

Lecture 48: Mean and variance of the number of runs; the one-sample runs test

Statistical Inference–II ■ Units IV–V: Nonparametric Inference

Kapil Kumar

Department of Statistics
Chaudhary Charan Singh University, Meerut

M.Sc. Statistics, Semester III
Academic Session 2026–2027

Previous lecture

Lecture 47: Runs: the null distribution of the number of runs.

Today

- Tests for Randomness Based on Runs
 - Conditional Distribution of the Number of Runs
 - One-Sample Runs Test about a Specified Centre

Theorem 48.1 (Mean and variance of the number of runs)

Under randomness, conditionally on (n_1, n_2) ,

$$\mathbb{E}(R) = 1 + \frac{2n_1n_2}{n}, \quad \text{Var}(R) = \frac{2n_1n_2\{2n_1n_2 - n\}}{n^2(n-1)}.$$

Proof.

For $i = 1, \dots, n-1$ let $I_i = 1$ if positions i and $i+1$ carry different symbols and $I_i = 0$ otherwise. Every change of symbol starts a new run, so

$$R = 1 + \sum_{i=1}^{n-1} I_i.$$

Single indicators. Positions $i, i+1$ read AB with probability $\frac{n_1}{n} \cdot \frac{n_2}{n-1}$ and BA with the same probability, so

$$\mathbb{P}(I_i = 1) = p := \frac{2n_1n_2}{n(n-1)}, \quad \mathbb{E}(R) = 1 + (n-1)p = 1 + \frac{2n_1n_2}{n}.$$

Proof (cont.)

Adjacent pairs ($j = i + 1$; there are $n - 2$ such pairs). $I_i I_{i+1} = 1$ exactly when positions $i, i + 1, i + 2$ read ABA or BAB :

$$\begin{aligned}\mathbb{E}(I_i I_{i+1}) &= \frac{n_1 n_2 (n_1 - 1) + n_2 n_1 (n_2 - 1)}{n(n-1)(n-2)} \\ &= \frac{n_1 n_2 (n-2)}{n(n-1)(n-2)} = \frac{n_1 n_2}{n(n-1)} = \frac{p}{2}.\end{aligned}$$

Non-adjacent pairs ($j \geq i + 2$; there are $(n-2)(n-3)/2$ such pairs). The four positions are distinct, and each of the two pairs must contain one A and one B (in either order), so

$$\mathbb{E}(I_i I_j) = q := \frac{4n_1 n_2 (n_1 - 1)(n_2 - 1)}{n(n-1)(n-2)(n-3)}.$$

Proof (cont.)

Second moment. Since $I_i^2 = I_i$,

$$\begin{aligned}\mathbb{E}\left(\sum_i I_i\right)^2 &= \sum_i \mathbb{E}(I_i) + 2 \sum_{i < j} \mathbb{E}(I_i I_j) \\ &= (n-1)p + 2(n-2)\frac{p}{2} + (n-2)(n-3)q.\end{aligned}$$

Substituting,

$$\begin{aligned}(n-1)p &= \frac{2n_1 n_2}{n}, & (n-2)p &= \frac{2n_1 n_2 (n-2)}{n(n-1)}, \\ (n-2)(n-3)q &= \frac{4n_1 n_2 (n_1 - 1)(n_2 - 1)}{n(n-1)}.\end{aligned}$$

Proof (cont.)

Hence, as $\text{Var}(R) = \text{Var}(\sum_i I_i) = \mathbb{E}(\sum_i I_i)^2 - \{(n-1)p\}^2$,

$$\begin{aligned}\text{Var}(R) &= \frac{2n_1n_2}{n} + \frac{2n_1n_2(n-2)}{n(n-1)} \\ &\quad + \frac{4n_1n_2(n_1-1)(n_2-1)}{n(n-1)} - \frac{4n_1^2n_2^2}{n^2}.\end{aligned}$$

Proof (cont.)

