0, then group 1 has larger distribution/measurement than group 2.
If CI of Δ has both bounds < 0, then group 1 has smaller distribution/measurement than group 2.
If CI of Δ has contains 0, then there is no significant difference between group 1 and group 2.</td></tr>
<tr><td>Choose an appropriate test</td><td>
<table>
<tr><th>Distribution</th><th>Statistic</th><th>Winner</th></tr>
<tr><td>Symmetric</td><td>Mean</td><td>Permutation</td></tr>
<tr><td>Symmetric</td><td>Median</td><td>Wilcoxon Rank Sum</td></tr>
<tr><td>Asymmetric</td><td>Mean</td><td>Wilcoxon Rank Sum</td></tr>
<tr><td>Asymmetric</td><td>Median</td><td>Permutation</td></tr>
</table>
</td></tr>
<tr><td colspan="2">Tests for three or more groups</td></tr>
<tr><td>ANOVA (non-parametric, permutation based)</td><td>Notation: Assume we have K groups
Let $X_{ij} = j^{th}$ observation from i^{th}group
Let n_i = sample size of i^{th}group
Let $\bar{X}_i$ = sample mean of i^{th}group
Let S_i^2 = sample variance of i^{th}group
Let N = overall sample size = $\sum_{i=1}^{k} n_i$
Let $\bar{X}$ = overall sample mean = $\frac{\sum_{i=1}^{k} n_i \bar{X}_i}{N}$

The idea is the same as parametric ANOVA. We compare the difference in means to the overall mean to the spread of each group.

The following measure the difference between groups:
SST = Sum of squared treatment = $\sum_{i=1}^{k} n_i\,(\bar{X}_i - \bar{X})^2$
$MST = \frac{SST}{k-1}$

The following measure the variances within each group:</td></tr>
</table>

	$SSE = \sum_{i=1}^{k}(n_i - 1)\, S_i^2$ = Sum of square errors $MSE = \frac{SSE}{N-k}$ **Test statistic:** $F_s = \frac{MST}{MSE}$ # When F_s is large => variance between groups is larger than within groups => means are significantly different ## When F_s is small => variance between groups is smaller than within groups => means are not significantly different **Assumptions (traditional):** 1) Random samples are taken from all k groups 2) All k groups are independent 3) $\sigma_1 = \sigma_2 = \cdots = \sigma_k$ equal variance (Levene's Test) 4) $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ independent and identically distributed (QQ plot and Shapiro-Wilks Test) # When the assumptions do not hold, we do not know what the distribution of F_s. But, we can find the permutation distribution. **Assumptions (non-parametric):** A random sample was taken from each group, groups independent. **Step 1: State the null and alternative hypothesis** $H_0: F_1(x) = F_2(x) = \cdots = F_k(x)\ \ v.s.\ \ H_1: F_i(x) \leq F_j(x)\ or\ F_i(x) \geq F_j(x)$ for some $i \neq j$ **Step 2: Calculate the observed test statistic** $F_{OBS} = \frac{MST}{MSE}$ **Step 3: Find the permutation p-value:** Possible permutations = $\frac{N!}{n_1!n_2!\dots n_k!}$ We can also use random permutations: 1) Randomly assign the N observations into the k groups, $R > 4000$ times 2) Calculate the R values of F_s, denote F_i 3) Our p-value is (# of $F_i \geq F_{OBS}$)/R **Step 4: If p-value < $\boldsymbol{\alpha}$, reject $\boldsymbol{H_0}$.**
Kroskall-Wallis (KW) Test (permutation based)	Kroskall-Wallis test uses ranks rather than the actual X_{ij} values. Has confidence interval. **Assumptions:**

