SPPCR
Version 1.0: January 2003
Estimates Allele Frequencies from Small Pool PCR Experiments

Barry W. Brown

This work supported by grants CA95567 and CA16672 from the National Cancer Institute.

The University of Texas, M.D. Anderson Cancer Center Department of Biomathematics, Box 237 1515 Holcombe Boulevard Houston, TX 77030

Contact: BWB at above address or bwb@mdanderson.org

1 Technicalities

1.1 Obtaining the Code

The source for this code (and all code written by this group) can be obtained from the following URL:

http://odin.mdacc.tmc.edu/anonftp/

2 Legalities

2.1 References

2.1.1 Incomplete Beta

DiDinato, A. R. and Morris, A. H. (1993) "Algorithm 708: Significant Digit Computation of the Incomplete Beta Function Ratios." *ACM Trans. Math. Softw. 18*,

360-373.

2.1.2 Cumulative Normal

Cody, W.D. (1993). "ALGORITHM 715: SPECFUN - A Portable FORTRAN Package of Special Function Routines and Test Drivers" *ACM Trans. Math. Softw. 19*, 22-32.

2.1.3 Inverse Normal

"Algorithm AS241" (1988) *Appl. Statist. 37*, NO. 3, 477-484.

2.1.4 Finding a Zero of a Monotone Function

Alefeld, G. E., Potra, F. A., Shi, Y. (1995) "Algorithm 748: Enclosing Zeros of Continuous Functions.", by G. E. Alefeld, F. A. Potra, YiXun Shi, *ACM Trans. Math. Softw., 21*, No. 3, 327-344

2.1.5 Base Random Number Generator

P. L'Ecuyer and S. Cote. (1991) "Implementing a Random Number Package with Splitting Facilities." *ACM Trans. on Math. Softw. 17*:1, pp 98-111.
We transliterated the Pascal of the reference to Fortran 95.

2.1.6 The Binomial Random Number Generator

Kachitvichyanukul, V. and Schmeiser, B. W. (1988) "Binomial Random Variate Generation." *Communications of the ACM, 31*: 216. (Algorithm BTPE.)

2.2 ACM Policy on Use of Code

Here is the software Policy of the ACM.

> Submittal of an algorithm for publication in one of the ACM Transactions implies that unrestricted use of the algorithm within a computer is permissible. General permission to copy and distribute the algorithm without fee is granted provided that the copies are not made or distributed for direct commercial advantage. The ACM copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Association for Computing Machinery. To copy otherwise, or to republish, requires a fee and/or specific permission.
>
> Krogh, F. (1997) "Algorithms Policy." *ACM Tran. Math. Softw. 13*, 183-186.

NO WARRANTY

WE PROVIDE ABSOLUTELY NO WARRANTY OF ANY KIND EITHER EXPRESSED OR IMPLIED, INCLUDING BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. THE ENTIRE RISK AS TO THE QUALITY AND PERFORMANCE OF THE PROGRAM IS WITH YOU. SHOULD THIS PROGRAM PROVE DEFECTIVE, YOU ASSUME THE COST OF ALL NECESSARY SERVICING, REPAIR OR CORRECTION.

IN NO EVENT SHALL THE UNIVERSITY OF TEXAS OR ANY OF ITS COMPONENT INSTITUTIONS INCLUDING M. D. ANDERSON HOSPITAL BE LIABLE TO YOU FOR DAMAGES, INCLUDING ANY LOST PROFITS, LOST MONIES, OR OTHER SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE OR INABILITY TO USE (INCLUDING BUT NOT LIMITED TO LOSS OF DATA OR DATA OR ITS ANALYSIS BEING RENDERED INACCURATE OR LOSSES SUSTAINED BY THIRD PARTIES) THE PROGRAM.

(Above NO WARRANTY modified from the GNU NO WARRANTY statement.)

2.3 Thanks to Our Supporters

The creation of this code was supported in part by the Core Grant CA11672 of the National Cancer Institute of NIH to M. D. Anderson Cancer Center. Other support was provided by the State of Texas and the Robert Herring Professorship in Cancer Research held by Dr. Brown.

3 Introduction

3.1 Caveats

- The generated answer file ¡something¿.ans will overwrite the contents of any file of the same name. The fix is to rename the file whose contents are to be saved.
- The program, if run twice on the same data, will give slightly different answers. That is because the random number generator used in obtaining the bootstrap answers is initialized through the system clock. Hence sets of random data will be generated in different runs of the program.

