"monte": When is n Sufficiently Large?

Jeffrey H. Gove*
Research Forester
USDA Forest Service
Northern Research Station
271 Mast Road
Durham, New Hampshire 03824 USA
e-mail: jgove@fs.fed.us or e-mail: jhgove@unh.edu

Friday 16th March, 2012
3:43pm

Contents

1 Introduction

In designing surveys or even the sampling methods themselves, one is often presented with the question of how many samples should be taken? This is a function of the variability of the population and the time it takes to execute the respective sampling technique on the population. Barrett and Goldsmith (1976) addressed the question "When is n sufficiently large" in their Monte Carlo analysis, based on the Central Limit Theorem (CLT). Given a population of individuals, drawing repeated samples in the Monte Carlo sense and invoking the central limit theorem is certainly an instructive way to address this question. In what follows, we use this concept to look at sample size issues for populations of sample (grid) points in sampling surface "sampSurf" objects. More information on this approach can be found, for example in Barrett and Nutt (1979).

*Phone: (603) 868-7667; Fax: (603) 868-7604.

The "monte" concept is really more of a subpackage within `sampSurf` and can be used in its current form for more general populations than those from "sampSurf" objects. Therefore, eventually, the classes detailed here will probably be separated from the `sampSurf` package and be located in a more general package, with perhaps only "sampSurf" extensions residing in `sampSurf`. In other words, support will always be available for "sampSurf" objects, but don't count on the code residing within the package name space (other than as an imported required package for `sampSurf`) in the future.

2 "monte" Class Structure Overview

There are several classes associated with the Monte Carlo CLT simulations implemented here. These are...

- "montePop:" This stores the population that the MC samples will be drawn from for confidence interval estimation.
- "monteSample:" This class holds the means, confidence intervals and other information from the individual MC samples drawn from the population. There are subclasses for normal theory ("monteNTSample") and bootstrap ("monteBSSample") methods.
- "monte:" The class the keeps tract of objects of the above classes for a give set of MC runs. There can be normal theory or bootstrap components, or both.

In what follows, each class will be detailed, including the class structures and generics used to create objects of the respective classes.

3 The "montePop" Class

Each object of class "monte" must have a population of values that form the basis for repeated sampling in the MC sense. The "montePop" class holds the information necessary in the context of "monte" objects.

```
R> showClass('montePop')
```

```
Class "montePop" [package "sampSurf"]

Slots:

Name:          mean          var         stDev             N         total
```

```
Class:       numeric       numeric       numeric       numeric       numeric

Name:        popVals zeroTruncated             n           fpc       varMean
Class:       numeric       logical       numeric       numeric       numeric

Name:          stErr   description
Class:       numeric     character
```

- *mean*: The population mean: $\mu = \frac{1}{N}\sum_{i=1}^{N} y_i$.
- *var*: The population variance: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)^2$.
- *stDev*: The population standard deviation: $\sigma = \sqrt{\sigma^2}$
- N: The number of observations in the population—the population size.
- *total*: The total for whatever attribute the population is concerned with: $\tau_x = \sum_{i=1}^{N} y_i$.
- *popVals*: A numeric vector containing the values, $y_i, i = 1, \ldots, N$, in the population.
- *zeroTruncated*: A logical scalar: `TRUE` if the population is zero-truncated (no zeros); `FALSE` otherwise. This was added to deal with "sampSurf" objects, which can be largely zero-inflated in all of the background grid cells where no inclusion zones exist, depending on how the tract size, population size, sampling design, etc., are chosen.

  Note: The following four slots can either contain the contents as described below, or be `NA` if no sample size information was provided...

- n: A numeric vector listing the different sample sizes that will be drawn from this object. If we will be drawning samples of size $n = 10, 20, 30$, then this would hold `c(10,20,30)`.
- *fpc*: The finite population correction factors for each sample size `n`. The correction is: $f_c = (N - n)/N$.
- *varMean*: The population variance of the mean, which is sample size dependent; *viz.*, $\sigma^2_{\bar{y}} = \frac{\sigma^2}{n} \times f_c$.
- *stErr*: The population standard error of the mean: $\sigma_{\bar{y}} = \sqrt{\sigma^2_{\bar{y}}}$.
- *description*: Some descriptive text about the object.

