Roulettician: *article #4*

A Test for Randomness

The randomness of a data sequence or series of trials is often taken for granted, but how do we know that a given sequence really *is* random? As I explained in previous articles, 'random' means that outcomes are unbiased and independent. In the second probability tutorial I described a simple way in which we can test for independence. In this article I'll introduce another, more sophisticated test which can be applied to *any* set of data, even when nothing is known about where it comes from. The test works purely on the *order* of the observations.

Significance Tests

Before describing the test and how to apply it, a word about the logic of this and similar tests. This is important so that you know how to properly interpret the results.

The basic idea is pretty simple and intuitive; we *assume* that the data sequence is random, perform the test, and if the result of the test would only occur a small percentage of the time if the data really *was* random, the result is said to be **significant**. The assumption that the data is random is called the **null hypothesis**, and is what the research is attempting to *disprove*.

There are 'levels' of significance which correspond to the percentage. For example, suppose that you suspect that a Roulette wheel is biased and that a particular sector is hitting more often than it should. Your null hypothesis is:

```
    'The wheel is unbiased and this sector will hit as often as any other'.
```

You then run an experiment (record spins), and summarize the data in the form of a statistic. If this statistic shows that your results would only occur 1% of the time if the null hypothesis was true (a fair wheel), then the result is declared significant at the 1% level.

Now, it's important to understand that this does *not* necessarily mean that we can deny our assumption that the wheel is fair. A word of caution is in order:

All we can say is that either something unusual has happened (probability 1 in 100), *or* our assumption of randomness is false.

Of course, if you did such an experiment enough times then you are going to get a 'significant' result occasionally, but in that case it wouldn't really be significant in the sense that you were hoping for.

Significance tests like this are for situations where we don't really understand, in any theoretical sense, what's going on. A science like physics is 'theoretical' in that there are laws, such as the principles of mechanics, which we can apply to any physical bodies and deduce consequences because the principles are invariable. On the other hand, in agriculture (significance tests were invented for farmers), medicine or psychology, there is very little deep theoretical knowledge.

Scientists usually don't understand very well why various drugs or medicines are effective, although they can tell, empirically, that some treatments are better than others. Significance tests are helpful for testing new treatments and can also help in finding out whether different factors are associated or correlated with each other.

The Runs Test

The **Runs Test** is used to test the independence of a set of data where the order in which the data was collected is preserved. This is consistent with the idea that independence is related to the regularity of outcomes; the more regularity there is in a set of data, the less the likelihood that it's independent. For example, which of the following sequences of R/B looks the most random?

1. **R R R R B B B B**
2. **R B B R R B R B**
3. **R B R B R B R B**

Most would agree that it's sequence 2; the others just look too regular. In the context of a data sequence, a **run** is a sequence of values (or labels, such as 'R', 'B') which is isolated by other sequences with different values or labels. For example, the following sequence of odd/even outcomes consists of 9 runs:

O EE O EEE OO E O EEEE OO

The runs test only applies to **binomial** data, meaning two outcomes. However, this isn't as restrictive as it first appears, because if there are more than two classes, you can combine some into one class. For example, for dozen or column outcomes, there are 3 classes, but combining any 2 of them reduces the data to binomial. E.g.:

D1 D2D2 D3 D1 D3 D1D1 D2D2

isn't binomial data, but if we combine dozen 2 and dozen 3 into one class (call it 'D'), then we get:

D1 DDD D1 D D1D1 DD

which *is* binomial, and there are 6 runs.

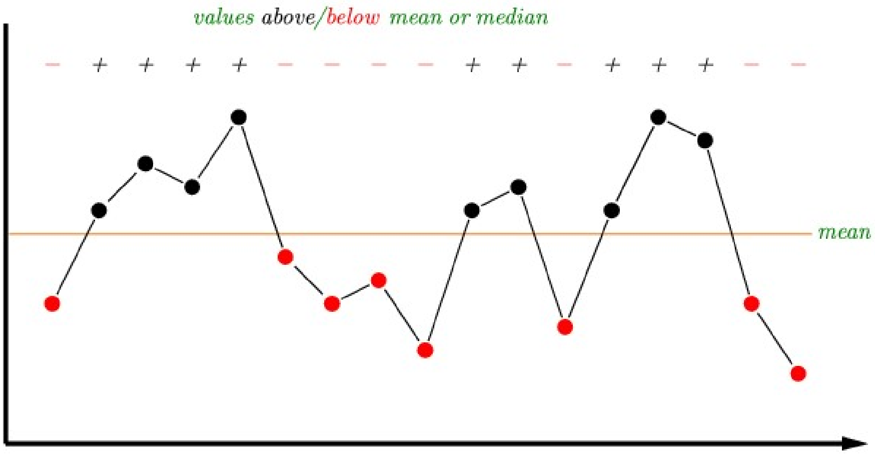
So far I've only been using *categorical* data, that is, data which has been put into a category or group such as red/black, or a dozen. But the runs test can also be used on *numerical* data. Of course, the numbers should have some meaningful interpretation, and it's up to you where they come from. A couple of suggestions:

