6. Nonparametric approach: Rank sum statistic
   a. Assume data are continuous.
   b. Rank all of the observations
      i. Under null hypothesis, all orderings have equal probabilities
   c. Compute \( W = \) sum of ranks in one of the groups (M, perhaps)
   d. Calculate expectation and variance
      i. \( W = \sum_{j=1}^{m+n} I_{j}j \) for \( I_{j} = \begin{cases} 1 & \text{if subject ranked } j \text{ is from } G; \\ 0 & \text{otherwise}. \end{cases} \)
      ii. \( \mathbb{E}[I_{j}] = n/(m+n) \)
      iii. \( \mathbb{E}[W] = \frac{n}{m+n} \sum_{j=1}^{m+n} j = \frac{n(m+n)(m+n+1)}{2(m+n)} = \frac{n}{m+n} \sum_{j=1}^{m+n} j \)
   iv. Variance is harder, since the \( I_{j} \) are not independent.
   e. If \( V \) is the sum of ranks for the other group, then \( V + W = (n + m)(n + m + 1)/2 \)
   f. Compare against normal.
      i. Proving this is difficult, since the addends are neither identically distributed nor independent.
         : 2.6-2.6.2
   g. Alternative formulation: \( W = \) sum of ranks of \( Y \)'s among whole sample
      \[ = \sum_{y} \#(\text{data points less than or equal to } y) = \sum_{y} \#(X \text{ values less than or equal to } y) + \sum_{y} \#(Y \text{ values less than or equal to } y) = U + n(n + 1)/2 \quad \text{for } U = \sum_{i=1}^{m} \sum_{j=1}^{n} I(X_{i} < Y_{j}) \]
      i. \( U \) is called Mann-Whitney Statistic.

\[ \text{Lecture 3} \quad 23 \quad \text{Lecture 3} \quad 24 \]

1. So
2. \[ \text{Var}[W] = \sum_{i=1}^{m+n} \sum_{j=1}^{m+n} a_{i}a_{j}\text{Cov}[I_{i}, I_{j}] \]
3. \[ = \sum_{i=1}^{m+n} a_{i}^{2}b_{1} + \sum_{i \neq j} a_{i}a_{j}b_{2} \]
4. \[ = (b_{1} - b_{2})(m+n)s_{2} + (n+m)^{2}b_{2}a^{2} \]
5. \[ = mn \frac{(m+n-1)}{(s_{2} - \bar{a}^{2})} \quad \text{(3)} \]
   for \( s_{2} = \sum_{i=1}^{m+n} a_{i}^{2}/(m+n) \).
   j. When \( a_{j} = j \), need \( \sum_{j=1}^{m+n} a_{j}^{2} \).
   i. Guess it must be cubic.
   ii. Look for function \( g(w) = aw^{3} + bw^{2} + cw + d \) so that \( g(0) = 0 \) and \( g(w) - g(w - 1) = w^{2} \).
   iii. Then \( d = 0 \), and \( w^{2} = (aw^{3} + bw^{2} + cw) - \)
   \[ = aw^{3} + 3aw^{2} - 3aw + a - bw^{2} + 2bw - b - cw + c \]
   \[ = 3aw^{2} - 3aw + a + 2bw - b + c. \]
   iv. So \( a = 1/3, \) \( b = 1/2, \) and \( c = 1/6 \).
   v. Then \( g(w) = w(2w + 1)(w + 1)/6. \)
   vi. Then \( \text{Var}[W] = mn(m+n+1)/12. \)

k. \( \text{P}_{m,n}[U = u] = f(u, m, n)/(\binom{m+n}{n}) \) for \( f(u, m, n) \) the number of ways that \( m \) symbols \( x \) and \( n \) symbols \( y \) can be written in a list to give \( U = u \).
   i. Number of these lists can be divided according to whether last symbol is \( x \) or \( y \).
   ii. So \( f(u, m, n) = f(u, m-1, n) + f(u - m, m, n - 1). \)
   iii. Can do recursively.
   : 2.7

7. Alternative test: let \( a_{j} \) be other scores.
   a. Under the null hypothesis, these scores with \( m \) assigned to \( x \) group and \( n \) assigned to \( y \) group, with all rearrangements having equal probability.
   i. Statistic still given by (1).
   ii. Moments still given by (2) and (3).
   iii. Results only minimally different from ANOVA on scores.
   c. Or exactly: List all possible ways to divide the \( m + n \) objects into two groups, one of size \( m \)
   i. There are \( \binom{m+n}{m} \) of these.
   ii. Calculate the test statistic for each rearrangement.
   iii. Calculate the number of rearrangements giving a test statistic as extreme or more extreme than the one we observe.
   iv. Divide by \( \binom{m+n}{m} \) to get \( p \)-value.
   d. Could use
   i. scores equal to expected value of order statistics from normal distribution: Normal scores.
   ii. scores calculated from the normal quantile function \( a_{j} = \Phi^{-1}(j/(n+m+1)) \): Van der Waarden scores.
   iii. scores equal to expected value of order statistics from exponential distribution: Exponential scores.
e. Calculating exact probabilities for score tests is hard.
   i. Unless scores are integers, you will have \( \binom{m+n}{n} \) possible values for the test statistic
   ii. Sorting through these to figure how many will be in critical region is very hard, even by computer.
   iii. You won’t find a simplifying formula.

