Qualification of R Univariate

Nick Lauerman

November 26, 2014

Abstract

The purpose of this suite of test is to compare the output of R 32 bit build version 3.1.2 with standard univariate test results provided by the National Inistute of Standards and Testing (NIST)

Contents

1 Setup

1.1 R

This seriess of tests is being run on R with only the "base" packages or libraries installed. The following commands are issued prior to runing hte tests for the reason stated

options(digits = 15) This is used to specify that 15 digits are to be displayed on numbers.

path Sets the path to the directory were the data sets are stored.

```
> options(digits = 15)
> path="~/R/workspace/qual/raw data/"
```

1.2 Computer Information

The degree of accuracy that can be expected from a computer if a function of several factors including the processor used. R provides a method to determine numeric tolerance based on David Goldberg (1991), âĂIJWhat Every Computer Scientist Should Know About Floating-Point ArithmeticâĂİ , ACM Computing Surveys, 23/1, 5âĂŞ48, also available via http://www.validlab.com/goldberg/paper.pdf.

This value can be treated as the error value for the computer and for accuracy beyond requires careful consideration.

```
> .Machine$double.eps ^ 0.5
```

```
[1] 1.49011611938477e-08
```

1.3 R Information

```
> version
```

```
               _
platform       i386-w64-mingw32
arch           i386
os             mingw32
system         i386, mingw32
status
major          3
minor          1.2
year           2014
month          10
day            31
svn rev        66913
language       R
version.string R version 3.1.2 (2014-10-31)
nickname       Pumpkin Helmet
```

1.4 Data Cleaning

all data sets downloaded from NIST as a DAT (ASCII Format) file were cleaned up to remove header information that was imbeded in the file. The file was than saved as a TXT file without any additional changes. This clean up was done to simplify the loading of the data into R.

2 Univariate Summary Background Information

Error Types:

We provide datasets with certified values for the mean, standard deviation, and (lag-1) autocorrelation coefficient to assess the accuracy

of Univariate Summary Statistic calculations in statistical software. Computational inaccuracy has 3 sources:

- truncation error;
- cancellation error;
- accumulation error.

Truncation error is the inexact binary representation error in storing decimal numbers according to the IEEE standard arithmetic. Of course, once these representational digits are lost, they cannot be recovered; their effect can at best be held constant, and at worst propagated to larger errors.

Cancellation error is an error that occurs when analyzing data that has low relative variation; that is, data with a high level of "stiffness". In "Assessing the Accuracy of ANOVA Calculations in Statistical Software" (Computational Statistics & Data Analysis 8 (1989), pp 325-332) Simon and Lesage noted that as the number of constant leading digits in a particular dataset increases and the data grows more nearly constant (i.e., the stiffness increases) accurate computation of standard deviations becomes increasingly difficult. This also holds for other similarly computed summary statistics, like the autocorrelation coefficient. In both cases computation is hindered by subtracting data from a mean quite close to the data, leaving behind the digits from the mantissa of each data element that are most likely to have been misrepresented.

Accumulation error (also as noted by Simon & Lesage) is the error that occurs in direct proportion to the total number of arithmetic computations, which in turn in this univariate case is proportional to the number of observations. This increases the accumulation of small errors, making accurate computations difficult.

Levels of Difficulty:

We include both generated and "real world" datasets so as to allow computational accuracy to be examined at different stiffness levels and different accumulation error levels. We have, in a fashion similar to the ANOVA datasets, drawn from the benchmark work of Simon and Lesage (1989), and have 4 "generated" data sets with the number of constant leading digits set to 7, 1, 7, and 8, respectively, and with the number of observations set to 3, 1001, 1001, and 1001, respectively. 5 "real world" datasets were borrowed from the dataset repository of the Dataplot Statistics/Graphics software system; two of these are from NIST statistical consulting, and the other 3 are "classic" general-interest sets drawn from outside NIST.

Datasets are ordered by level of difficulty (lower, average, and higher) according to their stiffness–the number of constant leading digits.

