V. Nonparametric One Way Analysis of Variance:

A. Definition of Problem:
1. Suppose $X_{ki}$ are samples from $K$ potentially different populations.
   a. $k$ indexes group, $k \in \{1, \ldots, K\}$.
   b. each sample including $n_k$ observations, $i \in \{1, \ldots, n_k\}$.
2. We wish to determine whether all populations are the same.
   a. $H_0 : F_1(x) = \cdots = F_K(x) \forall x$ vs
   b. $H_1 : F_i(x) \neq F_j(x) \forall x$ for some indices $i, j$, with strict equality at some $x$.
   b. Specifically, look for location differences
      i. assuming equal distributions up to a shift.
      ii. Let $\bar{R}_{ki}$ be the rank of $X_{ki}$ within the combined sample.
      iii. Then null hypothesis $E[\bar{R}_{ki}] = (N + 1)/2$ and
           $\text{Var}[\bar{R}_{ki}] = (N - n_k)(N + 1)/(12n_k)$.

3.1. ANOVA: heavily reliant on distribution of responses

3.2. Reference distribution is generated from permutation of ranks among groups, or groups among ranks. [Mark B sas] [Mark B R]

3.3. Contrasts: Tests for differences within set.
   1. We have $K(K - 1)/2$ chances to find a significant result.
   2. If done at nominal level, this will inflate experiment-wise error rate.
3. If nominal level is multiplied by number of possible comparisons performed ($\frac{1}{2}K(K - 1)$).
   a. will get usually a very conservative procedure.
   b. This is the Bonferroni procedure.
4. Fisher’s Least Significant Difference (LSD) method:
   a. Do Kruskal-Wallis test
   b. If Kruskal-Wallis test rejects $H_0$, then do test on each pairwise comparison between mean ranks
      i. Standard error for $\bar{R}_j - \bar{R}_i$ is
      $\sqrt{\frac{1}{12}N(N + 1)(1/n_j + 1/n_i)}$.
      ii. This is not the same as the Wilcoxon test for groups $i$ and $j$.
   c. This fails to control Type I error rate (aka size) if $K > 3$.
      i. Suppose $F_1(x) = F_2(x) = \cdots = F_{K-1}(x) = F_K(x - \Delta)$ for $\Delta \neq 0$.
      ii. Then null hypotheses $F_j(x) = F_i(x)$ are true, for $i, j < K$.
4. Average ranks for different groups are dependent.
   a. By extension to above,
   $\text{Var}(n_k\bar{R}_k + n_i\bar{R}_i) = (n_k + n_i)(N - n_j - n_i)(N + 1)/12$.
   b. Let $R_{ki} = n_k\bar{R}_k$ be total of ranks, rather than average of ranks. Then
   $\text{Cov}(R_{ki}, R_{lj}) = \frac{1}{2} [(\text{Var}(R_{ki} + R_{lj}) - \text{Var}(R_{ki}) - \text{Var}(R_{lj})]
   = [n_k(n_l + n_j) - n_k(n - n_i) - n_l(n - n_j) - n_l(n - n_i)](N + 1)/24
   = -[n_kn_l](N + 1)/12$.
   c. Approximate $H_0$ on the union of the critical regions for the $K$ tests.
   i. $F_j = \frac{\text{Var}(R_{ki} + R_{lj}) - \text{Var}(R_{ki})}{\text{Var}(R_{lj})}$.
   ii. Then null hypotheses $\bar{R}_j = \bar{R}_i$.
   iii. Can make $\Delta$ so large that Kruskal-Wallis test rejects equality of all distributions with probability close to 1.
   iv. Then at least one of the true null hypotheses is rejected on the union of the critical regions for the separate $(K - 1)(K - 2)/2$ tests.
   v. If $K = 3$, there is only one such test, and so no problem with multiple comparisons.