4 Interactive Data Entry

Here is an example of entering data from a SPPCR experiment interactively. When the program begins, it displays the following banner:

Begin Listing

```
                    SPPCR
                Version 1.0: January 2003

            Estimates Allele Frequencies from
                Small Pool PCR Experiments

                    Barry W. Brown

This work supported by grants xxxxxxxx and CA-16672 from the National
Cancer Institute.

Copyright 2003 for:
               The University of Texas, M.D. Anderson Cancer Center
               Department of Biomathematics, Box 237
               1515 Holcombe Boulevard
               Houston, TX      77030
Contact: BWB at above address or bwb@mdanderson.org
```

```
     This program may be freely  copied and distributed; it can be obtained
     from
                    http://odin.mdacc.tmc.edu/anonftp/

 Press the Return or Enter key to continue ...
<CR>
```

End Listing

The <CR> indicates that the user hit the enter key to continue the program.
Next, the program offers a choice of actions.

Begin Listing

```
           What do you want to do?

           (0) Exit this program

           (1) Read data extracted from a filemaker file.
               (See the documentation for the format)

           (2) Read data from a 'batch' input file
               (See the documentation for the format)

           (3) Interactively enter a data set.

           (4) Generate random data from parameters that you specify.

 > 3
```

End Listing

The program has four options, the first three of which are the entry of one data set in different formats. Examples of each format are provided below. The user selects option 3, the interactive entry of the data.
Here is a brief description of each of the four choices.

- **(1)** Data is entered from a file in a format peculiar to the researchers at M. D. Anderson Cancer Center. Users other than M. Siciliano and his crew are not likely to want to use this format.
- **(2)** Data is entered from a file created by an experimenter with his favorite editor.
- **(3)** Data is entered in response to prompts from the program.
- **(4)** Data is generated according to parameters specified by the user of the program. This option is usually chosen **only** by statisticians seeking to explore the methodology; it is not used by experimenters analysing their data.

Begin Listing

```
GENERAL INFORMATION ON INTERACTIVE DATA ENTRY:

RUNS: A run consists of a number of wells examined at a particular
genome equivalent amount of dna.  Different runs can have the same
amount of dna, but this is unusual.

ALLELES: Alleles are identified through integer numbers that we term
size.  The size is only the label for the allele -- it is not used in
any calculation.

You can enter information on alleles that are not seen at all in the
experiment if this makes transcription easier.  However, these alleles
are not included in the analysis.

GENERAL: Lists of numbers can be entered on several lines (the program
keeps reading until it has enough numbers then quits reading.  Numbers
themselves cannot span lines.  Thus 123 CANNOT be written as 1 at the
end of a line and 23 at the beginning of the next line.


Enter the number of runs (number of dna amounts):
Range (1 to 50)

> 4

Enter the number of alleles seen:
```

```
    Range (1 to 50)

> 10


    Enter the  integer sizes of the 10  alleles.  Each size can  be in the
    range of 1-999.

> 65  68  71  74  77  80  73  86  89  92
```

End Listing

The sizes of the alleles are merely labels identifying the alleles. We require integers for these labels because that is what is used by the scientists at MDACC. However, the numbers entered are not used in any calculations.

Begin Listing

```
    Enter the size of the two progenitor alleles.  If the subject is homozygous
    enter the size of the one progenitor allele twice.

    Enter two numbers.

> 86 92

    For  each  run (dna  amount),  enter the  dna  amount  in your  genome
    equivalent units.
     4 numbers should be entered -- several lines can be used if desired

> 2.20 1.80 1 0.4

    For each run (dna amount), enter the number of wells (replicates)
     4 numbers should be entered -- several lines can be used if desired

> 16 72 24 32
```

End Listing

Now the real data – the number of times each allele was seen in each run is entered. There are 4 runs hence 4 sets of data must be entered.
The computer prompts for each run by providing the ordinal number of the run, the DNA amount for the run as entered above, and the number of wells as previously entered. It prints the sizes of the first 20 alleles (or all of them if there were less than 20 – the usual case).