This ordering is simply meant to provide rough guidance for the user; producing correct results on a dataset of higher difficulty does not imply that your software will correctly solve all datasets of average or even lower difficulty. Of the 9 datasets, 6 (5 "real world" and 1 generated) datasets are of the lower level of difficulty, 2 (generated) are of average level of difficulty, and 1 (generated) is of higher level of difficulty.

Simple Remedial Action:

In computing general summary statistics, if you find your software giving less-than-desirable results in the calculation of the sample standard deviation, one simple remedial measure is to subtract the leading constant from all the observations in that dataset before analyzing it, and a second remedial measure is to assure yourself that the sample standard deviation is computed by the formula which first computes deviations about the mean before squaring and summing, as opposed to using the old desk calculator formula of a generation ago which involves the (computationally unstable) difference of 2 large numbers: the sums of squares of the raw data (uncentered) and the sum of the squared sample mean.

As noted in the General Background Information producing correct results for all datasets in this collection does not imply that your software will do the same for your own particular dataset. It will, however, provide some degree of assurance, in the sense that your package provides correct results for datasets known to yield incorrect results for some software

3 NumAcc4 Data Set

Using the NumAcc4.dat (http://www.itl.nist.gov/div898/strd/univ/numacc4.html) file.

Data Set Description

This generated/fabricated dataset consists of 1001 9-digit floating-point values: a single 10000000.2, followed by 500 pairings of 10000000.1 and 10000000.3. By construction, this data set has sample mean = 10000000.2 (exact); sample standard deviation = .1 (exact); and sample autocorrelation coef. = -0.999 (exact). The construction was carried out based on considerations described by Simon, Stephen D. and Lesage, James P. (1989): Assessing the Accuracy of ANOVA Caluclations in Statistical Software", Computational Statistics & data Analysis, 8, pp. 325-332.

Data set properties

Level of Difficulty Higher

Variables 1

Observations 101

First Observation 10000000.2

Expected Results (as certified)

Mean 10000000.2 (exact)

Standard Deviation 0.1 (exact)

population lag-1 autocorrelation coefficient -0.999 (exact)

```
> NumAcc4 <- read.table(file=paste0(path,"NumAcc4.txt"))
> mean(NumAcc4$V1)

[1] 10000000.2

> sd(NumAcc4$V1)

[1] 0.100000000558794

> acf(NumAcc4$V1, plot=F, lag.max=1)

Autocorrelations of series 'NumAcc4$V1', by lag

     0      1
 1.000 -0.999
```

4 NumAcc3 Data Set

Using the NumAcc3.dat (http://www.itl.nist.gov/div898/strd/univ/numacc3.html) file.

Data Set Description

This generated/fabricated dataset consists of 1001 8-digit floating-point values: a single 1000000.2, followed by 500 pairings of 1000000.1 and 1000000.3. By construction, this data set has sample mean = 1000000.2 (exact); sample standard deviation = .1 (exact); and sample autocorrelation coef. = -0.999 (exact). The construction was carried out based on considerations described by Simon, Stephen D. and Lesage, James P. (1989): Assessing the Accuracy of ANOVA Caluclations in Statistical Software", Computational Statistics & data Analysis, 8, pp. 325-332.

Data set properties

Level of Difficulty Average

Variables 1

Observations 101

First Observation 1000000.2

Expected Results (as certified)

Mean 10000000.2 (exact)

Standard Deviation 0.1 (exact)

population lag-1 autocorrelation coefficient -0.999 (exact)

```
> NumAcc3 <- read.table(file=paste0(path,"NumAcc3.txt"))
> mean(NumAcc3$V1)

[1] 1000000.2

> sd(NumAcc3$V1)

[1] 0.100000000034925

> acf(NumAcc3$V1, plot=F, lag.max=1)

Autocorrelations of series 'NumAcc3$V1', by lag

     0      1
 1.000 -0.999
```

5 NumAcc2 Data Set

Using the NumAcc2.dat (http://www.itl.nist.gov/div898/strd/univ/numacc2.html) file.