5. Tukey’s Honest Significant Difference (HSD) method:
   a. Studentized range distribution:
      i. Setup: $X_{ij} \sim N(\mu, \sigma^2/n_j)$ for $j \in \{1, \ldots, K\}$,
         $S^2 \sim \sigma^2X^2_{ni}$ independent, $n_j$ all equal.
      ii. If all $n_j$ are equal,
         $\text{max}_{1 \leq i < j < K}((X_{ij} - X_{ik})/(S/\sqrt{n_j}))$ has the studentized range distribution with $K$ and $m$ degrees of freedom, exactly.
      iii. If $n_j$ are not all equal,
         $\text{max}_{1 \leq i < j < K}((X_{ij} - X_{ik})/(S\sqrt{(1/2)(1/n_j + 1/n_k)}))$ has the studentized range distribution with $K$ and $m$ degrees of freedom, approximately.
         - It’s curious that this distribution wasn’t defined without the 1/2 under the square root, but it wasn’t.

   b. Approximate $\bar{R}_1, \ldots, \bar{R}_K$ as means of $n_1, \ldots, n_K$ normally distributed variables each with variance
      $(n_i(N - n_i)(N + 1)/12)/n_i^2 = (N - n_i)(N + 1)/(12n_i)$
      c. Approximate $\bar{R}_1, \ldots, \bar{R}_K$ as means of $n_1, \ldots, n_K$ independently normally distributed variables each with variance
      $(n_i(N - n_i)(N + 1)/12)/n_i^2 = (N - n_i)(N + 1)/(12n_i)$.
VI. Ordered Alternatives: Test $\theta_1 = \cdots = \theta_K$ vs $\theta_1 \leq \cdots \leq \theta_K$ with strict equality somewhere.

A. For $F_i(x) = F(x - \theta_i)$
B. Reduces parameter space to $\frac{1}{2}K$ of former size.
C. Use the Jonckheere-Terpstra Test: $J = \sum_{i<j} U_{ij}$

1. where $U_{ij}$ is MWW test for subsamples $i, j$.
2. Reject when $J$ is large.
D. This can be calibrated using asymptotic normality.
1. Expectation is $\sum_{i<j} n_i n_j / 2 = N^2 / 4 - \sum_i n_i^2 / 4$.
2. $\text{Var} [J] = \frac{1}{12} \sum_{i=2}^K \text{Var}[U_i] = \frac{1}{12} \sum_{i=2}^K n_i m_{i-1} (m_i + 1)$.
   a. $m_i = \sum_{j=1}^i n_j$.
   b. Hettmansperger has derivation, but you won’t learn much from reviewing it.

: 4.1-4.3

VII. Location tests for symmetric distributions and Definition of the Wilcoxon Signed–Rank Statistic:
A. Example: Paired data Testing
1. $H_0$: Two populations have the same continuous distribution versus (and differences are symmetric)
2. $H_A$: Two populations have same continuous distribution, shifted.
3. Paired samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$.
B. Suppose that $Z_i = X_i - Y_i$, which is often close to symmetrically distributed.
C. To derive a test under the assumption of symmetry, use definitions:
1. Let $R_j = \text{number of } Z_i \ni \{ 1 \text{ if item whose abs. value is ranked } j \text{ is } + \}
   \text{ and } 0 \text{ if item whose abs. value is ranked } j \text{ is } -$
2. Let $V_j = \begin{cases} 1 & \text{ if item whose abs. value is ranked } j \text{ is } + \\ 0 & \text{ if item whose abs. value is ranked } j \text{ is } - \end{cases}$
3. Let $S_j = \begin{cases} 1 & \text{ if item } j \text{ is } + \\ 0 & \text{ if item } j \text{ is } - \end{cases}$
4. Wilcoxon Signed rank test is $T = \sum j V_j = \sum R_j S_j$

Lecture 6

D. Under $H_0: \theta = 0$, $(S_j)$ and $(R_j)$ are independent random vectors.
1. Follows from the fact that $S_j$ and $|X_j|$ are pairwise independent under $H_0$.
2. $T = \sum j V_j$
3. $\mathbb{E}_0 [T] = \sum j \mathbb{E}_0 [V_j] = n(n+1)/4$
4. $\text{Var}_0 [T] = \sum j^2 \text{Var}_0 [V_j] = \sum j^2 / 4 = n(2n+1)(n+1)/24$.