

Heritage Windows Programs from the RT Package

Bryan F.J. Manly
Western EcoSystems Technology Inc.
Cheyenne, Wyoming
bmanly@west-inc.com

Contents

| | |
|--------------------------------|----|
| 1. Introduction..... | 2 |
| 2. The Heritage Programs. | 2 |
| 3. Running the Programs. | 5 |
| 4. Data Files..... | 6 |
| 5. An Example. | 9 |
| References..... | 11 |

1. Introduction

The RT (randomization testing) package was originally developed to carry out calculations described in the book *Randomization, Bootstrap and Monte Carlo Methods in Biology* (1st edition 1991, 2nd edition 1997, 3rd edition 2007). The package contained a number of Windows programs mainly for randomization tests, plus a program for choosing which analysis to do, a program for entering and transforming data and a program for calculating distance matrices. None of the programs in the package would run on a 64 bit computer.

The RT package is no longer available for purchase but the Windows programs for data analysis have been recompiled so that they run on 32 bit and 64 bit Windows computers and these are available to be freely downloaded from the Western EcoSystems technology website www.west-inc.com. These programs have been used by many people in the past and are believed to be free of errors. However, they are provided with no guarantees in that respect. All random number generation uses the algorithm proposed by Wichmann and Hill (1982).

2. The Heritage Programs

The heritage programs carry out the following range of randomization and Monte Carlo tests:

- RT1Samp is for one-sample methods: the paired comparison test or Fisher's (1966) randomization test that he first published in 1935, with the option of determining a randomization confidence interval for a population mean difference. With this program one column of an RT data file contains the data to be analysed.
- RT2Samp: is for two sample methods: the comparison of two sample means and the two sample variance ratio, with the option of determining a randomization confidence interval for a population mean difference. With this option two columns in an RT data file are analysed. One column must hold the sample numbers for the cases (1 or 2), and the other must hold the data values. The observed sample mean difference and the sample variance ratio are compared with the distributions obtained by randomly allocating data values to the two samples. Randomization confidence limits for the difference between the populations that the samples are drawn from can also be obtained.
- RTAnova for analysis of variance and related methods: one, two or three factor analyses, allowing for unequal numbers of replicates if there is only one factor, with tests for unequal error variance at different factor levels when this is possible, and randomization of either observations or residuals. One or more columns in the data file hold the factor levels and one column holds the data values. For example, for a two factor analysis one column holding the levels of factor one, one column holding the

levels of factor two, and one column holding data values must be chosen. There must be equal numbers of replicates if there are two or three factors. If possible, the Bartlett statistic and a variance range statistic are used to test for unequal residual variances. There is also a facility to carry out the analysis on transformed data. Two types of randomization are possible. First, the original observations can be randomized. In this case the test statistics are the sums of squares for the various effects and interactions as percentages of the total sum of squares, Bartlett's statistic for unequal residual variance, and the variance range as a percentage of the error mean square. Because randomizing the original observations does not change the total sum of squares, using percentage sums of squares is equivalent to using the sums of squares directly. The variance range is an alternative to Bartlett's test statistic that can be used when the observations are all the same for some factor combinations. Expressing this range as a percentage of the error mean square is intended to reduce the effect of changes in the mean with different factor combinations. The second method of randomization involves randomizing the residuals from the analysis of variance model, following the ideas of ter Braak (1992). For this approach to work well it is necessary to have scale free test statistics. For this reason, the statistics used to test the sums of squares for the various effects and interactions are F-ratios rather than the percentages of the total sum of squares as used when randomizing the original observations. Bartlett's statistic for unequal residual variance and the variance range as a percentage of the error mean square are used for testing for unequal residual variances at different factor levels.