- The number of 'gaps' between hits for a particular bet. E.g., you record spins and count the number of spins which occur between hits of street 4-6. They might be: 7, 15, 4, 1, 29, 8, 12, 17. That means you had to wait 7 spins before street 4-6 first hit, then 15 spins before it hit again, 4 spins before it hit again, and so on.
- The distance, counting clockwise or anticlockwise, between successive pockets on the wheel where the ball lands. E.g. the first number to hit is 5 and the 2nd number is 14. Counting clockwise from 5 to 14, there are 6 pockets, so the first number in your sequence is 6. The next pocket the ball lands in is 36. Again counting clockwise from 14, there are 25 pockets between it and 36, so the next number in the sequence is 25.

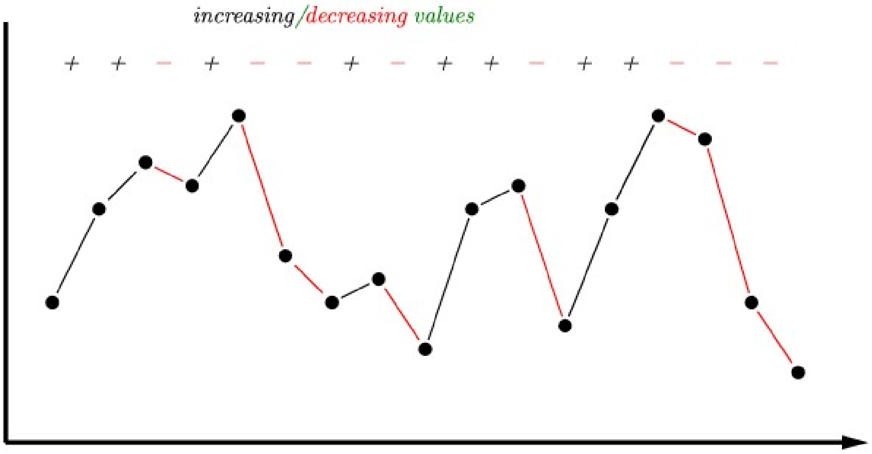
Ok, so we have our data sequence where the order of data elements is preserved, but the data *must* be binomial to apply the runs test. That means we should be able to allocate each numerical data element to one of two possible classes — how do we do that?

We have two options.

Option 1. Find the mean or median of the data set. If any item of data is above the mean or median, we give it a $+$ sign, if it's below the mean or median, it gets a $-$ sign. Any datum which falls exactly on the mean or median is ignored. Like this:



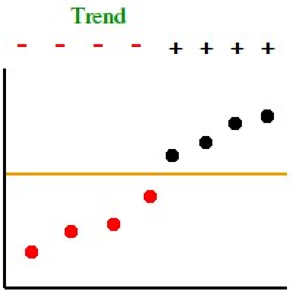
Option 2. Compare successive values of the sequence. If the next item of data is higher than the previous item, it gets a $+$ sign. If it's lower, it gets a $-$ sign. If successive items have the same numerical value, we skip that item and move to the next one. The plot below illustrates this scenario:

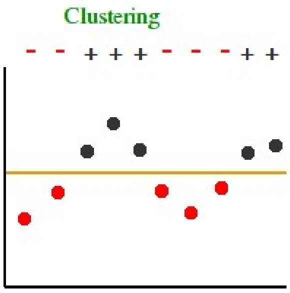


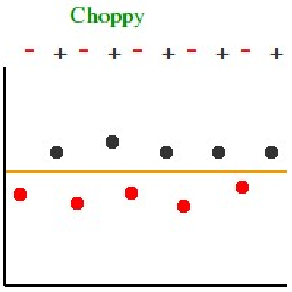
Notice that the shape of the plots are identical because the same data points were plotted in both cases, but the pattern of + 's and − 's are different. Of course, you'd expect this, because the classes are defined differently, but it does suggest that there is no such thing as an 'objectively' random data sequence; it depends what you're measuring the randomness with respect to.

