TWO-SAMPLE VALIDATION

**Background:** given data sets \( X_1, \ldots, X_n \), and \( Y_1, \ldots, Y_m \), a common validation problem is to test the hypothesis \( H_0 \) that the \( n + m \) RVs are iid: both data sets are from same \( F \).

Example Applications:
simulated vs. real data, control vs. experimental treatments.

**Two-Sample Rank Sum Test (Wilcoxon, Mann-Whitney)**

- Two-sample rank sum test procedure:
  a) sort the combined set \( X_1, \ldots, X_n, Y_1, \ldots, Y_m \);  
  b) define \( R_i \) to be rank of \( X_i \) in sorted set, with average rank used for ties;  
  c) compute the statistic \( r = \sum_{i=1}^{n} R_i \), rejecting \( H_0 \) for large \( r \) or small \( r \)
    (Note: average rank is \( \frac{n+m+1}{2} \), so \( E[R] = \frac{n(n+m+1)}{2} \));  
  d) compute \( p \)-value
    \[
    p = 2 \times \min \left( P_{H_0}(R \leq r), \ P_{H_0}(R \geq r) \right).
    \]
    Note: \( p \) is the probability that \( H_0 \) is correct; if \( p \) is small the hypothesis \( H_0 \) is rejected.

- Computation of \( p \): define \( P_{n,m}(r) = P_{H_0}(R \leq r) \) and use
  \[
  P_{n,m}(r) = \frac{n}{n+m} P_{n-1,m}(r-n-m) + \frac{m}{n+m} P_{n,m-1}(r),
  \]
  starting with
  \[
  P_{1,0}(k) = 0, \ k \leq 0, \ P_{1,0}(k) = 1, \ k > 0, \\
  P_{0,1}(k) = 0, \ k < 0, \ P_{0,1}(k) = 1, \ k \geq 0.
  \]
  Then \( P_{H_0}(R \geq r) = 1 - P_{n,m}(r-1) \).
  But recursive method is infeasible for large \( n, m \).
• The Normal approximation for large $n, m$: using

$$E_{H_0}[R] = \frac{n(n + m + 1)}{2},$$

$$V_{H_0}(R) = \frac{nm(n + m + 1)}{12};$$

a standardized normal is

$$Z = \frac{R - \frac{n(n + m + 1)}{2}}{\sqrt{\frac{nm(n + m + 1)}{12}}};$$

so if $r^* = \frac{r - \frac{n(n + m + 1)}{2}}{\sqrt{\frac{nm(n + m + 1)}{12}}}$, the approximate $p-$value is

$$p \approx \begin{cases} 2P\{Z < r^*\} & \text{if } r \leq \frac{n(n + m + 1)}{2} \\ 2P\{Z > r^*\} & \text{otherwise} \end{cases}.$$
TWO-SAMPLE TEST CONT.

- Two-Rank Sum Test Examples:

```matlab
X = [ 132 104 162 171 129 ]; n = 5; m = 10;
Y = [ 107 94 136 99 114 122 108 130 106 88 ];
XY = sort([ X Y ]); nm = n + m; r = 0;
for i = 1:n, r = r + find(X(i)==XY); end, disp(r) 55

% Simulation to Approximate p
for k = 1:1000, S = [1:nm]; rr = 0;
    % Determine rr for random subset
    for i = nm : -1 : nm-n+1, j = ceil(i*rand);
        rr = rr + S(j); S(j) = S(i);
    end, M(k) = rr <= r; N(k) = rr >= r;
end
disp( 2*min( mean(M), mean(N) ) ) % Simulated 0.0772
rs = (r-n*(nm+1)/2)/sqrt(n*m*(nm+1)/12); disp(rs) 1.8371
p = 2*(1 - normcdf(rs)); disp(p) % Normal approx. 0.066193
p = ranksum(X,Y); disp(p) % Exact 0.075258
```
TWO-SAMPLE TEST CONT.

X = [42 164 44 64 28 10 237 141 92 96]; n = 10;
Y = [16 48 104 43 63 297 167 61 206 172 8], m=11;
XY = sort([ X Y ]); nm = n + m; r = 0;
for i = 1:n, r = r + find(X(i)==XY); end, disp(r)
105

% Simulation to Approximate p
for k = 1:1000, S = [1:nm]; rr = 0;
    % Determine rr for random subset
    for i = nm : -1 : nm-n+1, j = ceil(i*rand);
        rr = rr + S(j); S(j) = S(i);
    end, M(k) = rr <= r; N(k) = rr >= r;
end
disp( 2*min( mean(M), mean(N) ) ) % Simulated
0.778
rs = (r-n*(nm+1)/2)/sqrt(n*m*(nm+1)/12); disp(rs)
-0.3521
p = 2*normcdf(rs); disp(p) % Normal approx.
0.7248
p = ranksum(X,Y); disp(p) % Exact
0.7513
cdfplot(X), hold on, cdfplot(Y)
Empirical CDF's for X and Y