- RTRegr for linear regression: simple and multiple regression, with randomization of either Y values or residuals. One column in the data file holds the dependent variable Y and the other column or columns hold the values for the X variable or variables. There can be up to 20 X variables and up to 5000 cases. Two types of randomization are possible. The first involves randomizing the order of the Y values and the second involves finding the regression residuals for the original data and then recalculating the regression equation on the residuals in a random order to determine significance. This yields a type of approximate randomization test that may have more power than is obtained by simply randomizing the Y values. The test statistics used with this option are (a) the t-statistics which are the estimated regression coefficients divided by their estimated standard errors, (b) F-statistics which are the extra sums of squares obtained by adding variables to the equation one at a time in order divided by the error mean square, and (c) the regression mean square divided by the error mean square.
- RTMant for matrix randomization tests: Mantel (1967) type tests with the distances between n objects in one matrix being related to the distances in up to nine other matrices. The distances are in a standard RT data file. The analysis is similar to the one given by the linear regression option except that randomization is by permuting the rows and columns of the dependent matrix. One column in the data matrix holds the distances for the dependent matrix. One or more other columns hold the distances for the other matrices. All the columns to be used are chosen from those available by the

user. With ten input matrices, up to 31,125 distances are allowed, corresponding to the distances between 250 cases.

- RTMant1 for matrix randomization tests allowing for a restricted randomization as described by Manly (2007, Example 9.3). The distances are in a standard RT data file, with one distance being related to up to five other distances with up to 4950 distances corresponding to up to 100 cases. For the restricted randomization up to ten groups are allowed.
- RTDist for calculating distance matrices for use with RTMant and RTMant1. Input variables can be used to calculate three types of distance matrices between cases. The distances can be: (a) Euclidean using either standardized (mean=0, SD= 1) or unstandardized variable values; (b) the variable values for a case can be treated as proportions with the distances being half the sum of absolute proportion differences; or (c) the distances can represent sample differences with 0 for cases in different samples and $1/(n-1)$ for two cases in the same sample, where n is the sample size.
- RTSpace for tests on spatial data: the Mead (1974) randomization test on counts in one or more blocks of 16 quadrats and a Monte Carlo nearest neighbour test for the whether points are randomly located in a rectangular region. The Mead test is on four by four blocks of 16 quadrats. The data must be in a standard RT file in four consecutive columns. The number of cases (rows) is four times the number of blocks, with a maximum of 256 rows corresponding to a maximum of 64 blocks. The first four rows of data must be the quadrat counts for block 1. Rows five to eight then contain the quadrat counts for block 2, rows nine to 12 contain the quadrat counts for block three, and so on. In other words, the data for the different blocks are stacked one upon another in four columns of the data matrix. However, it is possible to have only one block, in which case only four rows of data are needed. If the data file that is selected for Mead's test contains a number of rows that is not a multiple of four then it is possible to do the test with the excess rows of data ignored. For example, a data set containing 14 rows of data can be analysed as consisting of three blocks, with the last two rows ignored. The second test that is allowed by the spatial data option is the Monte Carlo test described by Manly (2007, Section 10.4). This involves comparing the first ten mean nearest neighbour distances between a set of points in a rectangular area with the distributions of these mean distances that are found if the points are allocated to completely random positions within the rectangle. The random points are allocated one by one and if desired each point can have an inhibition distance surrounding it. The data for this option must be in a standard RT data file, with one column holding the X coordinates and one column holding the Y coordinates of the observed points. Each case then represents one point in the rectangular region.
- RTTS for time series tests: tests for serial correlations, the von Neumann ratio, various trend statistics, periodogram ordinates, and tests for whether the times between events are what is expected if events occur at a random choice of potential times. The data

must be in a standard RT data file with the number of cases (rows) being the number of observations on the time series. Either one or two columns must be chosen for analysis. With observations that are equally-spaced in time it is simply necessary to specify the column that holds these observations. With an irregular time series the column holding the observations and the column holding the observation times must be specified. If observations are equally-spaced in time then the observation times are taken to be 1, 2, 3, and so on. If observations are not equally-spaced in time then an equally-spaced series will be obtained by interpolation for those analyses that require this. In these cases, the total number of interpolated observations will be made the same as the initial number of observations but observation times will be spread evenly between the first and last observation. Randomization tests will then involve randomizing the original observations and repeating the interpolation rather than randomizing the interpolated values for the original data. The tests that can be carried out are described by Manly (2007, Chapter 11).