Note that each object of class "montePop" is designed to hold one, and only one population.

3.1 Object creation

There is a constructor generic to create objects of class "montePop". While it is simple enough to create objects using `new`, it is recommended to use the constructor to minimize the chances of creating an invalid object. The constructor has the same name as the class and can create an object from an **R** numeric vector, or from a "sampSurf" object...

```
R> x = rnorm(100)
R> x.mp = montePop(x)
R> x.mp
```

```
Population...
  Mean = 0.032465716
  Variance = 1.298729
  Standard Deviation = 1.1396179
  Total = 3.2465716
  Size (N) = 100
  Zero-truncated = FALSE
```

The result of the above run is a "montePop" population object with the slots defined above assigned when the signature object of the constructor is of class "numeric" (vector).

In the following, we generate a population object from a "sampSurf" object...

```
R> smTract = Tract(c(x=30,y=30), cellSize=0.5)
R> smbuffTr = bufferedTract(8,smTract)
R> agauge = angleGauge(6)
R> SS.hps = sampSurf(10, smbuffTr, 'horizontalPointIZ', angleGauge=agauge,
+                    estimate='volume')
```

```
Number of trees in collection = 10
Heaping tree: 1,2,3,4,5,6,7,8,9,10,
```

```
R> (hps.pop = montePop(SS.hps, zeroTruncate = TRUE, n = c(10,20,30)))
```

```
Population...
  Mean = 13.878831
  Variance = 51.840625
```

```
  Standard Deviation = 7.2000434
  Total = 19583.031
  Size (N) = 1411
  Zero-truncated = TRUE
  Sample sizes (n) = 10, 20, 30
  Finite population corrections = 0.9929, 0.9858, 0.9787
  Variance of the mean = 5.1473222, 2.5552909, 1.6912805
  Standard error of the mean = 2.2687711, 1.5985277, 1.3004924
```

In this example we have specified that samples of size $n = (10, 20, 3)$ will eventually be drawn from the population. This prompts the constructor to calculate the finite population correction, population variance of the mean and standard error of the mean associated with these sample sizes as is demonstrated in the summary of the object. This option is not one that would probably be used on its own just to create a "montePop" object, as it would limit the eventual use of the contents to these sample sizes. Where it becomes very useful is in the "monte" object construction, where the sample sizes are an intrinsic component of a given Monte Carlo experiment.

One very important point to keep in mind when using the zero-truncated population and subsequent MC sampling routine is the following. The sampling surface method takes the mean of the surface estimates *including the background cells*, which have zero value, to compute the estimate. When we truncate the zeros in the background, we now will have a population mean that will be larger—perhaps substantially so, depending upon the inclusion zone coverage—than the unbiased estimate we get from running the `sampSurf` method above. For example, the horizontal point sampling population we just created has the following statistics (see also Figure 1)...

```
R> summary(SS.hps)


Object of class: sampSurf
------------------------------------------------------------
sampling surface object
------------------------------------------------------------

Inclusion zone objects: horizontalPointIZ
Measurement units = metric
Number of trees = 10
True tree volume = 5.4377694 cubic meters
True tree basal area = 0.58541245 square meters
True tree surface area = 81.218394 square meters
True tree biomass = NA
True tree carbon = NA

Estimate attribute: volume
```

```
Surface statistics...
  mean = 5.4397308
  bias = 0.0019613078
  bias percent = 0.036068242
  sum = 19583.031
  var = 66.229076
  st. dev. = 8.1381248
  cv % = 149.60529
  surface max = 35.502872
  total # grid cells = 3600
  grid cell resolution (x & y) = 0.5 meters
  # of background cells (zero) = 2189
  # of inclusion zone cells = 1411
```

Note that the standing tree volume on the tract is 5.438 m^3, and the unbiased estimate given by the mean over all grid cells of the sampling surface is 5.44. However, we see in the "montePop" object that the population mean is 13.88. The difference is obvious, and the source of the difference should now be apparent—though it probably was initially without this lengthy example. Finally, note that the size of the population in the "montePop" object is $N = 1411$, which is the number of cells covered by inclusion zones in the above summary.