Data Set Description

This generated/fabricated dataset consists of 1001 2-digit floating-point values: a single 1.2, followed by 500 pairings of 1.1 and 1.3. By construction, this data set has sample mean = 1.2 (exact); sample standard deviation = 0.1 (exact); and sample autocorrelation coef. = -0.999 (exact). The construction was carried out based on considerations described by Simon, Stephen D. and Lesage, James P. (1989): Assessing the Accuracy of ANOVA Caluclations in Statistical Software", Computational Statistics & data Analysis, 8, pp. 325-332.

Data set properties

Level of Difficulty Average

Variables 1

Observations 1001

First Observation 1.2

Expected Results (as certified)

Mean 1.2 (exact)

Standard Deviation 0.1 (exact)

population lag-1 autocorrelation coefficient -0.999 (exact)

```
> NumAcc2 <- read.table(file=paste0(path,"NumAcc2.txt"))
> mean(NumAcc2$V1)

[1] 1.2

> sd(NumAcc2$V1)

[1] 0.1

> acf(NumAcc2$V1, plot=F, lag.max=1)

Autocorrelations of series 'NumAcc2$V1', by lag

     0      1
 1.000 -0.999
```

6 NumAcc1 Data Set

Using the NumAcc1.dat (http://www.itl.nist.gov/div898/strd/univ/numacc1.html) file.

Data Set Description

This generated/fabricated dataset consists of three 8-digit integers differing only in the least significant digit. The data set is: 10000002, 10000001, and 10000003. By construction, this data set has sample mean = 10000002 (exact); sample standard deviation = 1 (exact); and sample autocorrelation coef. = -0.5 (exact). The construction was carried out based on considerations described by Simon, Stephen D. and Lesage, James P. (1989): Assessing the Accuracy of ANOVA Caluclations in Statistical Software", Computational Statistics & data Analysis, 8, pp. 325-332

Data set properties

Level of Difficulty Lower

Variables 1

Observations 3

First Observation 10000001

Expected Results (as certified)

Mean 10000002 (exact)

Standard Deviation 1 (exact)

population lag-1 autocorrelation coefficient -0.5 (exact)

```
> NumAcc1 <- read.table(file=paste0(path,"NumAcc1.txt"))
> mean(NumAcc1$V1)

[1] 10000002

> sd(NumAcc1$V1)

[1] 1

> acf(NumAcc1$V1, plot=F, lag.max=1)

Autocorrelations of series 'NumAcc1$V1', by lag

   0    1
 1.0 -0.5
```

7 Michelso Data Set

Using Michelso.dat(http://www.itl.nist.gov/div898/strd/univ/michelso.html) file.

Data Set Description

This "real world" dataset is the result of the classic study conducted by Michelson on the speed of light in air in 1879. The response variable is speed of light (in millions of meters per second). The data was included as part of a larger study by Dorsey, Ernest N. (1944) on the velocity of light as reported in the Transactions of the American Philosophical Society.

Data set properties

Level of Difficulty Lower

Variables 1

Observations 100

First Observation 299.85

Expected Results (as certified)

Mean 299.852400000000

Standard Deviation 0.0790105478190518

population lag-1 autocorrelation coefficient 0.535199668621283

```
> Michelso <- read.table(file=paste0(path,"Michelso.txt"))
> mean(Michelso$V1)

[1] 299.8524

> sd(Michelso$V1)

[1] 0.0790105478190507

> acf(Michelso$V1, plot=F, lag.max=1)

Autocorrelations of series 'Michelso$V1', by lag

    0     1
1.000 0.535
```

8 Mavro Data Set

Using Mavro.dat(http://www.itl.nist.gov/div898/strd/univ/mavvo.html) file.

Data Set Description

This "real world" dataset is the result of a study by Radu Mavrodineaunu, a chemist at the National Institute of Standards & Technology (NIST). The purpose of the study was to determine a certified transmittance value that may be attached to the particular of filter under study. The 50 transmittance valuess were collected equi-spaced in time at a sampling rate of 10 observations per second.