- RTMult for multivariate tests: the comparison of several samples using Wilks' lambda statistic, the sum of $\log(F)$ values, and Romesburg's sum of squares statistic E. The data must be in a standard RT data file, with one column in the file holding sample numbers. The columns (variables) to be analysed can be chosen by the user. The sum of squares statistic E depends on the scales used for variables. There is therefore a facility to standardize each of the variables to a mean of zero and a variance of one before starting the analysis. This will not affect the tests using Wilk's lambda statistic and the sum of $\log(F)$ values.
- RTChiSq for multivariate tests on count data: the comparison of several samples in terms of counts. The data consist of counts in C categories for each of R samples. The R samples are in G (two or more) groups, and randomization is used to test the hypothesis that the allocation of samples to groups was at random. The test statistic is the sum of the usual Pearson chi-squared values calculated within groups, and the significance level is the proportion of statistics for randomized data (with samples randomly allocated to groups) that are less than or equal to the observed statistic, because low chi-squared sums correspond to relative homogeneity within groups. the test is described in more detail by Manly (2007, Section 12.4).

3. Running the Programs

The RT heritage programs should run on any computer with a Windows operating system. It is important, however, that when the programs are run the data files are in the same folder as the executable files.

The output from all of the programs is recorded in the file RT.LOG. If several analyses are carried out then all the output goes in order into this file which is a text file that can be looked at and edited if necessary with any text editor.

4. Data Files

The RT programs require the data to be input as text files as indicated in Table 1. The first line contains a descriptive title. The second line contains the number of cases, followed by the number of variables. The third and subsequent lines contain the data values in the order: first variable for case 1, second variable for case 1, ..., last variable for case 1, first variable for case 2, second variable for case 2, ..., last variable for last case. All data values must be separated by new lines, spaces, commas or tabs, but are otherwise in a free format. It is probably simplest to prepare the data in a spreadsheet and copy and past into Notepad for any final editing. That should separate the data values by tabs and new lines.

Each of the programs has its own restrictions on the size of data sets that are stated when the program starts. For example, the program RT1Samp for one sample tests allows up to 50,000 data values to be input and up to 5,000 values for randomization tests.

Table 1 The standard RT data format. An example is provided on the left with comments on the right. All numbers are separated by new lines, spaces, commas or tabs. For most analyses, the maximum number of cases is 1,000 and the maximum number of variables is 50. However, these limits are reduced in some cases because of the storage required for particular calculations.

| An Example Data Set | Title for the data |
|---------------------|--|
| 10 3 | There are ten cases (rows) and three variables (columns) |
| 1 10 1 | The first row of data with values for three variables |
| 1 1 99 | . |
| 1 3 100 | . |
| 1 1 1 | . |
| 2 11 95 | . |
| 2 121 90 | . |
| 2 2 2 | . |
| 1 5 14 | . |
| 1 111 3 | . |
| 2 7 7 | The last row of data with three variables |

A number of data sets are provided with the heritage programs. These all have the text format shown in Table 1 with more description of the data provided by Manly (2007). The data files are as follows:

- Zeamays.dat is Charles Darwin's data from a paired comparison of the offspring of cross-fertilized and self-fertilized *Zea mays*. There is one variable, which is the height of cross-fertilized offspring minus the height of self-fertilized offspring, in eights of an

inch over 12 inches. There are 15 cases (paired differences). This set of data serves as test data for one sample methods using the program RT1Samp.