Begin Listing

```
You are entering the number of times each allele seen:
--------------------
Run number   1 dna amount (your ge)   2.20 number of wells  16
Enter 10 numbers, each between 0 and  16.
--------------------
The sizes of the first 20 alleles are printed below:
  65  68  71  74  77  80  73  86  89  92
 >  1   0   0   0   1   2   1  14   1  14


You are entering the number of times each allele seen:
--------------------
Run number   2 dna amount (your ge)   1.80 number of wells  72
Enter 10 numbers, each between 0 and  72.
--------------------
The sizes of the first 20 alleles are printed below:
  65  68  71  74  77  80  73  86  89  92
 >  0   1   2   5   1   2   2  37   2  38


You are entering the number of times each allele seen:
--------------------
Run number   3 dna amount (your ge)   1.00 number of wells  24
Enter 10 numbers, each between 0 and  24.
--------------------
The sizes of the first 20 alleles are printed below:
  65  68  71  74  77  80  73  86  89  92
 >   0   0   0   0   1   1   3  11   0   6


You are entering the number of times each allele seen:
--------------------
```

```
Run number   4 dna amount (your ge)   0.40 number of wells  32
Enter 10 numbers, each between 0 and  32.
--------------------
The sizes of the first 20 alleles are printed below:
  65  68  71  74  77  80  73  86  89  92
 >  0 0 0 0 0 0 0 0 0 1
```

End Listing

That is it! The data is entered. The program provides the following information:

Begin Listing

```
 Processing input file with the following name:
     interactive.data
 Results will be written to the screen and to file:
     interactive.ans

 Press the Return or Enter key to continue ...
```

End Listing

The input file name means nothing here since the data was entered from interactively.
The results of the analysis are written to the screen of the user and also to a file named 'interactive.ans' in the current directory of the user. **If there is a file called 'interactive.ans' before the program is run, it will be destroyed. Consequently, such a file should be renamed before running the program. This same comment holds for all methods of data entry.**

5 The Results of the Calculation.

Identical text is written to the screen and to the answer file. We display the answer file with comments.

Begin Listing

****

DATA

Number of alleles: 10
Number of runs (dna amounts):   4

Sizes of progenitor alleles: (86 , 92)

Number of Times Each Allele Seen

DNA Exp Unit	DNA G.E.	N Well	ALLELE SIZE 65	68	71	74	77	80	73	86	89	92
4.40	2.18	16	1	0	0	0	1	2	1	14	1	14
3.60	1.78	72	0	1	2	5	1	2	2	37	2	38
2.00	0.99	24	0	0	0	0	1	1	3	11	0	6
0.80	0.40	32	0	0	0	0	0	0	0	0	0	1

End Listing

The first portion of the file merely echos the data entered with one exception. The second column of the table (labelled DNA G.E.) uses the calibration information printed later to calculate the actual DNA amount in genome equivalents for each run.

Begin Listing

****

Mu[i]: Calibration Times Frequency of Allele i

Allele Size	Initial Estimate	Final Estimate	Mean Bootstrap Estimate	Asymptotic Standard Error	Bootstrap Standard Error

65	0.0037	0.0025	0.0024	0.0025	0.0024
68	0.0010	0.0025	0.0025	0.0025	0.0024
71	0.0020	0.0050	0.0051	0.0035	0.0036
74	0.0050	0.0127	0.0127	0.0057	0.0056
77	0.0100	0.0075	0.0076	0.0044	0.0044
80	0.0149	0.0127	0.0127	0.0057	0.0054
73	0.0223	0.0152	0.0155	0.0062	0.0062
86	0.2449	0.2212	0.2209	0.0288	0.0261
89	0.0056	0.0075	0.0076	0.0044	0.0044
92	0.2161	0.2077	0.2080	0.0277	0.0258

End Listing

The next portion of the answer file contains information on the $\mu_i$for each allele i. Conceptually, $\mu_i$is the average number of alleles of size $i$in each well in a run at one experimenter unit of DNA. Computationally, $\mu_i$is the calibration parameter (number of allele-equivalents per experimenter DNA unit) multiplied by the frequency of allele i.

Experimenters will generally ignore this portion of the file. It is provided primarily to allow experimenters the information needed to test the difference in different allele frequencies in a single sample.

The first column of the table shows the size (label) of each allele. The second column shows the initial estimate of μ_i. This estimate is the average over runs of the estimate for each run – there is a simple formula for the estimate for each run.