3.2 Plotting the object

Currently only histograms are supported for "montePop" objects. The command is...

```
R> hist(hps.pop)
```

4 The "monteSample" Class

This is a general class for holding information on the MC samples drawn from a "montePop" population object. It is a virtual class; therefore, you must use one of the subclasses that have been tailored to normal theory or bootstrap resampling described below—or create your own subclass for a new application.

```
R> showClass('monteSample')
```

```
Virtual Class "monteSample" [package "sampSurf"]
```

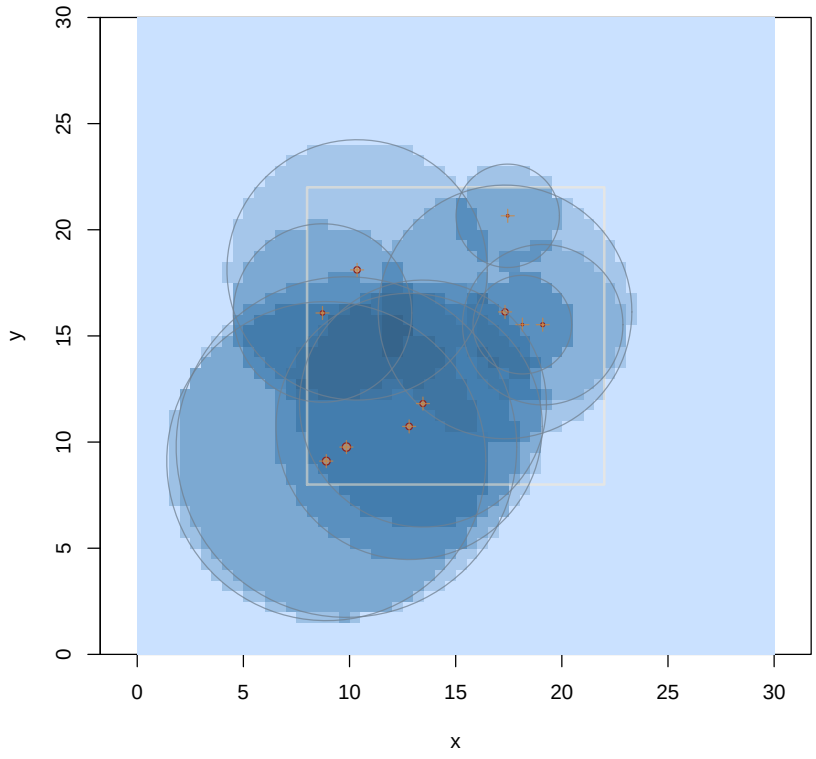

Figure 1: The horizontal point sampling "sampSurf" representation.

```
Slots:

Name:   mcSamples           n      alpha    replace    ranSeed        fpc
Class:    numeric    numeric    numeric    logical    numeric    numeric

Name:       means       vars     stDevs   varMeans     stErrs   lowerCIs
Class: data.frame data.frame data.frame data.frame data.frame data.frame

Name:    upperCIs     caught  caughtPct      stats
Class: data.frame data.frame    numeric data.frame

Known Subclasses: "monteNTSample", "monteBSSample"
```

- *mcSamples*: A scalar numeric specifying the number of Monte Carlo samples drawn from the population.