- Mandible.dat is data on the length of males and female mandibles for the golden jackal. There are two variables: the sample number (1 for males, 2 for females), and the lengths in mm. There are 20 cases (jackals). This set of data serves as test data for the comparison of two samples using RT2Samp.
- Lizards1.dat, which is part of the data from a study by Powell and Russell (1984, 1985) on the diet of the eastern long horned lizard *Phrynosoma douglassi brevirostre* at a site near Bow Island, Alberta. There are 10 variables: the month (1 to 4 for June to September), and gut contents in milligrams dry biomass for the nine prey categories of ants, non-formicid Hymenoptera, Homoptera, Hemiptera, Diptera, Coleoptera, Lepidoptera, Orthoptera, and Arachnida. There are 24 cases (adult male and yearling female lizards). This set of data serves as test data for one factor analysis of variance (with the consumption figures for one prey category) using RTAnova, or for multivariate analysis (with consumption figures for two or more prey categories), with month as the factor using RTMult.
- Lizards2.dat, which is also the part of Powell and Russell's (1984, 1985) data on the eastern long horned lizard. There are 24 cases (lizards), and three variables. Variable 1 is a month number (1 for June, 2 for July, 3 for August, and 4 for September); variable 2 is a size indicator (1 for adult males and yearling females, and 2 for adult females); and variable 3 is the consumption of Orthoptera in milligrams of dry biomass). These data can be used for a two factor analysis of variance with replication using RTAnova.
- E&O.dat, which holds data for ratios of numbers of Ephemeroptera (E) to numbers of Oligochaetes (O) in samples from New Zealand streams. There are 108 cases (samples) and four variables. Variable 1 is the stream number from 1 to 3; variable 2 is the position of the sample in the stream (1 for bottom, 2 for middle, and 3 for top); variable 3 is the season (1 for summer, 2 for autumn, 3 for winter, and 4 for spring); and variable 4 is the ratio E/O. These data can be used for a three factor analysis of variance with replication using RTAnova
- Editha.dat, which holds McKechnie *et al.*'s (1975) data on colonies of *Euphydryas editha*. There are 18 cases (colonies) and five variables. Variable 1 is the percentage of the hexokinase (Hk) 1.00 mobility gene in the colony; variable 2 is the reciprocal of the altitude in thousands of feet; variable 3 is the annual precipitation in inches; variable 4 is the maximum annual temperatures in Fahrenheit degrees; and variable 5 is the minimum annual temperature in Fahrenheit degrees. These data can be used for simple and multiple linear regression to account for variation in the Hk 1.00 frequency using the other variables using RTRegr.

- Earwigs.dat, which holds Popham and Manly's (1969) earwig distance matrices. There are 28 cases ('distances' between pairs of continents) and three variables. Variable 1 is the values for earwig species similarities between continents, variable 2 is the 'jump' distances between the continents in their present positions, and variable 3 is the 'jump' distances between the continents for their positions before continental drift took place. The ordering of all distances is as follows, where d_{ij} indicates the distance from continent i to continent j : d_{21} , d_{31} , d_{32} , ..., d_{86} , and d_{87} (the lower triangular part of the distance matrix, with the diagonal omitted, read row by row). These data can be used as test data for the program RTMant.
- Swede1.dat, which is counts of Swedish pine saplings. There are 16 cases (rows) and four variables (columns). The first four rows contain the counts for one four by four block of quadrats. The next four by four block of quadrats is in rows five to eight, and so on until the last block is in rows 13 to 16. Thus the counts in different four by four blocks are placed one under another in four columns of the data matrix. These data can be examined for spatial randomness using Mead's (1974) randomization test using the program RTSpace.
- Swede2.dat, which gives the positions of the Swedish pine saplings in terms of X-Y coordinates. There are 71 cases (saplings) and two variables. Variable 1 is the horizontal distance from the origin (at the bottom left-hand corner of the figure) in units of 0.625 m. Variable 2 is the vertical distance from the origin, using the same unit of measurement. These data can be tested for the spatial randomness of the saplings with a Monte Carlo method using the program RTSpace.
- Extinct.dat, which holds data on the extinction of marine genera. There are 48 cases (geologic ages) and five variables. Variable 1 is a geologic age number from 1 to 48; variable 2 is the time of the end of the geologic age in millions of years since 265 million years before the present; variable 3 is the estimated percentage of marine genera becoming extinct during the age, variable 4 are the times of mass extinctions according to Raup (1987), with 1 for a mass extinction or otherwise 0; and variable 5 is the times of mass extinctions based on deviations from a regression of logarithms of extinction rates against time, again with 1 for a mass extinction or otherwise 0. These data can be used for time series tests for serial correlation, trend and periodicity with the program RTTS.
- Broad-2b.dat, which is yearly grain yields from plot 2B of the Broadbank field at Rothamstead Experimental Station. These yields are for the years 1852-1925, as extracted from Table 5.1 of Andrews and Hertzberg (1985). There are 74 cases (years) and two variables. Variable 1 is the year (1852 to 1925), and variable 2 is the wheat yield. These data can be used for time series tests for serial correlation, trend and periodicity with the program RTTS.