The third column (Final Estimate) shows the maximum-likelihood estimate of μ_i. This is the preferred estimate using the data from all runs together.

The fourth column (Mean Bootstrap Estimate) shows the mean estimate from 1000 random data sets generated from the data. The values in the third and fourth column should be quite similar, and they are in this example.

The fifth column (Asymptotic Standard Error) shows the standard error of the Final Estimate calculated using asymptotic formulae.

The final column (Bootstrap Standard Error) is the same standard error calculated using the random data replicates. In case of disagreement between the values of the fifth and sixth columns, the sixth column (Bootstrap) is preferred since it is calculated without the assumption that the number of wells is large.

Begin Listing

```
**********************************************************************

                        CALIBRATION

**********************************************************************

One experimenter's units is   0.4945 genome equivalents
         95% CI (   0.4182,   0.5708 )
One genome equivalent is   2.0222 experimenter's units
         95% CI (   1.7520,   2.3911 )



Estimate of calibration constant:                      0.4945
Mean bootstrap estimate of calibration constant:       0.4950

Asymptotic estimate of SE of calibration constant:     0.0420
Bootstrap estimate of SE of calibration constant:      0.0389
```

End Listing

This section of output shows the estimated relation between experimenter DNA units and the amount of amplifiable DNA in genome equivalents.

Begin Listing

```
**********************************************************************

             TOTAL FREQUENCY OF MUTANT ALLELES

**********************************************************************

                 Untransformed Frequency

Estimate:                 0.132785  Asymptotic SE: 0.025006
Mean Bootstrap Estimate:  0.133811  Bootstrap SE:  0.025431

              Arc_sin Transformed Frequency
```

```
Estimate:                  0.745968  Asymptotic SE: 0.073688
Mean Bootstrap Estimate:   0.745910  Bootstrap SE:  0.075430

95% Confidence Limits: (   0.0868,   0.1868 )
                From Transformed Bootstrap Values
```

End Listing

This section of the report shows the estimated total mutant frequency. The untransformed frequency is the raw frequency – proportion of the total. The transformed frequency shows the arcsin transform of the frequency; the arcsin transform makes the estimate more nearly normal and stabilizes the variance. The transformed frequency is **not** an estimate of the total proportion of mutant alleles – this estimate is 0.13, **not** 0.74.
The transformed frequency should be used in comparing total mutation frequency of two samples.

Begin Listing

```
**********************************************************************

**********************************************************************
```

Allele Size		Estimate of Frequency	Mean Bootstrap Estimate	Asymptotic Standard Error	Bootstrap Standard Error	Lo 95% Conf	Hi 95% Conf
65	M	0.005043	0.004905	0.005036	0.004766	0.000086	0.022755
68	M	0.005038	0.005002	0.005031	0.004953	0.000118	0.023205
71	M	0.010122	0.010229	0.007136	0.007288	0.000306	0.033492
74	M	0.025654	0.025607	0.011386	0.011176	0.007466	0.054310
77	M	0.015236	0.015425	0.008758	0.008838	0.001948	0.040783
80	M	0.025655	0.025762	0.011386	0.011099	0.007653	0.053823
73	M	0.030770	0.031403	0.012448	0.012536	0.010344	0.061579
86	P	0.447257	0.446176	0.042437	0.037947	0.373522	0.522178
89	M	0.015267	0.015477	0.008775	0.009099	0.001695	0.042069
92	P	0.419958	0.420013	0.042115	0.038076	0.346315	0.495439

```
**********************************************************************

**********************************************************************
```

End Listing

The final (unlabelled) section of the answer file gives information on each allele. The first column of the table is the size (label) of the allele. The second is either 'M' for mutant or 'P' for progenitor.
The third and fourth column are the estimate of the allele frequency and the mean bootstrap estimate of the same. The two numbers should be close in value.
The fifth and sixth columns are the asymptotic and bootstrap estimates of the standard error of the estimate of the allele frequency. Again, if the numbers differ, the bootstrap is preferred.
The final two columns show 95% confidence limits on the allele frequency.

6 Batch Data Entry

The 'batch' data format was designed to make it easy for the user to input data using his/her favorite text editor. Once the file is written, the computer dialog to process it is very simple. The only thing that the user need specify is the name of the file containing the data. Here is the dialog – the answers are the same as above.