Whichever option we choose (it might even be both), the numerical data series will have been transformed into a binomial series, and we can apply the runs test to it.

The runs test counts the number of runs in a data sequence. This number is a good indication of whether the data is random or not, because if there are too few or too many runs relative to the length of the data sequence, a lack of independence between values is suggested. So the lack of randomness may take various forms, some of which are illustrated below.







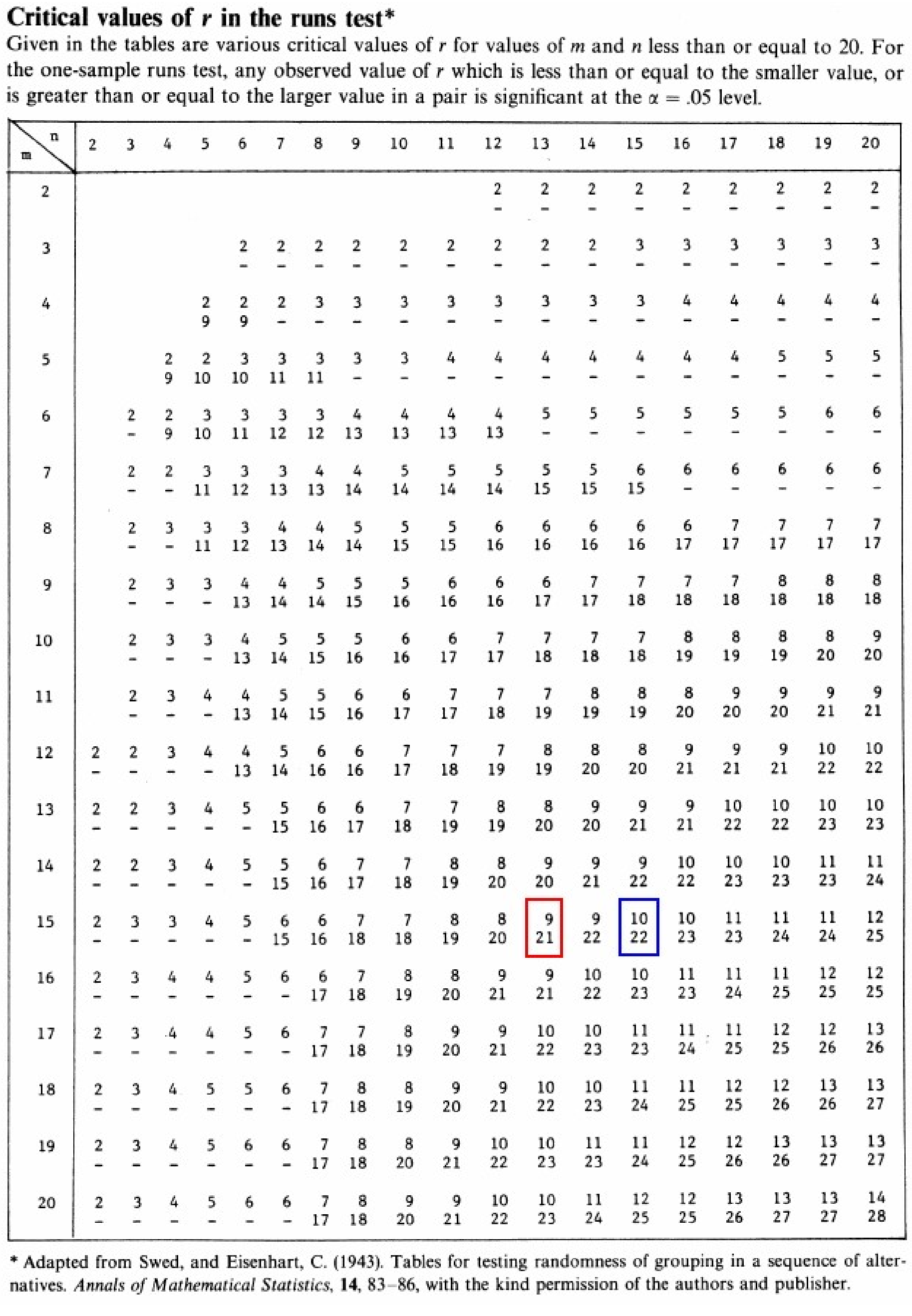
Applying the Test

Given a data series, we need some criteria for deciding when the number of runs is so extreme that we should reject the null hypothesis (which is that the data is random).