f. Scores may be chosen to be optimal against certain distributions:
   i. Normal scores are optimal against normal observations.
   ii. Exponential scores are optimal against exponential distribution.
   iii. Original ranks are optimal against logistic

8. Data-Dependent Scoring
   a. Ties:
      i. Assign tied observations average ranks.
      ii. Treat ranks as general scores as above.
      iii. Use normal approximation based on variance
         • Variance is different from before.
         • You can do the summation in closed form (formula given in book), but you won’t learn much from this.
         • Tie correction should be done even with tied observations fall in the same group.
      iv. In this case, the exact analysis is impossible, since we can’t handle the additional randomness arising from potential ties in values.

   b. Instead of scores, use original data
      i. Then test statistic is the difference in means that you would ordinarily use from 401,
      ii. but sampling distribution is generated by rearranging the group members.
      c. Unlike previous analyses, distribution of test statistic depends on observed data
         i. Different pattern of ties would result in different scores.

   H: p. 165

B. Efficiency
1. We can do this for test statistics \( T \) such that \((T - \mu(\theta))/(\sigma(\theta)/\sqrt{\text{sample size}}) \approx N(0,1)\).
2. In order to do ARE computations, we need to specify how \( m \) and \( n \) move together.
   a. Let \( m = \lambda N, n = (1 - \lambda)N \), for \( \lambda \in (0,1) \).
3. Applied pooled \( T \) test, MWW
   a. Pooled \( T \) test: approximately based on sample mean.
      i. Large sample version behaves as though variance of the data are known; we’ll make this assumption.
      ii. \( \mu(\theta) = \theta \)
      iii. \( \Var[T] = \rho^2 \left( \frac{1}{n} + \frac{1}{m} \right) = \rho^2 \left( \frac{1}{N(1 - \lambda)} + \frac{1}{N\lambda} \right) \)
      and \( \sigma(\theta) = \rho \sqrt{\frac{1}{N(1 - \lambda)} + \frac{1}{N\lambda}} = \frac{\rho}{\sqrt{N(1 - \lambda)}} \), for \( \rho^2 \)
      the variance of each observation.
   iv. Normal, unit variance: \( e = \sqrt{\lambda(1 - \lambda)} \).

<table>
<thead>
<tr>
<th>Test</th>
<th>Pooled ( T )</th>
<th>MWW</th>
<th>ARE</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>( \mu(\theta) = \theta )</td>
<td>( \mu(\theta) = \Phi(\theta/\sqrt{2}) )</td>
<td>( \frac{\rho}{\sqrt{2}} )</td>
</tr>
<tr>
<td>( \sigma(\theta) = \sqrt{\Var[X_i]} \times )</td>
<td>( \sigma(\theta) = \sqrt{\lambda(1 - \lambda)} )</td>
<td>( 3 )</td>
<td>( \frac{3}{\eta} )</td>
</tr>
<tr>
<td>Normal</td>
<td>( \mu'(0) = 1 )</td>
<td>( \mu'(0) = (2\sqrt{\pi})^{-1} )</td>
<td>( \frac{3}{\eta} )</td>
</tr>
<tr>
<td>( \sigma(0) = \sqrt{\lambda(1 - \lambda)} )</td>
<td>( \sigma(0) = \sqrt{\lambda(1 - \lambda)} )</td>
<td>( \frac{\pi^2}{9} )</td>
<td>( \frac{10}{9} )</td>
</tr>
<tr>
<td>( e = (\sqrt{3}/\pi)\zeta )</td>
<td>( e = (\sqrt{3}/\pi)\zeta )</td>
<td>( \zeta )</td>
<td>( \zeta )</td>
</tr>
</tbody>
</table>

\( \zeta = (\lambda(1 - \lambda))^{1/2} \)

1. If one distribution is more spread out than another, then outside points are from one group and inside points are from the other.
2. Siegel-Tukey Test:
   a. Rank points: Min, then max, then second to max, then second to min, third to min, third to max, and continue alternating.
   b. Do Wilcoxon test.
3. Ansari-Bradley Test (I like this one better):
   a. Number from outside in, with extremes getting equal rank.
   b. Disadvantage Re Siegel-Tukey: Can’t use off-the-shelf Wilcoxon tail calculations
c. Advantage Re Siegel-Tukey: exactly invariant to reflection.