- *Cepaea1.dat*, which is Cain and Sheppard's (1950) data on the snail *Cepaea nemoralis*. There are 17 cases (colonies of *C. nemoralis*) and 11 variables. Variable 1 is a habitat indicator (1 = downland beech, 2 = oakwood, 3 = mixed deciduous woods, 4 = hedgerows, 5 = downside long coarse grass, and 6 = downside short turf). Variables 2 to 11 are percentages of ten different colour and banding types in the colonies in the order YUB, YMB, YFB, YOB, PUB, PMB, PFB, BOB, BUB and BB, where the first letter of the code stands for the colour (Y = yellow; P = pink; B = brown), and the other letters stand for the banding type (UB = unbanded; MB = mid-banded; FB = fully banded; OB = other banded; B = banded). These data can be used for one factor analysis of variance on the individual shell types using the program RTAnova or for multivariate tests using RTMult to see whether there is evidence of a difference between habitats in the distribution of shell types.
- *Cepaea2.dat*, which is again based on Cain and Sheppard's (1950) *Cepaea nemoralis* data with 17 colonies and 11 variables. However in this data set the data for variables 2 to 11 are the counts for the different shell types rather than percentages. This makes this data set suitable for use with the program RTChiSq.

5. An Example

All of the options give different output. However, the example that will now be considered indicates the general form of input and output for RT. This example involves using the program RTTS to run a randomization test on the times of extinction events in the data file *Extinct.dat* that is described above.

After starting the program RTTS a data file name is requested. The response *Extinct.dat* then leads to the following display:

```
Data: Extinction of Marine Gen # Cases: 48 # Var: 5 This case: 1
-----
V 1 1.000          V 2 .0000          V 3 22.00          V 4 .0000
V 5 .0000
-----
```

Look at another case (L), choose an analysis (A) or quit (Q)?

If the response made is A at this point then the following output is provided. Comments on the responses to questions are shown in italics on the right.

Is the time series equally-spaced (y/n)? n *Not equally-spaced*

The variable that holds the time series or the 0-1 values indicating the absence or presence of an event must now be chosen.

Which variable is this? 4 *Variable 4 is to be analysed*

Which variable holds observation times? 2

Variable 2 chosen

Number of randomizations required? 1000

You must define the SKIP factor. For example SKIP = 1 means print for every randomization; SKIP = 10 means print every tenth result, etc.

SKIP factor (1-1000)? 10

Results from every 10th randomization printed

Random number seed (1-30000)? 1278

Must be an integer in the range shown

After these questions are answered a menu for choosing an analysis is provided, as shown below. For this example choice D is made, which leads to the further output that is shown.

```
#####
#                               #
#           MENU                 #
#                               #
# (A) Tests on serial correlations and the von Neumann #
#       ratio                    #
#                               #
# (B) Trend tests                #
#                               #
# (C) Periodogram to test for cycles #
#                               #
# (D) Test on times between events #
#                               #
# (E) Test using Stothers periodicity model #
#                               #
# (Q) Quit                       #
#                               #
#####
```

Choice (A,B,C,D,E or Q)? D

```
#####
# Test for cycles in the times of m events where these #
# events are constrained to occur at m out of n particular #
# times. The test statistics used are based on times #
# between events. The full distribution of these is #
# considered, and the mean and variance. #
#####
```