If there are not many elements in the data series, we can use a table of values which have been formulated by the statisticians who developed this test. Once we have our data sequence, we just count the number of runs and look in the table to see whether we should reject the null hypothesis (i.e., whether there is some evidence that the data is non-random, or unusual).

The table is valid for a significance level of 5%, which means that if the number of runs in our sequence is outside or equal to the interval boundaries given in the table, then it occurs with probability 1 in 20 (or less).

The table should be used when the number of elements in either of the classes is less than or equal to 20. If this is not the case, then we use a formula. But first, I'll give an example of how to use the table (shown below).



Critical values of *r* in the runs test*

Given in the tables are various critical values of *r* for values of *m* and *n* less than or equal to 20. For the one-sample runs test, any observed value of *r* which is less than or equal to the smaller value, or is greater than or equal to the larger value in a pair is significant at the $\alpha = .05$ level.

m \ n	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2											2 –	2 –	2 –	2 –	2 –	2 –	2 –	2 –	2 –
3					2 –	2 –	2 –	2 –	2 –	2 –	2 –	2 –	2 –	3 –	3 –	3 –	3 –	3 –	3 –
4				2 9	2 9	2 –	3 –	3 –	3 –	3 –	3 –	3 –	3 –	3 –	4 –	4 –	4 –	4 –	4 –
5			2 9	2 10	3 10	3 11	3 11	3 –	3 –	4 –	4 –	4 –	4 –	4 –	4 –	4 –	5 –	5 –	5 –
6		2 –	2 9	3 10	3 11	3 12	3 12	4 13	4 13	4 13	4 13	5 –	5 –	5 –	5 –	5 –	5 –	6 –	6 –
7		2 –	2 –	3 11	3 12	3 13	4 13	4 14	5 14	5 14	5 14	5 15	5 15	6 15	6 –	6 –	6 –	6 –	6 –
8		2 –	3 –	3 11	3 12	4 13	4 14	5 14	5 15	5 15	6 16	6 16	6 16	6 16	6 17	7 17	7 17	7 17	7 17
9		2 –	3 –	3 –	4 13	4 14	5 14	5 15	5 16	6 16	6 16	6 17	7 17	7 18	7 18	7 18	8 18	8 18	8 18
10		2 –	3 –	3 –	4 13	5 14	5 15	5 16	6 16	6 17	7 17	7 18	7 18	7 18	8 19	8 19	8 19	8 20	9 20
11		2 –	3 –	4 –	4 13	5 14	5 15	6 16	6 17	7 17	7 18	7 19	8 19	8 19	8 20	9 20	9 20	9 21	9 21
12	2 –	2 –	3 –	4 –	4 13	5 14	6 16	6 16	7 17	7 18	7 19	8 19	8 20	8 20	9 21	9 21	9 21	10 22	10 22
13	2 –	2 –	3 –	4 –	5 –	5 15	6 16	6 17	7 18	7 19	8 19	8 20	9 20	9 21	9 21	10 22	10 22	10 23	10 23
14	2 –	2 –	3 –	4 –	5 –	5 15	6 16	7 17	7 18	8 19	8 20	9 20	9 21	9 22	10 22	10 23	10 23	11 23	11 24
15	2 –	3 –	3 –	4 –	5 –	6 15	6 16	7 18	7 18	8 19	8 20	9 21	9 22	10 22	10 23	11 23	11 24	11 24	12 25
16	2 –	3 –	4 –	4 –	5 –	6 –	6 17	7 18	8 19	8 20	9 21	9 21	10 22	10 23	11 23	11 24	11 25	12 25	12 25
17	2 –	3 –	4 –	4 –	5 –	6 –	7 17	7 18	8 19	9 20	9 21	10 22	10 23	11 23	11 24	11 25	12 25	12 26	13 26
18	2 –	3 –	4 –	5 –	5 –	6 –	7 17	8 18	8 19	9 20	9 21	10 22	10 23	11 24	11 25	12 25	12 26	13 26	13 27
19	2 –	3 –	4 –	5 –	6 –	6 –	7 17	8 18	8 20	9 21	10 22	10 23	11 23	11 24	12 25	12 26	13 26	13 27	13 27
20	2 –	3 –	4 –	5 –	6 –	6 –	7 17	8 18	9 20	9 21	10 22	10 23	11 24	12 25	12 25	13 26	13 27	13 27	14 28