	A random sample was taken from each group, groups independent. **Step 1: State the null and alternative hypothesis** $H_0: F_1(x) = F_2(x) = \cdots = F_k(x) \;\; v.s. \;\; H_1: F_i(x) \leq F_j(x) \; or \; F_i(x) \geq F_j(x)$ for some $i \neq j$ **Step 2: Calculate the test statistic** $KW_{OBS} = \frac{1}{S_R^2} \sum_{i=1}^{k} n_i \left(\bar{R}_i - \frac{N+1}{2}\right)^2$ Where S_R^2 = variance of ranks, regardless of groups, $\bar{R}_i$ = mean rank of each group # This form of KW test works whenever ties are present or not **Step 3: Calculate the approximate permutation p-value** p-value = (# of $KW_i \geq KW_{OBS}$)/R **Step 4: If p-value < α, reject H_0.**
Large Sample Approximation to Kroskall-Wallis Test	If the n_i's are large, but an assumption of ANOVA is violated, we may use a large sample approximation. Motivation: In traditional ANOVA, we know that SST/σ_ϵ^2 is distributed X^2 with df = k – 1. Now, replace X_{ij} with R_{ij}, which is the corresponding ranks, we can see: $SST_R = \sum_{i=1}^{k} n_i \left(\bar{R}_i - \frac{N+1}{2}\right)^2$ But the normalizing constant for the X^2 distribution has changed (since we are using R_{ij}) $E[c(SST_R)] = k - 1$ (since we know $E[X_{k-1}^2] = k - 1$), which gives $c = 1/S_R^2$ This gives our test statistic as: $KW = \frac{1}{S_R^2} \sum_{i=1}^{k} n_i \left(\bar{R}_i - \frac{N+1}{2}\right)^2 \sim X_{k-1}^2$ **Assumptions:** A random sample was taken from each group, groups independent, combined sample size at least 30. **Step 1: State the null and alternative hypothesis** $H_0: F_1(x) = F_2(x) = \cdots = F_k(x) \;\; v.s. \;\; H_1: F_i(x) \leq F_j(x) \; or \; F_i(x) \geq F_j(x)$ for some $i \neq j$ **Step 2: Calculate the test statistic** $KW = \frac{1}{S_R^2} \sum_{i=1}^{k} n_i \left(\bar{R}_i - \frac{N+1}{2}\right)^2$ Where S_R^2 = variance of ranks, regardless of groups, $\bar{R}_i$ = mean rank of each group

	Step 3: Calculate the p-value p-value = P($X^2_{k-1} > KW$) **Step 4: If p-value < α, reject H_0.**
Asymptotic Bonferroni and Tukey cutoffs (Corrections for multiple comparisons)	**Assumptions:** A random sample was taken from each group, groups independent, combined sample size at least 30. **Bonferroni (BON) cutoff:** BON = $Z_{1-\frac{\alpha}{2g}}\sqrt{S_R^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$ **Tukey (HSD) cutoff:** HSD =$q_\alpha(k, df = N-k)\sqrt{S_R^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$ **Parametric version of Tukey:** We reject H_0 if $\left\lvert\bar{X}_i - \bar{X}_j\right\rvert \geq q_\alpha(k, df = N-k)\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$ If $\left\lvert\bar{R}_i - \bar{R}_j\right\rvert > BON\ or\ HSD$, we may conclude that the specific group have significant different average ranks.
Permutation cutoff for Bonferroni and Tukey	**Assumptions:** A random sample was taken from each group, groups independent. There are $\binom{k}{2}$ possible permutations. Compare the p-values to $\frac{\alpha}{g}$. Step 1: Randomly shuffle each observation into a group, R>4000. Step 2: Pick a comparison measure, T_{ij}. Common values are $\lvert\bar{X}_i - \bar{X}_j\rvert$, $\lvert\bar{R}_i - \bar{R}_j\rvert$, $\lvert median_i - median_j\rvert$, $\frac{\bar{X}_i-\bar{X}_j}{\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}}$ Step 3: For each R permutation, calculate $Q_{ij} = \max\ \lvert T_{ij}\rvert$ Step 4: Let $q^*(\alpha)$ be the (1-α)100^{th} percentile of Q_{ij}. Then groups i and j are significant different if $\left\lvert T_{ij}^{OBS}\right\rvert > q^*(\alpha)$. We can also find the p-value = (# of $Q_{ij} \geq \lvert T_{ij}^{OBS}\rvert$)/R. If p-value < α, groups i and j are significant different.