Begin Listing

```
What do you want to do?

(0) Exit this program

(1) Read data extracted from a filemaker file.
    (See the documentation for the format)

(2) Read data from a 'batch' input file
    (See the documentation for the format)

(3) Interactively enter a data set.
```

```
          (4) Generate random data from parameters that you specify.

 > 2

     Enter the name of the file containing the data in 'batch' format.

 > Tumor2.bat
 Processing input file with the following name:
     /export/home/odin/bwb/SOFTWARE/OURS/sppcr/DOCS/Tumor2.bat
 Results will be written to the screen and to file:
     Tumor2.ans

 Press the Return or Enter key to continue ...
```

End Listing

Here is a listing of the contents of file 'Tumor2.bat'. It's format is discussed below.

Begin Listing

```
NALLELE 10

NRUN 4

NWELL 16 72 24 32

ALLELESIZES 65 68 71 74 77 80 73 86 89 92

PROGENITOR  86 92

RUN 1.1 1 0 0 0 1 2 1 14 1 14
RUN 0.9 0 1 2 5 1 2 2 37 2 38
RUN 0.5 0 0 0 0 1 1 3 11 0  6
RUN 0.2 0 0 0 0 0 0 0  0 0  1
```

End Listing

Each line of input is introduced by a character string that identifies it. Following this identifier is an ordered set of numbers that constitute the data; the numbers are separated from the identifier and from each other by spaces.
The identifiers and corresponding information must appear in the order shown. The identifiers themselves can be in upper or lower case. Lists of numbers can span several lines, but a single number cannot be split over lines. (For example, the characters '12' endling a line followed by the characters '34' beginning the next line will be interpreted as two numbers: 12 and 34.)
The items below must occur in the specified order.

1. NALLELE This keyword is followed by a single integer: the total number of different alleles seen.

2. NRUN This keyword is followed by a single integer: the number of DNA amounts examined.

   A single run by this definition can be broken into several runs for data entry – it makes no difference to the results. Example: a single run with DNA amount 2 (in experimenter units), the number of wells 20, and a single allele seen 5 times is equivalent to two runs with DNA amount 2, the number of wells 10 and 10 and the number of times the allele is seen being 3 and 2.

3. NWELL is followed by NRUN numbers. This entry specifies the number of wells (replicates) for each run.

4. ALLELESIZES is followed by NALLELE integers which are the labels for the alleles.

5. PROGENITOR is followed by two integers which must be on the ALLELESIZES list. These are the labels of the progenitor alleles. If the subject is homozygous for the progenitor allele, the single size should be entered twice.

6. RUN is followed by NALLELE+1 numberic entries. The first entry is the DNA amount in experimenter units; this entry can contain a decimal point. The other entries are integers specifying the number of times each allele is seen out the NWELL replicates. There must be NRUN entries starting with the keyword RUN.

7 Filemaker File Data Entry

This form of data entry was devised to allow local M. D. Anderson investigators to analyse data using Filemaker files in which they keep the data. We make no claim that this form will be found useful by others who might want to modify the program to read data in a format more useful to their particular system.
The dialog is quite similar to that for batch data entry:

Begin Listing

```
        What do you want to do?

        (0) Exit this program

        (1) Read data extracted from a filemaker file.
            (See the documentation for the format)

        (2) Read data from a 'batch' input file
            (See the documentation for the format)

        (3) Interactively enter a data set.

        (4) Generate random data from parameters that you specify.

> 1

    Enter the  name of  the  file  containing the data  in   the
    standard format for this analysis.

> Tumor2
```

End Listing

The user specifies that data is to be read in filemaker format and provides the name of the file containing the data. Here is a listing of the contents of the file, Tumor2:

Begin Listing

```
"86","92","16","1.10","65","1","68","0","71","0","74","0",<CONT>

"77","1","80","2","83","1","86","14","89","1","92","14","95","0"
"86","92","72","0.90","65","0","68","1","71","2","74","5",<CONT>

"77","1","80","2","83","2","86","37","89","2","92","38","95","0"
```

```
"86","92","24","0.50","65","0","68","0","71","0","74","0",<CONT>

"77","1","80","1","83","3","86","11","89","0","92","6","95","0"
"86","92","32","0.20","65","0","68","0","71","0","74","0",<CONT>

"77","0","80","0","83","0","86","0","89","0","92","1","95","0"
```