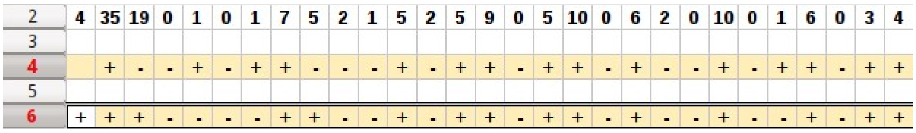
* Adapted from Swed, and Eisenhart, C. (1943). Tables for testing randomness of grouping in a sequence of alternatives. *Annals of Mathematical Statistics*, **14**, 83–86, with the kind permission of the authors and publisher.

Suppose you collect data on the gap lengths (number of spins between successive hits) for a certain sector of the wheel covering 6 numbers. Here's the sequence (remember — it's crucial that the *order* in which the data was collected is preserved):

$$4, 35, 19, 0, 1, 0, 1, 7, 5, 2, 1, 5, 2, 5, 9, 0, 5, 10, 0, 6, 2, 0, 10, 0, 1, 6, 0, 3, 4$$

Since this is a sequence of numbers (not a sequence of categories such as R/B), we need transform the data into a binary (two-valued) sequence. We could do this by looking for values above/below the mean or median, or compare successive values and classify by an increase or decrease.

Let's do both. First compare successive values; if there is a decrease we assign the element a $-$, and if an increase assign it a $+$. This is shown in row 4 below. Next, compare values with the median which is 2.5 (I calculated this in a spreadsheet using the `median()` function). If the sequence value is greater than 2.5 add a plus, otherwise add a minus. The resulting series is shown in row 6.



2	4	35	19	0	1	0	1	7	5	2	1	5	2	5	9	0	5	10	0	6	2	0	10	0	1	6	0	3	4
3																													
4		+	-	-	+	-	+	+	-	-	-	+	-	+	+	-	+	+	-	+	-	-	+	-	+	+	-	+	+
5																													
6	+	+	+	-	-	-	-	+	+	-	-	+	-	+	+	-	+	+	-	+	-	-	+	-	-	+	-	+	+

We are ready to use the table to determine whether the number of runs is significant. Referring to row 4, there are 15 $+$'s, 13 $-$'s, and 19 runs. The leftmost column and top row in the table refer to m and n respectively, which are labels for the classes (it doesn't matter which you call m and n , but let's arbitrarily decide that $m = +$ and $n = -$).

Now it's just a matter of reading down the column to 15, the number of $+$'s, and across to 13, the number of $-$'s. At the intersection of this row and column you'll see a pair of numbers which represent the 'critical' values (marked by the red box). If the number of runs in your data sequence is between these values (not inclusive), then there is no evidence, at the 5% significance level, that the sequence is random.

Since the number of runs is 19, and this is greater than 9 and less than 21, we cannot say that this sequence is anything other than random.

Now we'll look at values above/below the median. Refer to row 6 above. There are 15 $+$'s and 15 $-$'s which make up 18 runs. Find the intersection of row 15 and column 15 in the table. The critical values are between 10 and 22 (marked by the blue box), so because 18 falls between these values, again there is no evidence that the sequence is anything other than random, or that outcomes are varying systematically.

A Formula

When your data sequence increases to a length such that m and n are greater than 20 (the maximum values in the table), you can use formulas for the mean and standard deviation of the number of runs, which are based on the well known bell curve. Then we calculate a z-score, which gives the number of standard deviations from the theoretical mean of the sequence, in terms of the number of runs.