Press <ENTER> to continue

Number of events = 8

Order statistics for the times between events with observed series

| | | | | | | | | | | | |
|---|--------|---|--------|---|--------|---|--------|---|--------|---|--------|
| 1 | 19.000 | 2 | 26.000 | 3 | 27.000 | 4 | 27.000 | 5 | 40.000 | 6 | 50.000 |
| 7 | 53.000 | | | | | | | | | | |

Mean time between events = 34.5714
 SD of time between events = 13.1511
 Press <ENTER> to continue

At this point the randomizations begin, with some printing of test statistics for every tenth random allocation of the events to the potential times. After all the randomizations are completed the following summary tables are produced.

| Order stat. | Obs. value | Lower tail % | Upper tail % | Order stat. | Obs. value | Lower tail % | Upper tail % |
|----------------|---------------|-----------------|-----------------|----------------|---------------|-----------------|-----------------|
| 1 | 19.000 | 99.30 | 1.30 | 2 | 26.000 | 99.20 | 1.00 |
| 3 | 27.000 | 93.60 | 8.10 | 4 | 27.000 | 67.50 | 35.80 |
| 5 | 40.000 | 77.90 | 26.80 | 6 | 50.000 | 60.70 | 48.70 |
| 7 | 53.000 | 15.20 | 86.00 | | | | |

| | Observed | Lower tail % | Upper tail % |
|---------------------------|----------|-----------------|-----------------|
| Mean time between events | 34.571 | 73.20 | 29.10 |
| SD of time between events | 13.151 | 6.00 | 94.10 |

Press <ENTER> to continue.

When the enter key is pressed the program stops and the output is stored in the file RT.log. The interpretation of the results of this analysis is discussed by Manly (2007, Example 9.4) and will not be repeated here.

References

- Andrews, D.F. and Herzberg, A.M. (1985). *Data*. Springer-Verlag, New York.
- Cain, A.J. and Sheppard, P.M. (1950). Selection in the polymorphic land snail *Cepaea nemoralis*. *Heredity* 4: 275-94.
- Fisher, R.A. (1966). *The Design of Experiments*. Oliver and Boyd, Edinburgh.
- Manly, B.F.J. (2007). *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 3rd Edition. Chapman and Hall/CRC, Boca Raton.
- Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research* 27: 209-20.
- McKechnie, S.W., Ehrlich, P.R. and White, R.R. (1975). Population genetics of *Euphydryas* butterflies. I. genetic variation and the neutrality hypothesis. *Genetics* 81: 571-94.
- Mead, R. (1974). A test for spatial pattern at several scales using data from a grid of contiguous quadrats. *Biometrics* 30: 295-307.
- Popham, E.J. and Manly, B.F.J. (1969). Geographical distribution of the Dermoptera and the continental drift hypothesis. *Nature* 222: 981-2.

- Powell, G.L. and Russell, A.P. (1984). The diet of the eastern short-horned lizard (*Phrynosoma douglassi brevirostre*) in Alberta and its relationship to sexual size dimorphism. *Canadian Journal of Zoology* 62: 428-40.
- Powell, G.L. and Russell, A.P. (1985). Growth and sexual size dimorphism in Alberta populations of the eastern short-horned lizard, *Phrynosoma douglassi brevirostre*. *Canadian Journal of Zoology* 63: 139-54.
- Raup, D.M. (1987). Mass extinctions: a commentary. *Palaeontology* 30: 1-13.
- Romesburg, H.C. (1985). Exploring, confirming and randomization tests. *Computers and Geosciences* 11: 19-37.
- Stothers, R. (1979). Solar activity cycle during classical antiquity. *Astronomy and Astrophysics* 77: 121-7.
- ter Braak, C. (1992). Permutation versus bootstrap significance tests in multiple regression and ANOVA. In *Bootstrapping and Related Techniques* (eds. K.H. Jockel, K.H. Rothe and W. Sendler), pp. 79-85. Springer-Verlag, Berlin.
- Wichmann, B.A. and Hill, I.D. (1982). Algorithm AS 183: an efficient and portable pseudo-random number generator. *Applied Statistics* 31: 188-90. (Correction in *Applied Statistics* 33: 123, 1984.)