Data set properties

Level of Difficulty Lower

Variables 1

Observations 50

First Observation 2.00180

Expected Results (as certified)

mean 2.00185600000000

Standard Deviation 0.000429123454003053

population lag-1 autocorrelation coefficient 0.937989183438248

```
> Mavro <- read.table(file = paste0(path,"Mavro.txt"))
> mean(Mavro$V1)

[1] 2.001856

> sd(Mavro$V1)

[1] 0.000429123454003085

> acf(Mavro$V1, plot=F, lag.max=1)

Autocorrelations of series 'Mavro$V1', by lag

    0     1
1.000 0.938
```

9 Lew Data Set

Using Lew.dat (http://www.itl.nist.gov/div898/strd/univ/lew.html) file.

Data Set Description

This "real world" dataset is the result of a study by H. S. Lew of the Structures Division of the Center for Building Technology at the National Institute of Standards & Technology (NIST). The purpose of the study was to characterize the physical behavior of steel-concrete beams under periodic load. The response variable is deflection (from a rest point) of the steel-concrete beam. The 200 observations were collected equi-spaced in time.

Data set properties

Level of Difficulty Lower

Variables 1

Observations 200

First observation -213

Expected Results (as certified)

Mean -177.435000000000

Standard Deviation 277.332168044316

population lag-1 autocorrelation coefficient -0.307304800605679

```
> Lew <- read.table(file=paste0(path,"Lew.txt"))
> mean(Lew$V1)

[1] -177.435

> sd(Lew$V1)

[1] 277.332168044316

> acf(Lew$V1, plot=F, lag.max=1)

Autocorrelations of series 'Lew$V1', by lag

     0      1
 1.000 -0.307
```

10 Lottery

Using Lottery.dat (http://www.itl.nist.gov/div898/strd/univ/lottery.html) file.

Data Set Description

This dataset consists of 218 3-digit numbers (from 000 to 999) resulting from the state of Maryland's Pick-3 Lottery. The data was collected for the 32-week period September 3, 1989 to April 14, 1990. One 3-digit random number was drawn per day, 7 days per week for most weeks, but 6 or 5 days per week for other weeks. Interesting data-analytic questions involving the dataset are 1) are the lottery numbers uniformly distributed? and 2) is there serial correlation between lottery numbers?

Data set properties

Level of Difficulty Lower

Variables 1

Observations 218

First Observation 162

Expected Results (as certified)

Mean 518.958715596330

Standard Deviation 291.699727470969

population lag-1 autocorrelation coefficient -0.120948622967393

```
> Lottery <- read.table(file=paste0(path,"Lottery.txt"))
> mean(Lottery$V1)
```

```
[1] 518.95871559633
```

```
> sd(Lottery$V1)
```

```
[1] 291.699727470969
```

```
> acf(Lottery$V1, plot=F, lag.max=1)
```

```
Autocorrelations of series 'Lottery$V1', by lag

     0      1
 1.000 -0.121
```

11 Pi Digits Data Set

Using PiDigits.dat (http://www.itl.nist.gov/div898/strd/univ/pidigits.html) file.

Data Set Description

This dataset consists of the first 5000 digits of the mathemtatical constant pi (= 3.1415926535897932384...). These 5000 digits were reported in Mathematics of Computation, January 1962, page 76. Interesting number-theoretic questions involving pi digits are 1) are the digits uniformly distributed? and 2) is there serial correlation between successive digits?

Data set properties

Level of Difficulty Lower

Variables 1

Observations 5000

All variables are all single digits.

Expected Results (as certified)

Mean 4.53480000000000

Standard Deviation 2.86733906028871

population lag-1 autocorrelation coefficient -0.00355099287237972

```
> PiDigits <- read.table(file=paste0(path,"PiDigits.txt"))
> mean(PiDigits$V1)
```

```
[1] 4.5348
```

```
> sd(PiDigits$V1)
```

```
[1] 2.86733906028871

> acf(PiDigits$V1, plot=F, lag.max=1)

Autocorrelations of series 'PiDigits$V1', by lag

     0      1
 1.000 -0.004
```