If you're not sure what all those terms mean, and frankly, don't want to know, then all is not lost, because you can use the calculator below. Just enter the values of m , n , and r and click the button. However, for the sake of completeness, and for those who might want to use the formula in their own spreadsheet or program, here is the formula for the z-score and its components:

$$z = \frac{r - \mu_r}{\sigma_r}$$

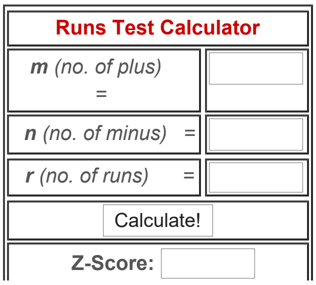
Where r is the number of runs, and μ_r , the average number of runs is given by:

$$\mu_r = \frac{2mn}{N} + 1 \text{ , where } N = m + n \text{ ,}$$

and σ_r , the standard deviation of the number of runs, is given by:

$$\sigma_r = \sqrt{\frac{2mn(2mn - N)}{N^2(N-1)}}$$

It's important that you only use this formula (or the calculator) when the numbers of elements in the classes are such that you can't use the table (if m or n is more than 20), otherwise the results will be misleading. Here's the calculator:



Runs Test Calculator

m *(no. of plus)* =	
n *(no. of minus)* =	
r *(no. of runs)* =	

Z-Score:

What the Z-Score means

Looking up the critical values in the table tells us whether we're entitled to reject or not reject the null hypothesis at the 5% level of significance, but the calculator gives us a number — what does this number mean?

There are two considerations: the magnitude of the number and the sign of it. We'll consider the sign first (whether it's positive or negative). Remember that if the result of the runs test is 'significant', then either there are too many or not enough runs, relative to the length of the data sequence.

- If there are *too few* runs, the formula (or calculator) will return a negative number.
- If there are *too many* runs, the formula (or calculator) will return a positive number.

Try it. Put $m = 20$, $n = 21$, and $r = 12$ into the calculator. You'll get a value of -3 . The minus sign indicates that this data sequence is 'streaky' — there are not enough runs. A sequence of R/B outcomes which fits the description might look like this:

RRR BBB RRR B RRRRRR BBBB RRR B RRR BBBBBBB RR BBBBB

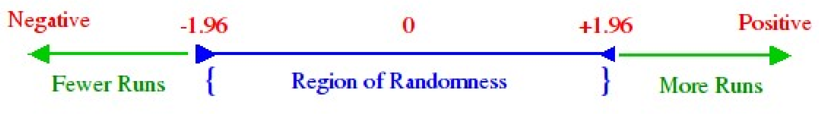
There are 20 reds, 21 blacks, and 12 runs. On the other hand, try leaving the values of m and n the same but change the number of runs to 30. The calculator will return a value of 2.69. This is a relatively large positive number, indicating that the data sequence is 'choppy' — there are too many runs. These values might correspond to a R/B sequence like this:

RR B R BB R B R B R BB R B RR B R B RR B R B RR BB R B R B R BB RR BBB

Ok, so that's the sign of the z-score dealt with. The *magnitude* of the number (how big it is, regardless of its sign) tells us the degree of significance we should attach to it. A larger z-score will indicate a greater departure from randomness or independence than a smaller number (alternatively, larger z-scores indicate more unusual events).

We can relate the z-score to the 5% significance level used in the table. A z-score of magnitude 1.96 corresponds to a 5% level, meaning that there is only 1 chance in 20 of getting such a score. So if the z-score is higher than or equal to 1.96, there is some evidence that the sequence is not random (or the sequence is relatively rare, with respect to the number of runs).

The diagram below may help to give a feel for what the Z-Score means, and how to interpret it.



The 'region of randomness' is represented by the blue interval. Outside it, in both directions, the results become significant at the 5% level. Here are some examples of possible scores and how to interpret them:

- Z-Score $= 0.34$: Less than 1.96, so there is no evidence of non-randomness.
- Z-Score $= -1.25$: Less than 1.96, so there is no evidence of non-randomness.
- Z-Score $= 2.3$: More than 1.96 and positive, so there is some evidence of non-randomness in the direction of too many runs.
- Z-Score $= 0$: Less than 1.96, so there is no evidence of non-randomness. In fact, a score of 0 indicates that the number of runs is right on the average.
- Z-Score $= -2.7$: More than 1.96 and negative, so there is evidence of non-randomness in the direction of not enough runs.

The runs test is not the only test for randomness; there are many others. In order to test the fitness of a random number generator (used for gaming purposes, perhaps), it's necessary to apply many such tests. But for checking the independence of a sequence of binary data, the runs test is popular, flexible, and easy to apply.