Kroskall-Wallis v.s. Permutation	**The KW test will have higher power than a permutation test when:** 1) Outliers are present 2) The distribution of one or more groups is skewed 3) The distribution of one or more groups has “heavy tails”
Test for linear relationship	
Parametric test for correlation	**Assumptions:** 1) Pairs are independent (random selection of pairs) 2) (x_i, y_i) are distributed bivariate normal, where $r = \frac{1}{n-1}\sum_{i=1}^{k}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$ Let ρ denote the population correlation between numeric variables X and Y. We measure n pairs of data, (x_i, y_i). **Step 1: State the null and alternative hypothesis** $H_0: \rho = 0 \;\; v.s. \;\; H_1: \rho \neq 0$ $H_0: \rho \geq 0 \;\; v.s. \;\; H_1: \rho < 0$ $H_0: \rho \leq 0 \;\; v.s. \;\; H_1: \rho > 0$ **Step 2: Calculate the test statistic** $t_s = r\sqrt{\frac{n-2}{1-r^2}}$ **Step 3: Calculate the p-value** $H_1: \rho \neq 0 \;\; => \;\; p-value = 2P(t > \|t_s\|)$ $H_1: \rho < 0 \;\; => \;\; p-value = P(t < t_s)$ $H_1: \rho > 0 \;\; => \;\; p-value = P(t > t_s)$ **Step 4: If p-value < α, reject H_0.** **We can also create linear regression line and a test for the slope:** True model: $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ Least square line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1\bar{x}$, $\hat{\beta}_1 = r\left(\frac{s_y}{s_x}\right)$, $\beta_1 = \bar{y} - \hat{\beta}_1\bar{x}$ **Assumptions:** 1) Pairs are randomly sampled/independent 2) $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ independent and identically distributed

	Step 1: State the null and alternative hypothesis $H_0: \beta_1 = 0 \;\; v.s. \;\; H_1: \beta_1 \neq 0$ $H_0: \beta_1 \geq 0 \;\; v.s. \;\; H_1: \beta_1 < 0$ $H_0: \beta_1 \leq 0 \;\; v.s. \;\; H_1: \beta_1 > 0$ **Step 2: Calculate the test statistic** $t_s = \hat{\beta}_1 \sqrt{\frac{\sum(x_i - \bar{x})^2}{MSE}} \sim t(df = n-2)$, where $MSE = \frac{\Sigma(y_i - \hat{y}_i)^2}{n-2}$ **Step 3: Calculate the p-value** $H_1: \beta_1 \neq 0 \;\; => \;\; p - value = 2P(t > \|t_s\|)$ $H_1: \beta_1 < 0 \;\; => \;\; p - value = P(t < t_s)$ $H_1: \beta_1 > 0 \;\; => \;\; p - value = P(t > t_s)$ **Step 4: If p-value < α, reject H_0.** # If β_1 or $\rho = 0$ => no linear relationship between X and Y ## If β_1 or $\rho < 0$ => negative linear relationship between X and Y ### If β_1 or $\rho > 0$ => positive linear relationship between X and Y
Permutation test for the slope	**Common reasons we use a non-parametric test:** 1) Outliers present (violates normality) 2) Non constant variance (violates normality) 3) Small sample size (may not be able to conclude normal) **Assumptions:** A random sample of pairs of data was taken. **Step 1: State the null and alternative hypothesis** $H_0: \beta_1 = 0 \;\; v.s. \;\; H_1: \beta_1 \neq 0$ $H_0: \beta_1 \geq 0 \;\; v.s. \;\; H_1: \beta_1 < 0$ $H_0: \beta_1 \leq 0 \;\; v.s. \;\; H_1: \beta_1 > 0$ **Step 2: Calculate the observed test hypothesis** $\hat{\beta}_1^{OBS}$ = estimated least-squares slope = $r\frac{S_y}{S_x}$ Step 3: Calculate the permutation p-value

