VIII. Two-Way Non-Parametric Analysis of Variance:
A. $X_{kli}$, $k \in \{1, \ldots, K\}$, $l \in \{1, \ldots, L\}$; $i \in \{1, \ldots, M_{kl}\}$; $X_{kli} \sim F_{kl}$ and are independent. Generally $F_{kl} = F_l(\cdot - \theta_k)$
   1. Here $k$ indexes treatment
   2. Here $l$ indexes a blocking factor.
   3. Don’t add KW statistics per block to get $\chi^2$
      a. Calculate separate KW statistics for each block
      b. Since statistics from separate blocks are independent, add KW statistics to get $\chi^2_{(K-1)}$ degree of freedom statistic.
      c. Problem: Does not penalize different treatment orderings within block, and so has low power.
B. Key point: Under null, every ordering within block is equally likely, and alternative shifts treatments.
C. Do inference based on ranks within block.
   1. Rank within blocks: $R_{kli}$ is rank of $X_{kli}$ within $X_{1li}, \ldots, X_{KIM_{kl}}$.
   2. Sums of ranks across blocks are dependent.
      a. Ranks that make up each sum within block are independent.
      b. Using means makes the test remain sensible when design is unbalanced.
   3. In the balanced case, the test statistic
      
      $\begin{align*}
      R_{k..} &= \sum_{l=1}^{L} \sum_{i=1}^{M_{kl}} R_{kli} \\
      T &= 12L \sum_{k=1}^{K} \left[ \bar{R}_{k..} - \frac{1}{2} (MK + 1) \right]^2 / [K(MK + 1)]
      \end{align*}$
      
      $K - 1$ was replaced by $K$ in the denominator to account for dependence among group rank means.
   4. Number of replicates impacts complication of formula.
      a. Higgins does case with $M = 1$;
      b. Conover allows for balanced replication, indexed by $k$.
   5. Can also do this test via permutation, within block and across treatments.
   6. Can examine for where differences in means occur
      1. Can use the variances and covariances to show that under
         the null $\frac{(\bar{R}_{k..} - \bar{R}_{..})}{\sqrt{K(LK + 1)/6}}$ is approximately normal.
      2. Then use earlier multiple comparisons approaches to adjust.
         a. Bonferroni
         b. Fisher’s LSD
         c. Tukey’s HSD
F. You can use scoring methods as before
   1. Ex. normal, Savage, etc.
   2. Can use average ranks to adjust for ties
   3. Observed data values, exactly to give permutation distribution.
   4. Number of replicates impacts complication of formula.
      a. Higgins does case with $M = 1$;
      b. Conover allows for balanced replication, indexed by $k$.
H. Same techniques may be used for unbalanced designs as well, but calculations are more tedious. (Bernard and van Elteren (1953) *Indagationes Math.*): 4.6
I. Ordered Alternatives:
   1. Recall Friedman’s statistic
      $K \propto \sum_{k=1}^{K} [\bar{R}_{k..} - 1/2(MK + 1)]^2$
      a. $\bar{R}_{k..}$ are mean rankings for treatment $k$ when observations are ranked within blocks.
      b. Any deviation in any treatment in any direction is treated similarly.
   2. If one wants to look specifically for ordered alternatives, use instead $P = \sum_{k=1}^{K} (k - (K + 1)/2) \bar{R}_{k..}$.
      a. Book uses rank sums rather than rank means, and is specific to the non-replicated case.
      b. Using means makes the test remain sensible when design is unbalanced.
      c. Test is called Page’s test (for E.B. Page, not Jimmy Page).
   3. Null expectation is $\sum_{k=1}^{K} (k - (K + 1)/2) E [\bar{R}_{k..}]$
   4. Null variance is $\sum_{k=1}^{K} \sum_{m=1}^{M} (k - (K + 1)/2)(m - (K + 1)/2) Cov [\bar{R}_{k..}, \bar{R}_{m..}]$.
      : 5.1-5.3
   5. Bivariate Methods: $(X_i, Y_i) \sim f_{XY}(x, y)$; determine whether there is association between these variables.
A. Test whether \( f_{X,Y}(x,y) = f_X(x)f_Y(y) \).