- *n*: A numeric vector listing the different sample sizes recorded in the object that have been drawn from a "montePop" population object. So, if we have drawn samples of size $n = 10, 20, 30$, then this would hold `c(10,20,30)`. The associated names should always be of the form `c('n.10','n.20','n.30')`.
- *alpha*: The two-tailed alpha level for which confidence intervals have been calculated. I.e., for the 95% confidence level ($\alpha = 0.05$) `alpha = 0.05`.
- *replace*: `TRUE` if the samples have been drawn from the population with replacement, `FALSE` otherwise.
- *ranSeed*: The random number seed as a numeric vector. Please see the **R** documentation on `.Random.seed` for information on the format of this slot. Note that it is *not* a simple scalar integer "seed", but a vector of integers containing the state of the random number generator at the beginning of the simulations.
- *fpc*: The finite population correction factors for each sample size `n`. The correction is: $f_c = (N - n)/N$.
- *means*: A data frame with `mcSamples` rows, and one column for each of the sample sizes in the `n` slot of the object. What is stored here depends on the subclass object type, so please see the definitions below for these slots.

  Note: The next six slots have the same dimensions as the `means` slot.

- *vars*: Contains the individual sample variances for each sample for both "monteNTSample" and "monteBSSample" subclasses: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$.
- *stDevs*: Contains the individual sample standard deviations for each sample for both "monteNTSample" and "monteBSSample" subclasses: $s = \sqrt{s^2}$.
- *varMeans*: Contains the individual sample variance of the mean for each sample for both "monteNTSample" and "monteBSSample" subclasses: $s^2_{\bar{y}} = \frac{s^2}{n} \times f_c$.
- *stErrs*: Contains the individual standard errors for each sample for both "monteNTSample" and "monteBSSample" subclasses: $s_{\bar{y}} = \sqrt{s^2_{\bar{y}}}$.
- *lowerCIs*: Contains the individual lower limit for the confidence intervals. This is defined differently for the "monteNTSample" and "monteBSSample" subclasses.
- *upperCIs*: Contains the individual upper limit for the confidence intervals. This is defined differently for the "monteNTSample" and "monteBSSample" subclasses.
- *caught*: Contains a flag where `TRUE` means the confidence interval caught the population mean and `FALSE` means it failed to catch the population mean. Taking column sums, therefore (since `TRUE == 1` and `FALSE == 0`) will give the number of intervals that caught the population mean for each sample size. This is used to calculate the next slot below.
- *caughtPct*: The percentage of times the confidence intervals caught the population mean as calculated from the data frame in the `caught` slot of the object.

- *stats*: A summary data frame with rows as the *average* of each column (i.e., over all MC samples) from the information in the data frames above (`means`, `vars`, `stDevs`, `varMeans`, `stErrs`, `lowerCIs`, and `upperCIs`). The interpretation of some of the rows depends on the subclass object as has been mentioned above.

4.1 The "monteNTSample" class

This class holds information for classic normal theory confidence intervals under simple random sampling. It adds only one slot to the "monteSample" superclass, `t.values`. Some of the other slot definitions that depend on the type of intervals are also defined below for this class.

- *t.values*: The $t_{n-1}^{1-\alpha/2}$ Student's t values for each sample size `n` with two-tailed α-level `alpha`.
- *means*: The data frame contains the individual means for all `mcSamples` by `length(n)` samples drawn from the population. Taking column means gives the overall mean for each of the sample sizes. The sample mean is: $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.
- *lowerCIs*: This is the usual normal theory lower limit for each sample: $\bar{y} - t_{n-1}^{1-\alpha/2} s_{\bar{y}}$, where t is Student's t-value and $s_{\bar{y}}$ is the standard error of the mean for the sample.
- *upperCIs*: This is the usual normal theory upper limit for each sample: $\bar{y} + t_{n-1}^{1-\alpha/2} s_{\bar{y}}$, where t is Student's t-value and $s_{\bar{y}}$ is the standard error of the mean for the sample.

The `stats` slot averages over all Monte Carlo samples and therefore has the usual interpretation for each row in the data frame.

4.1.1 Object creation

The constructor for objects of this class has the same name as the class. Below we create an object from a sampling surface for several sample sizes, n, and a small number of MC samples for illustration.