Over the common denominator $n^2(n-1)$, the numerator is $2n_1n_2$ times

$$\begin{aligned} & n(n-1) + n(n-2) + 2n(n_1-1)(n_2-1) - 2n_1n_2(n-1) \\ &= 2n^2 - 3n + 2n n_1n_2 - 2n(n_1 + n_2) + 2n - 2n n_1n_2 + 2n_1n_2 \\ &= 2n^2 - 3n - 2n^2 + 2n + 2n_1n_2 \quad (\text{using } n_1 + n_2 = n) \\ &= 2n_1n_2 - n. \end{aligned}$$

Therefore $\text{Var}(R) = 2n_1n_2(2n_1n_2 - n)/\{n^2(n-1)\}$. □

For moderate samples,

$$Z = \frac{R - \mathbb{E}(R)}{\sqrt{\text{Var}(R)}}$$

is approximately standard normal.

A continuity correction should be applied in the direction of the relevant tail:

- for a lower-tail probability use $R + \frac{1}{2}$,
- for an upper-tail probability use $R - \frac{1}{2}$.

Exact critical values of R for $n_1, n_2 \leq 20$: Tables in Appendix B of the lecture notes.

One-sample runs test about a specified centre

Given observations $X_1, \dots, X_n$ in their natural time or spatial order and a cut point c (either a hypothesised median η_0 fixed in advance, or the sample median), replace each observation by $+$ if $X_i > c$ and by $-$ if $X_i < c$; omit observations equal to c . The number of runs is tested using the preceding distribution, conditionally on n_1 and n_2 .

One-sample runs test

1. State H_0 : the observations occur in random order.
2. Convert the observations into two symbols using a specified cut point, commonly a hypothesised or sample median.
3. Count n_1 , n_2 and the number of runs R .
4. Use the exact conditional distribution for small samples or the normal approximation for moderate samples.
5. Reject for very small or very large values of R in a two-sided test.

Example: a one-sample runs test

Example 48.2

Consider the sign sequence

$$+, +, +, +, -, -, -, -, +, +.$$

Test the hypothesis of randomness.

Solution

Here $n_1 = 6$, $n_2 = 4$ and $R = 3$. Under randomness,

$$\mathbb{E}(R) = 1 + \frac{2(6)(4)}{10} = 5.8, \quad \text{Var}(R) = \frac{2(6)(4)\{2(6)(4) - 10\}}{10^2(9)} = 2.0267.$$

The observed number of runs is below its expectation, suggesting clustering.

Example: a one-sample runs test (cont.)

Solution (cont.)

Exact test. By the Theorem on the null distribution of the number of runs, with $\binom{10}{6} = 210$,

$$\mathbb{P}(R = 2) = \frac{2\binom{5}{0}\binom{3}{0}}{210} = \frac{2}{210},$$

$$\mathbb{P}(R = 3) = \frac{\binom{5}{1}\binom{3}{0} + \binom{5}{0}\binom{3}{1}}{210} = \frac{5 + 3}{210} = \frac{8}{210},$$

so the exact lower-tail probability is

$$\mathbb{P}(R \leq 3) = \frac{10}{210} = 0.0476.$$

Example: a one-sample runs test (cont.)

Solution (cont.)

Normal approximation. With continuity correction,

$$z = \frac{3 + 0.5 - 5.8}{\sqrt{2.0267}} = -1.616 \quad \text{and} \quad \Phi(-1.616) = 0.0531,$$

close to the exact value; for samples this small the exact value should be used.

Conclusion. Against the one-sided alternative of clustering, the exact p -value $0.0476 < 0.05$, so randomness is rejected at the 5% level.

Example: a one-sample runs test (cont.)

Solution (cont.)

Against a two-sided alternative (clustering or alternation), doubling gives $2(0.0476) = 0.095$; but the null distribution of R is not symmetric, so it is better to add the probabilities of all values at least as extreme as the one observed.

Here only $R = 9$ lies as far above $\mathbb{E}(R) = 5.8$ as $R = 3$ lies below it, and

$$\mathbb{P}(R = 9) = \frac{\binom{5}{4}\binom{3}{3} + \binom{5}{3}\binom{3}{4}}{210} = \frac{5}{210},$$

giving the exact two-sided value $15/210 = 0.0714$.

Either way randomness is not rejected at the 5% level against a two-sided alternative. The alternative must therefore be fixed before the data are examined.

Practice (Lecture Notes, Chapter 3)

- Short-answer questions (Section 3.14): Q6.
- Exercises (Section 3.15): 3.9, 3.10.

Next lecture

Lecture 49: Runs up and down; the Wald–Wolfowitz two-sample runs test.