1. Want test not to require knowledge of \( f_{X,Y} \), or even null values.
2. Against alternative that, vaguely, large values of \( X \) are associated with large values of \( Y \) (or vice versa).

B. Also, measure strength of association.

C. Parametric Approach: Pearson Correlation

1. \( r_p = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}} \).

2. This gives the slope of the least squares line fitting \( Y \) to \( X \), after scaling both variables by standard deviation.

3. Cauchy-Schwartz says that this is always in \([-1,1]\).
   a. Perfect positive or negative linear association is reflected in 1 or -1 resp.

4. Inference:
   a. \((X_j,Y_j)\) independent bivariate normal, unit variance, \( j = 1, n \).
   b. Numerator of \( r_p^2 \) is multiple of \( \chi^2 \).
   c. Denominator of \( r_p^2 \) is multiple of \( \chi^2_{n-2} \).
   d. Numerator square is one of squares in denominator
   e. So \( t = \sqrt{n - 2}r_s/\sqrt{1 - r_s^2} \sim t_{n-2} \), even when \( p \neq 0 \).

D. Variance of \( r_p \):

1. Fix the \( X \)'s, and rearrange the \( Y \)'s.
2. For permutation distribution of \( Y \)'s, note \( \text{Cov}(Y_1,Y_2) = \sum_{i\neq j} \frac{1}{n(n-1)}(y_i - \bar{y})(y_j - \bar{y}) \)
3. \( = \sum_{i,j} \frac{1}{n(n-1)}(y_i - \bar{y})(y_j - \bar{y}) - \frac{1}{n(n-1)}(y_i - \bar{y})^2 \)
4. \( = -\frac{1}{n-1}\sigma_Y^2 \),
   for \( \sigma_Y^2 = \sum \frac{1}{n(n-1)}(y_i - \bar{y})^2 \).

3. So \( \text{Var}\left[\sum (X_j - \bar{X})Y_j\right] = \sum (X_j - \bar{X})^2 \text{Var}[Y_j] \)
   + \( \sum (X_i - \bar{X})(X_j - \bar{X})\text{Cov}[Y_j,Y_i] \)
   = \( \sum (X_j - \bar{X})^2\sigma_Y^2 \)
   - \( \left[\sum_j(X_i - \bar{X})(X_j - \bar{X}) - \sum_i(X_i - \bar{X})^2\right]\sigma_Y^2 \)
   = \( \sum (X_j - \bar{X})^2\sigma_Y^2 + \sum_i(X_i - \bar{X})^2\sigma_Y^2 / (n-1) \)
   = \( n^2\sigma^2_Y\sigma_Y^2 / (n-1) \)

4. So \( \text{Var}[r_p] = 1/(n-1), \) under permutation distribution.

E. Look at rank correlation (Spearman, 1904.)

1. Under the null each \( X \) rank should be equally likely to be associated with each \( Y \) rank.

2. Under the alternative, extreme ranks should be associated with each other.

3. Let \( R_j \) be the rank of the \( Y \) value associated with \( X(j)_i \).

4. Define the Spearman Rank correlation as
   \[ r_s = \frac{\sum_j (j - (n+1)/2)(R_j - (n+1)/2)}{\sqrt{\left(\sum_j (j - (n+1)/2)^2\right)\left(\sum_j (R_j - (n+1)/2)^2\right)}} \]

5. That is, \( r_s \) is Pearson correlation on ranks.

6. Sums of squares in the denominator have the same value for every data set, and numerator can be simplified:
   a. \( \sum_j (j - (n+1)/2)^2 = \sum j^2 - n(n+1)^2 / 4 \)
   \( \quad = n(n+1)(2n+1)/6 - n(n+1)^2/4 \)
   \( \quad = n(n^2 - 1) / 12. \)
   b. \( \sum_j (R_j - (n+1)/2)^2 \) is same.
   c. \( \sum_j (j - (n+1)/2)(R_j - (n+1)/2) = 0. \)
   d. \( r_s = (12/n(n^2 - 1))\sum_j (j - (n+1)/2)R_j \)

7. Values for perfect agreement:
   a. \( 1 \) if variables share exact same ordering, and
   b. \( -1 \) if ordering is exactly opposite.
   c. Cauchy-Schwartz says these are extremes.

