Runs Test for Randomness

Definition: A **Run** is a sequence of data having the same characteristics (such as below the median). This sequence is preceded and followed by data with a different characteristic or by no data. For our purposes we will consider data with only two characteristics.

Rationale:
Suppose we have a sequence of 9 items, called H (high) and L(Low). Further suppose there are 4 H and 5 L. There are 9!/(4!*5!) = 126 different arrangements of H and L.

If we were to list all the cases, there would be
1 case with 9 runs (L H L H L H L H L)
The probability of 9 runs would be 1/126 = 0.007937 or 0.7%

8 cases with 8 runs (L L H L H L H L H) (LH L LH L H L H)
(L H L H LL H L H) (L H L H L H LL H)
(H L LH L H L H L) (H L H L H L H L L)
(H L H LL H L H L) (H L H L H LL H L
The probability of 8 runs would be 2/126 = 0.06349 or 6.3%

Obviously we don't want to do this forever! Yes there are tables for small values, and there is a normal approximation to use when the table is not sufficient.

Assumptions

1. The sample data are arranged according to some scheme (such as time series)
2. The data falls into two separate categories (such as above and below a specific value).
3. The runs test is based on the order in which the data occur; not on the frequency of the data.

Notation:
n_1 = number of elements in the sequence with characteristic 1
n_2 = number of elements in the sequence with characteristic 2
G = number of runs.

For a small sample, simply use the table in the back of the book. However this table assumes a significance level of 5%.

For a large sample (at least one of the samples is greater than 20) or a different significance level use the following:

$$m_G = \frac{2n_1n_2}{n_1+n_2}+1$$

$$s_G = \sqrt{\frac{(2n_1n_2)(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}}$$

$$TS = \frac{G - m_G}{s_G}$$

The Critical value is from the z table.

Example 1.

I flipped a coin 20 times (yes, really)

HH TT HHH T H TT H T HH T HHHH

Is this pattern random?

I got H 13 times, T 7 times. There are 11 runs

The table in the back of the book shows 5 and 14 as bounds. 11 is between these values, so we conclude that the data is random.

Example 2

I used Excel to simulate rolling a pair of dice 50 times and recording the results. Do "doubles" come up randomly?

(I used 1 for doubles, 0 for not doubles)

0	0	1	0	1	1	0	0	0	0
0	0	1	0	1	0	0	0	0	0
0	1	0	0	0	0	0	0	1	0
1	0	0	1	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0

10 doubles, 40 not doubles, 17 runs

$$m_G = \frac{2n_1n_2}{n_1 + n_2} + 1 = \frac{2 \cdot 10 \cdot 40}{10 + 40} + 1 = \frac{800}{50} + 1 = 17$$

$$s_G = \sqrt{\frac{(2n_1n_2)(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}} = \sqrt{\frac{(2 \cdot 10 \cdot 40)(2 \cdot 10 \cdot 40 - 10 - 40)}{(10 + 40)^2 (10 + 40 - 1)}} = \sqrt{\frac{(800)(750)}{(50)^2 (49)}} = 2.213$$

. $TS = \frac{G - m_G}{s_G} = \frac{17 - 17}{2.212} = 0$

Well, I really don't have to worry about the alpha level. This is legitimately random!

Example 3.

I looked in Bookshelf 98 for some data and found the per capita totals of money in circulation. Consider whether it increased or decreased. Is the pattern random?

year	money	change
1910	34.07	
1915	33.01	-1
1920	51.36	1
1925	41.56	-1
1930	36.74	-1
1935	43.75	1
1940	59.4	1
1945	191.14	1
1950	179.03	-1
1955	182.9	1
1960	177.47	-1
1965	204.14	1
1970	265.39	1
1975	380.08	1
1980	558.28	1
1985	778.58	1
1990	1028.71	1
1995	1531.39	1

Number decreases -5
Number increases - 12
Number of runs 8

Table says bounds are 3 and 10, so this is random.

$$m_G = \frac{2n_1 n_2}{n_1 + n_2} + 1 = \frac{2 \cdot 5 \cdot 12}{5 + 12} + 1 = 8.1$$

$$s_G = \sqrt{\frac{(2n_1 n_2)(2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}} = \sqrt{\frac{(2 \cdot 5 \cdot 12)(2 \cdot 5 \cdot 12 - 5 - 12)}{(5 + 12)^2 (5 + 12 - 1)}} = \sqrt{\frac{12360}{4624}} = 1.63$$

$$TS = \frac{8 - 8.1}{1.63} = 0.06$$

Still confirmed to be random! I’m surprised.

