

The mean and variance of the number of runs

Under the null hypothesis H_0 of the runs test, we have X_1, \dots, X_m and Y_1, \dots, Y_n , where all $m + n$ variables are i.i.d. with the same continuous distribution. Let R be the number of runs. Hogg and Tanis, 2d ed., p. 550, say that the following can be shown:

Proposition 1. [] (a) Under H_0 , the expectation

$$E_0 R = \frac{2mn}{m+n} + 1.$$

(b) Under H_0 , the variance

$$\text{Var}_0(R) = \frac{2mn(2mn - m - n)}{(m+n)^2(m+n-1)}.$$

Proof of (a). Let $R_j = 1$ if a run starts at the j th position among the $m + n$ order statistics of the combined sample, otherwise $R_j = 0$. Then $R = \sum_{j=1}^{m+n} R_j$. We have $R_1 \equiv 1$, so $E_0 R_1 = 1$. For $j = 2, \dots, m+n$, $E_0 R_j$ is the probability of having a Y in the $(j-1)$ st position and an X in the j th, or vice versa. Thus

$$E_0 R_j = 2 \binom{m+n-2}{m-1} / \binom{m+n}{n} = 2mn / [(m+n)(m+n-1)]$$

and

$$E_0 R = \sum_{j=1}^{m+n} E_0 R_j = 1 + (m+n-1)E_0 R_2 = 1 + \frac{2mn}{m+n},$$

proving (a).

For (b) we have $E_0(R^2) = \sum_{j=1}^{m+n} E_0(R_j^2) + 2 \sum_{1 \leq i < j \leq m+n} E_0(R_i R_j)$. For each j , we have $E_0(R_j^2) = E_0(R_j)$ as found in part (a). We also have $E_0(R_1 R_j) = E_0(R_j)$ for any $j \geq 2$.

For $3 \leq j \leq m+n$ we have by Pascal's identity

$$\begin{aligned} E_0(R_{j-1} R_j) &= \left[\binom{m+n-3}{m-1} + \binom{m+n-3}{m-2} \right] / \binom{m+n}{m} \\ &= \binom{m+n-2}{m-1} / \binom{m+n}{m} = \frac{mn}{(m+n)(m+n-1)}. \end{aligned}$$

For $2 \leq i < i+2 \leq j \leq m+n$ we have

$$E_0(R_i R_j) = 4 \binom{m+n-4}{m-2} / \binom{m+n}{m} = \frac{4m(m-1)n(n-1)}{\prod_{k=0}^3 (m+n-k)}.$$

Combining terms and multiplying by appropriate coefficients gives

$$\begin{aligned} E_0 R^2 &= 1 + \frac{2mn}{m+n} + \frac{4mn}{m+n} + \frac{2mn(m+n-2)}{(m+n)(m+n-1)} + \\ &\quad \left[\binom{m+n-1}{2} - (m+n-2) \right] \frac{8m(m-1)n(n-1)}{\prod_{k=0}^3 (m+n-k)} \\ &= 1 + \frac{8mn}{m+n} - \frac{2mn}{(m+n)(m+n-1)} + \frac{4m(m-1)n(n-1)}{(m+n)(m+n-1)} \\ &= 1 + \frac{8mn}{m+n} + \frac{4mn(mn-m-n) + 2mn}{(m+n)(m+n-1)}. \end{aligned}$$

Thus

$$\begin{aligned} \text{Var}_0 R &= E_0(R^2) - (E_0 R)^2 = \frac{4mn}{m+n} + \\ &\quad + \frac{4mn(mn-m-n) + 2mn}{(m+n)(m+n-1)} - \frac{4m^2 n^2}{(m+n)^2} \end{aligned}$$

which equals $mn/[(m+n)^2(m+n-1)]$ times

$$\begin{aligned} &4[m^2 + 2mn + n^2 - m - n] + 4[m^2 n + mn^2 - m^2 - 2mn - n^2] + \\ &\quad + 2m + 2n - 4[m^2 n + mn^2 - mn] = 4mn - 2m - 2n, \end{aligned}$$

and the expression for $\text{Var}_0 R$ in (b) follows, Q.E.D.