8. Under \( H_0 \), \( E[r_s] = 0. \)
   a. expected value of each numerator summand is 0

9. Can examine both this and Pearson correlation under permutation

F. Construct a new measure based on counts of concordant and discordant pairs

1. Consider the population quantity
   \[ \tau = 2P\left[ (X_i - X_j)(Y_i - Y_j) > 0 \right] - 1, \]
   a. called Kendall’s Tau.
   b. Reflects the probability of a concordant pair minus the probability of a discordant pair.

2. Denote the number of concordant pairs by
   \[ Q = \sum_{i<j}s(R_i - R_j) = \sum_{i<j}s((X_j - X_i)(Y_j - Y_i)). \]
   \( Q^\dagger = n(n-1)/2 - Q \) is number of discordant pairs; equals number of rearrangements necessary to make all pairs concordant.
   a. \( s(u) = 1 \) if \( u > 0 \)
   b. \( 0 \) otherwise.
   b. Number of discordant pairs is \( n(n-1)/2 - Q^\dagger \).
   c. Estimate \( \tau \) by the excess of concordant over discordant pairs, divided by maximum:
   \[ r_\tau = \frac{(n(n-1)/2 - Q^\dagger) - Q^\dagger}{n(n-1)/2} = 1 - \frac{4Q^\dagger}{n(n-1)} \]

3. Then \( E[Q^\dagger] = 1/2n(n-1)p_1 \), for \( p_1 = P\left[ (X_1 - X_2)(Y_2 - Y_1) > 0 \right], \)
4. \( E[r_\tau] = 1 - 2p_1 \)
5. Null value of \( p_1 \) is half.
6. \( \text{Var} [Q^\dagger] = \sum_{i<j} \text{Var} [Z_{ij}] + \sum^* \text{Cov} [Z_{ij}, Z_{kl}] \)

a. \( \sum^* \) the sum over \( i < j \), \( k < l \), 3 distinct indices.

i. Consists of \( 1/4n^2(n-1)^2 - 1/2n(n-1) - 1/4n(n-1)(n-2)(n-3) = 1/4n(n-1)(n^2-n^2+5n-6-2) = n(n-1)(n-2) \) terms.

b. \( Z_{ij} = s((X_i - X_j)(Y_j - Y_i)) \).

c. \( \text{Var} [Q^\dagger] = \sum_{i<j} \text{Var} [Z_{ij}] + \sum^* \text{Cov} [Z_{12}, Z_{13}] \)

\[ = \sum_{i<j} (p_1 - p_2^2) + \sum^* (p_3 - p_2^2) \]

\[ = 1/2n(n-1)(p_1 - p_2^2) + n(n-1)(n-2)(p_3 - p_2^2) \]

i. \( p_3 = P [(X_1 - X_2)(Y_2 - Y_1) \geq 0, (X_1 - X_3)(Y_3 - Y_1) \geq 0] \)

d. Null value of \( p_3 \) is \( \frac{5}{18} \).

i. Examine all 36 combinations of two sets of permutations of 1, 2, 3.

e. \( \text{Var}_0 \ [Q^\dagger] = 1/2n(n-1) + n(n-1)(n-2) \frac{5}{18} = n(n-1)(n^2 + n^2 - n^2) = n(n-1)(5 + 2n)/72 \).

f. \( \text{Var}_0 \ [r_2] \) is \( 2(2n+5)/(9n(n-1)) \).