Example 4

Excel does NOT have this test built in. Just a quick Minitab run…

Bluman 14-95. A supervisor records the number of employees absent over a 30 day period. Test for randomness.

We have to check above and below the mean. If we were doing this by hand, we would first calculate the mean – value is 13.733. Next we could turn the data into + and – signs, count runs, etc.

However, I'll use Minitab first and then verify manually.
>Stat>Non Parametric> Runs test.
Check above/below the mean

Data Display

```
     27      6     19     24     18     12     15     17     18     20
      0      9      4     12      3      2      7      7      0      5
     32     16     38     31     27     15      5      9      4     10
```

Runs Test: C1

```
    K =    13.7333


    The observed number of runs =   8
    The expected number of runs =  15.9333
    14 Observations above K   16 below
              The test is significant at  0.0031
```

+ 0 +++ 0 ++++ 00000000000 ++++++ 0000

$$m_G = \frac{2n_1n_2}{n_1+n_2}+1=\frac{2\cdot 14\cdot 16}{14+16}+1=15.9$$

$$s_G = \sqrt{\frac{(2n_1n_2)(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}=\sqrt{\frac{(2\cdot 14\cdot 16)(2\cdot 14\cdot 16-14-16)}{(14+16)^2(14+16-1)}}=2.68$$

$$TS=\frac{8-15.9}{2.68}=-2.95$$

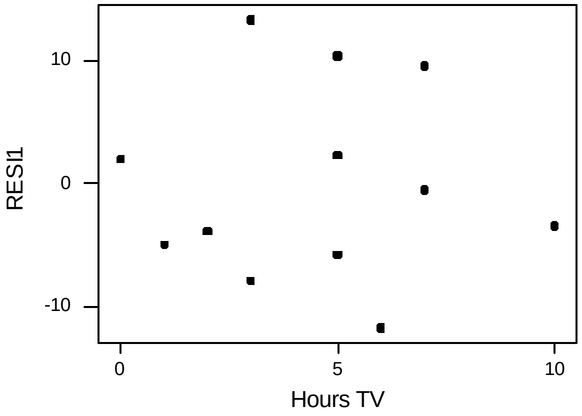
If alpha = 0.01, CZ = +/- 2.575,
Randomness is rejected..

Example 5

One last use for this test. Remember simple linear regression? When we look at the residuals (difference between predicted and actual values), we hope the results are random. Recall the television and test scores data. We perform a regression and look at a residual plot.

```
Row  Hours TV   Test     RESI1

  1         0     96    2.0300
  2         1     85   -4.9026
  3         2     82   -3.8352
  4         3     74   -7.7678
  5         3     95   13.2322
  6         5     68   -5.6330
  7         5     76    2.3670
  8         5     84   10.3670
  9         6     58  -11.5655
 10         7     65   -0.4981
 11         7     75    9.5019
 12        10     50   -3.2959
```



Looks Random, but it is? Let's check above and below zero.

Minitab results

```
    RESI1

    K =      0.0000


    The observed number of runs =   8
    The expected number of runs =   6.8333
     5 Observations above K    7 below
 * N Small -- The following approximation may be invalid
              The test is significant at  0.4662
              Cannot reject at alpha = 0.05
```

Manual Results

+ 000 + 0 ++ 00 + 0
8 runs, 7 0, 5 –
Table in back of book has bounds 3 and 11. Do not reject randomness.

Comment: since there are duplicate x values, this really isn't quite right. If the data was a time series, with only one occurrence of each time, it would be correct. I am trying to illustrate another way to use this procedure.