Please note that while we can create objects in the following manner, it is preferable to use the method `monte` to do so as discussed later. The reason for this is that objects of class "monte" store everything needed to reconstruct the samples, including the population they came from, which is not present in this class of MC results.

```
R> hps.nts = monteNTSample(hps.pop, n = c(10,20,30), mcSamples=100)
R> hps.nts
```

```
Number of Monte Carlo samples = 100
Sample sizes: n = 10, 20, 30
Sample summary statistics (mean values)...
              n.10       n.20       n.30
mean    13.6193251 13.7816539 13.8006514
var     51.9325699 52.3250914 51.3714150
stDev    7.0122578  7.1518746  7.1113143
VarMean  5.1564515  2.5791709  1.6759727
stErr    2.2095988  1.5878335  1.2844659
lowerCI  8.6208653 10.4582803 11.1736236
upperCI 18.6177850 17.1050276 16.4276791

Percentage of confidence intervals (95%) that caught the population mean...
n.10 n.20 n.30
  93   94   93
```

4.2 The "monteBSSample" class

This subclass of "monteSample" handles bootstrap confidence intervals in the MC setting. At present, only bootstrap "BCa" confidence intervals are calculated. The general idea is as follows. First, draw a sample from the population of interest, just like in the normal theory case. Then run R bootstrap resample replicates and calculate the mean and confidence interval endpoints from the distribution of bootstrap means of the replicates. This is repeated for each MC sample and for each sample size. The bootstrap intervals provide a nonparametric alternative to the normal theory intervals and may be more valid when the distribution of sample means is non-Gaussian. Thus, the bootstrap is really a nested—or second-stage—set of Monte Carlo iterations for each first-stage MC iteration and sample size.

The class adds two new slots to the "monteSample" superclass as shown below. In addition, a few of the other slot definitions that depend on the type of intervals are also defined below for this class.

- *degenerate*: It may happen that, especially for small n, some of the samples drawn from the population can be degenerate (all the same value). When this happens, all of the bootstrap resamples will also be degenerate, and confidence interval estimation is impossible since it is based on the distribution of the bootstrap sample means. This slot is a numeric vector with the number of degenerate samples for each sample size in the `n` slot of the object.
- R: The number of bootstrap sample replications.
- *means*: The data frame contains the overall bootstrap sample means for each of the `mcSamples` by `length(n)` samples drawn from the population. The overall bootstrap sample mean is

defined here as the mean of the R individual (second-stage) bootstrap sample means for each case. Taking column means gives the overall mean for each of the sample sizes. *Note:* formula.

- *lowerCIs*: This is the lower "BCa" confidence interval endpoint for the $1 - \alpha/2$ confidence level. It is calculated from the distribution of bootstrap sample means that is created in bootstrap sampling for each MC sample and sample size, n.
- *upperCIs*: This is the upper "BCa" confidence interval endpoint for the $1 - \alpha/2$ confidence level. It is calculated from the distribution of bootstrap sample means that is created in bootstrap sampling for each MC sample and sample size, n.

The `stats` slot again averages over all Monte Carlo samples in each column of the data frames as defined above. Note, however, that only the `means`, `lowerCIs` and `upperCIs` have a meaning that differs from those of class "monteNTSample". Therefore, the other rows in `stats` contain the usual Monte Carlo averages, not Monte Carlo averages based on bootstrapping results.

4.2.1 Object creation

As with the normal theory subclass, objects can be generated with a constructor of the same name. It is, however, preferable to use the `monte` constructor in general.

```
R> hps.bss = monteBSSample(hps.pop, n = c(10,20,30), mcSamples=100, R=50)
R> hps.bss
```

```
Number of bootstrap samples = 50
Number of Monte Carlo samples = 100
Sample sizes: n = 10, 20, 30
Sample summary statistics (mean values)...
              n.10       n.20       n.30
mean    13.6373037 13.8521311 14.0440540
var     51.9438373 52.1066343 54.0596715
stDev    6.9716304  7.1218430  7.2921056
varMean  5.1575702  2.5684028  1.7636760
stErr    2.1967970  1.5811659  1.3171209
lowerCI  9.9890499 11.0119686 11.6412114
upperCI 18.5040527 17.3315162 16.8828345

Percentage of confidence intervals (95%) that caught the population mean...
n.10 n.20 n.30
  88   94   93
```

The example uses a much smaller number of bootstrap iterations than is normally recommended in practice, just for illustration here.