<table>
<tr><td></td><td>There are n ways to pair the first y_i with an x_i, then n-1 ways to pair the second y_i with an x_i, etc. There are $n!$ possible permutations.
1) Permute the data and calculate $\hat{\beta}_1^i$
2) Repeat for either
i) All $n!$ permutations
ii) R > 3000 random permutations
3) The actual or estimated permutation p-values are
$H_1: \beta_1 \neq 0 \;\; => \;\; \frac{\# \, of \, |\hat{\beta}_1^i| \geq |\hat{\beta}_1^{OBS}|}{n!}$ (actual) or $\frac{\# \, of \, |\hat{\beta}_1^i| \geq |\hat{\beta}_1^{OBS}|}{R}$ (estimated)
$H_1: \beta_1 < 0 \;\; => \;\; \frac{\# \, of \, \hat{\beta}_1^i \leq \hat{\beta}_1^{OBS}}{n!}$ (actual) or $\frac{\# \, of \, \hat{\beta}_1^i \leq \hat{\beta}_1^{OBS}}{R}$ (estimated)
$H_1: \beta_1 > 0 \;\; => \;\; \frac{\# \, of \, \hat{\beta}_1^i \geq \hat{\beta}_1^{OBS}}{n!}$ (actual) or $\frac{\# \, of \, \hat{\beta}_1^i \geq \hat{\beta}_1^{OBS}}{R}$ (estimated)

Step 4: If p-value < α, reject H_0.</td></tr>
<tr><td>Large Sample approximation to permutation test for the slope</td><td>Assumptions:
A random sample of pairs of data was taken, combined sample size at least 30.

Step 1: State the null and alternative hypothesis
$H_0: \rho = 0 \;\; v.s. \;\; H_1: \rho \neq 0$
$H_0: \rho \geq 0 \;\; v.s. \;\; H_1: \rho < 0$
$H_0: \rho \leq 0 \;\; v.s. \;\; H_1: \rho > 0$

Step 2: Calculate the test statistic
$Z_s = \frac{r-0}{1/\sqrt{n-1}} = r\sqrt{n-1} \sim N(0,1/\sqrt{n-1})$

Step 3: Calculate the p-value
$H_1: \rho \neq 0 \;\; => \;\; 2P(Z > |Z_s|)$
$H_1: \rho < 0 \;\; => \;\; P(Z < Z_s)$
$H_1: \rho > 0 \;\; => \;\; P(Z > Z_s)$

Step 4: If p-value < α, reject H_0.</td></tr>
<tr><td>Spearman's Rank Correlation</td><td>Let $R(X_i)$ = rank for x_i, $i = 1, \dots, n$; $R(Y_i)$ = rank for y_i, $i = 1, \dots, n$.
$\bar{R}(x)$ = average rank of x_i, $S_{R(Y)}$ = standard deviation of rank of Y_i

Step 1: State the null and alternative hypothesis
$H_0: \rho_s = 0 \;\; v.s. \;\; H_1: \rho_s \neq 0$
$H_0: \rho_s \leq 0 \;\; v.s. \;\; H_1: \rho_s > 0$</td></tr>
</table>