End Listing

There are four lines to the file; <CONT> indicates that the line was continued for convenience in printing. Each line contains the information for one run.
Each line contains numbers in quotation marks; the fields are separated by commas. The program replaces the quotation marks and commas by blanks, leaving blank delimited numbers.
The first two numbers on each line are the progenitor allele sizes. The next (third) number is the number of wells for the current run, i.e., NWELL. The next (fourth) number is the DNA amount in experimenter units,
Succeeding numbers come in pairs. The first number of each pair is the label of an allele; the second number of the pair is the number of wells in which it was seen.

8 Generating Random Data

This option is used primary for the statistician who is investigating the properties of this statistical analysis. Instead of analysing actual data, the answers are entered, and data sets are randomly generated and analysed.

Begin Listing

```
What do you want to do?

(0) Exit this program

(1) Read data extracted from a filemaker file.
    (See the documentation for the format)

(2) Read data from a 'batch' input file
    (See the documentation for the format)

(3) Interactively enter a data set.
```

```
          (4) Generate random data from parameters that you specify.

 > 4
     Enter (1) The number of dna amounts to be simulated
           (2) The number of alleles to be simulated
3 5

     This program does not allow  the number of wells (replicates) per
     run to vary from run to run.

     Enter the number of wells per run for the simulated data.

 > 20
 Enter the frequencies of the   5 alleles.
  The frequencies will be normalized to add to one.

0.4 0.1 0.4 0.05 0.05

     The is one DNA amount per run and you have spocified   3 runs for this
     experiment.

     Enter   3 DNA amounts, one for each run.
 > 2 4 6

     The  calibration  parameter  is  the  number  of  g.e.'s  in  the
     experimenter's DNA unit.

     Enter the calibration parameter for the simulated data.

 > 1

     Enter the indices (ordinal number) of the two progenitor alleles.
     If  the simulated  sample  is homozygous,  enter  the same  index
     twice.

 > 1 3
```

End Listing

The first and third alleles are the (simulated) progenitor alleles.

Begin Listing

```
    Do you  want bootstrapping to  use the values you  entered rather
    than those from the first data set generated?

    Enter 'y' for this option else 'n'.

Please enter one of [yn]: > y
```

End Listing

If you choose 'no' for this option, the following action occurs. One random data set is generated from the parameters entered. The bootstrap estimates are generated from this data set just as if it were entered as data.
If you choose 'yes', the bootstrap estimates are generated from the original parameters that you entered and not from the first data set generated.

Begin Listing

```
    Do  you want  the  simulated values  (mu, calibration,  frequency
    estimation) written to a file for further processing?

    Enter 'y' for this option else 'n'.

    This  option is  usually  used only  by  those investigating  the
    properties of  the statistical method  -- not by  those analysing
    data.

Please enter one of [yn]: > y
```

End Listing

This option allows all of the simulated (bootstrap) answers to be written to a file for analysis by a different program.

Begin Listing

```
    ********************************************************************

                PARAMETERS ENTERED FOR GENERATING DATA

    ********************************************************************

 The number of runs (dna amounts) simulated is:   3

 The number of alleles simulated is:   5


 The number of wells per run is the same for all runs.
 The common  number of wells is:    20


                      Allele Frequencies Used to Generate the Data

(  1) 0.400000  (  2) 0.100000  (  3) 0.400000  (  4) 0.050000  (  5) 0.050000


 The calibration factor is the number of genome equivalents
 in one experimenter's unit.

 The calibration factor used to generate the data is:    1.000

 The indices of the progenitor alleles are:    1   3



 Processing input file with the following name:
     generated.data
 Results will be written to the screen and to file:
     generated.ans
 Simulation results will be written to the file:
     generated.sim

 Press the Return or Enter key to continue ...
```

End Listing

The program echoes the parameters entered for data generation. It then lists the relevent files.
The input file name – 'generated.data' is a red herring. The program neither reads nor writes a file of that name.
The results are written to a file, 'generated.ans'. Any previous contents of this file will be overwritten.
The random data generated for the bootstrap is written to a file, 'generated.sim'. Any previous contents of this file will be overwritten.
The ans file has exactly the same format as the one described above. It is listed here without comment.