4.2.2 Plotting the object

Currently only histograms are supported for "monteSample" subclass objects. The command is, e.g.,...

```
R> hist(hps.bss)
```

The histograms will be illustrated in the next section using objects of class "monte".

5 The "monte" Class

We have described all of the component classes that go into a possible set of Monte Carlo samples from a population of interest. The "monte" class combines the above class structures into slots in its structure. The constructor for class "monte" constructs the individual objects for the classes discussed above and then constructs the "monte" object.

The class structure is shown as follows...

```
R> showClass('monte')
```

```
Class "monte" [package "sampSurf"]

Slots:

Name:                  pop            estimate           NTsamples
Class:           montePop           character monteNTSampleOrNULL

Name:            BSsamples         description
Class: monteBSSampleOrNULL           character
```

By now, most of what follows should be self-explanatory.

- *pop*: An object of class "montePop".

- *estimate*: In the case of "sampSurf" objects, this is the attribute for which the surface has been estimated.
- *NTsamples*: An object of class "monteNTSample", or `NULL` if non-existent.
- *BSsamples*: An object of class "monteBSSample", or `NULL` if non-existent.
- *description*: Some descriptive text about the object.

One thing to note is that the object can contain either normal theory or bootstrap information or both. This will be illustrated below in the constructor method.

5.1 Object creation

Currently there are three methods for constructing objects of class "monte". The signature argument in each case wants a "population" specification from which to draw the repeated samples. The signature argument can be either a "numeric" vector, a "montePop" population object, or a "sampSurf" object. In all cases, the method for the "montePop" object is ultimately called to do the work, the other two are just wrappers to generate "montePop" objects from the population that was specified in the signature argument. Here we demonstrate the method that takes a signature argument of class "sampSurf". Information on other arguments shown below and available in general in the `monte` generic is found in the help pages (methods?monte).

```
R> mo = monte(SS.hps, zeroTruncate = TRUE, n = c(10, 20, 30, 50), mcSamples=200,
+             R=120)
R> mo


Estimate attribute = volume

Population...
  Mean = 13.878831
  Variance = 51.840625
  Standard Deviation = 7.2000434
  Total = 19583.031
  Size (N) = 1411
  Zero-truncated = TRUE
  Sample sizes (n) = 10, 20, 30, 50
  Finite population corrections = 0.9929, 0.9858, 0.9787, 0.9646
  Variance of the mean = 5.1473222, 2.5552909, 1.6912805, 1.0000722
  Standard error of the mean = 2.2687711, 1.5985277, 1.3004924, 1.0000361

Normal theory results...
```

```
Number of Monte Carlo samples = 200
Sample sizes: n = 10, 20, 30, 50
Sample summary statistics (mean values)...
              n.10       n.20       n.30       n.50
mean    14.0033580 13.8703030 13.7394658 13.9433716
var     53.8024032 50.3734979 50.1580482 53.0612797
stDev    7.1188799  7.0060921  7.0200081  7.2468975
VarMean  5.3421096  2.4829743  1.6363871  1.0236202
stErr    2.2431960  1.5554674  1.2679739  1.0065438
lowerCI  8.9288961 10.6146724 11.1461680 11.9206461
upperCI 19.0778200 17.1259336 16.3327637 15.9660971

Percentage of confidence intervals (95%) that caught the population mean...
n.10 n.20 n.30 n.50
94.0 93.0 92.5 94.5

Bootstrap results...
Number of bootstrap samples = 120
Number of Monte Carlo samples = 200
Sample sizes: n = 10, 20, 30, 50
Sample summary statistics (mean values)...
              n.10       n.20       n.30        n.50
mean    13.7711868 13.9322734 13.5942826 13.84910593
var     51.9840520 51.6668689 49.4517588 52.14454489
stDev    6.9966650  7.0845314  6.9625577  7.17650637
varMean  5.1615632  2.5467262  1.6133446  1.00593516
stErr    2.2046855  1.5728822  1.2575971  0.99676695
lowerCI 10.1002541 11.0950624 11.3059091 11.99951953
upperCI 18.6909923 17.4476325 16.4363188 16.07485653

Percentage of confidence intervals (95%) that caught the population mean...
n.10 n.20 n.30 n.50
90.0 92.5 93.5 96.0
```