	$H_0: \rho_s \geq 0 \;\; v.s. \;\; H_1: \rho_s < 0$ **Step 2: Calculate the test statistic** $r_s = \frac{1}{n-1} \sum_{i=1}^{k} \left(\frac{R(x_i) - \bar{R}(x)}{s_{R(x)}}\right)\left(\frac{R(y_i) - \bar{R}(y)}{s_{R(y)}}\right)$ **Step 3: Calculate the p-value** $H_1: \rho_s \neq 0 \;\; => \;\; 2P(r_s^* > \|r_s\|)$ $H_1: \rho_s > 0 \;\; => \;\; P(r_s^* > r_s)$ $H_1: \rho_s < 0 \;\; => \;\; P(r_s^* < r_s)$ # If $H_1: \rho_s < 0, P(r_s^* < -c) = P(r_s^* > c)$ **Step 4: If p-value < α, reject H_0.**
Kendall's Tau	Kendall's Tau does not use ranks directly, but also does not use the original data. Suppose we looks at a pair of $(x_i, \; y_i)$, say $(x_1, \; y_1)$ and $(x_2, \; y_2)$. 1) If as X increases, Y tends to increase, then we should see $x_1 > x_2 \;\; => \;\; y_1 > y_2$. 2) If as X increases, Y tends to decrease, then we should see $x_1 > x_2 \;\; => \;\; y_1 < y_2$. **We use this to describe "discordant" and "concordant" pairs.** Concordant pairs: If $X_i < X_j \;\; => \;\; Y_i < Y_j$, or equivalently $(X_i - X_j)(Y_i - Y_j) > 0$ (or $X_i > X_j \;\; => \;\; Y_i > Y_j$) Discordant pairs: If $X_i < X_j \;\; => \;\; Y_i > Y_j$, or equivalently $(X_i - X_j)(Y_i - Y_j) < 0$ (or $X_i < X_j \;\; => \;\; Y_i > Y_j$) If most pairs are concordant => positive linear relationship If most pairs are discordant => negative linear relationship **The "population" value of Kendall's Tau is** $\tau = 2P\left[(X_i - X_j)(Y_i - Y_j) > 0\right] - 1$, which is a rescaled probability of concordant pairs. If all pairs are concordant, $\tau = 1$. If all pairs are discordant, $\tau = -1$. If exactly half are concordant, half are discordant, $\tau = 0$. There are $\binom{n}{2}$ total pairs $(X_i, \; X_j), (Y_i, \; Y_j)$ then $U_{ij} = 1$ if $(X_i - X_j)(Y_i - Y_j) > 0$ (concordant) $U_{ij} = \frac{1}{2}$ if $(X_i - X_j)(Y_i - Y_j) = 0$ (tied) $U_{ij} = 0$ if $(X_i - X_j)(Y_i - Y_j) < 0$ (discordant)

	Let $V_i = \sum_{j=i+1}^{n} U_{ij} =$ # of concordant pairs for i^{th} value $(x_i,\ y_i)$. # $j = i + 1$ ensures that we are never comparing the same pair. $r_\tau = \frac{2[\sum_{i=1}^{n-1} V_i]}{\binom{n}{2}} - 1$
Exact Hypothesis Test for τ	**Step 1: State the null and alternative hypothesis** $H_0: \tau = 0 \;\; v.s. \;\; H_1: \tau \neq 0$ $H_0: \tau \leq 0 \;\; v.s. \;\; H_1: \tau > 0$ $H_0: \tau \geq 0 \;\; v.s. \;\; H_1: \tau < 0$ **Step 2: Calculate test statistic** $r_\tau = \frac{2[\sum_{i=1}^{n-1} V_i]}{\binom{n}{2}} - 1$ **Step 3: Calculate the p-value** $H_1: \tau \neq 0 \;\; => \;\; 2P(r_\tau^* > \|r_\tau\|)$ $H_1: \tau > 0 \;\; => \;\; P(r_\tau^* > r_\tau)$ $H_1: \tau < 0 \;\; => \;\; P(r_\tau^* < r_\tau)$ **Step 4: If p-value < α, reject H_0.**
Permutation test for τ	Same step 1 and 2 as exact hypothesis test for τ **Step 3: Calculate the p-value** $H_1: \tau \neq 0 \;\; => \;\; (\# \, r_\tau^* \geq \|r_{\tau OBS}\|)/R$ $H_1: \tau > 0 \;\; => \;\; (\# \, r_\tau^* \geq r_{\tau OBS})/R$ $H_1: \tau < 0 \;\; => \;\; (\# \, r_\tau^* \leq r_{\tau OBS})/R$ **Step 4: If p-value < α, reject H_0.**
Asymptotic Approximation for τ	The following formula can be used with or without ties. Let $s_i =$ # of ties for the i^{th} tied value of X Let $t_i =$ # of ties for the i^{th} tied value of Y
When to use which correlation	1) When there are no outliers, and the distribution is approximately symmetric (but with low sample size) use a permutation test for the slope. 2) When outliers are present in the data, use Spearman's of Kendall's.

	3) Kendall and Spearman tend to have similar results, but Spearman tends to have higher power at low sample sizes, and Kendall has higher power in large sample sizes.