Begin Listing

```
     ************************************************************************

                 PARAMETERS ENTERED FOR GENERATING DATA

     ************************************************************************

 The number of runs (dna amounts) simulated is:   3

 The number of alleles simulated is:   5


 The number of wells per run is the same for all runs.
 The common  number of wells is:    20


                      Allele Frequencies Used to Generate the Data


(  1) 0.400000  (  2) 0.100000  (  3) 0.400000  (  4) 0.050000  (  5) 0.050000


 The calibration factor is the number of genome equivalents
 in one experimenter's unit.

 The calibration factor used to generate the data is:    1.000

 The indices of the progenitor alleles are:    1   3
```

****

DATA

Number of alleles: 5
Number of runs (dna amounts): 3

Sizes of progenitor alleles: (1 , 3)

Number of Times Each Allele Seen

DNA Exp Unit	DNA G.E.	N Well	ALLELE SIZE 1	2	3	4	5
2.00	2.23	20	13	1	13	2	3
4.00	4.47	20	18	5	16	6	1
6.00	6.70	20	16	11	20	8	4

****

Mu[i]: Calibration Times Frequency of Allele i

Allele Size	Initial Estimate	Final Estimate	Mean Bootstrap Estimate	Asymptotic Standard Error	Bootstrap Standard Error
1	0.4563	0.4135	0.4078	0.0684	0.0711
2	0.0769	0.0883	0.0999	0.0216	0.0221
3	0.8093	0.4989	0.4113	0.0868	0.0714
4	0.0757	0.0802	0.0499	0.0202	0.0166
5	0.0438	0.0360	0.0491	0.0127	0.0152

****

CALIBRATION

```
**********************************************************************

One experimenter's units is   1.1169 genome equivalents
         95% CI (   0.9109,   1.3230 )
One genome equivalent is   0.8953 experimenter's units
         95% CI (   0.7558,   1.0979 )



Estimate of calibration constant:                        1.1169
Mean bootstrap estimate of calibration constant:         1.0179

Asymptotic estimate of SE of calibration constant:       0.1151
Bootstrap estimate of SE of calibration constant:        0.1052



**********************************************************************

             TOTAL FREQUENCY OF MUTANT ALLELES

**********************************************************************

                 Untransformed Frequency

Estimate:                  0.183109  Asymptotic SE: 0.029714
Mean Bootstrap Estimate:   0.196411  Bootstrap SE:  0.031654

              Arc_sin Transformed Frequency

Estimate:                  0.884365  Asymptotic SE: 0.076829
Mean Bootstrap Estimate:   0.915827  Bootstrap SE:  0.080134

95% Confidence Limits: (   0.1265,   0.2475 )
                From Transformed Bootstrap Values

**********************************************************************


**********************************************************************
```

```
Allele    Estimate  Mean      Asymptotic  Bootstrap Lo 95%   Hi 95%
Size      of        Bootstrap  Standard    Standard  Conf     Conf
          Frequency Estimate  Error       Error

    1  P  0.370228  0.400049  0.049299    0.051371  0.274037 0.471926
    2  M  0.079073  0.098688  0.019538    0.022187  0.043963 0.123336
    3  P  0.446663  0.403541  0.052559    0.051320  0.346025 0.549559
    4  M  0.071844  0.049244  0.018295    0.016222  0.038352 0.114760
    5  M  0.032193  0.048479  0.011511    0.015021  0.012332 0.060980

**********************************************************************


**********************************************************************
```

End Listing

The file 'generated.sim contains 3000 lines which are composed of 3 lines per simulation repeated for 1000 simulations. Here are the first few lines of the file.

Begin Listing

```
Cal      1.0195
Mu     0.3446  0.1269  0.4545  0.0517  0.0418
Freq  0.3380 0.1245 0.4458 0.0507 0.0410
Cal      0.8381
Mu     0.3145  0.0785  0.3433  0.0458  0.0561
Freq  0.3752 0.0936 0.4096 0.0547 0.0669
```

End Listing

The first of the three lines is identified by 'Cal'. The number following is the estimated calibration factor for the simulated data.
The second of the three lines contains the identifier 'Mu'. The values following are the estimates of the mu's for the simulated data (one per allele).
The final line is identified by 'Freq' and contains the estimated frequencies of the alleles in the simulated data.