The output from the above summary is fairly lengthy, as it includes the summaries from the individual objects and both normal theory and bootstrap intervals were calculated (`type = 'both'` by default). Note that the sample size fields in the "montePop" object are present because `n` is an argument to `monte`. Please note that only a few bootstrap ($R = 120$) and Monte Carlo samples (200) have been used here for illustration, more are often recommended in practice.

5.2 Plotting the object

We can plot histograms of the different results from a "monte" object. Since it is possible to create histograms of several of the component slots, you must specify which one you want to display. The options are `type = c('normal', 'bootstrap', 'population')`...

```
R> hist(mo, type='boot')
```

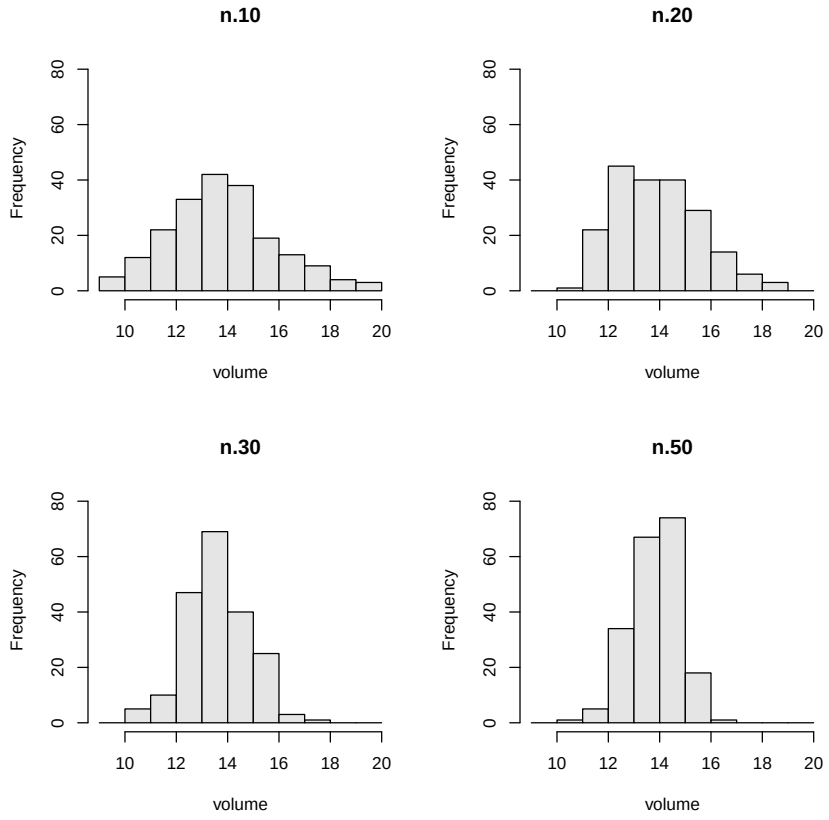

Figure 2: Histogram of a "monte" object showing the bootstrap results for each sample size.

In Figure 2 all sample sizes, n, that were requested are displayed. In general, the formal argument (`n`) to the function accepts a subset or all (the default) of the sample sizes present in the "monte" object. This will work for both the normal theory and bootstrap histograms, and is, of course, not applicable to the population histogram.

The histograms can also be plotted independently by simply referring to the individual slot objects in the "monte" object. For example, the above histogram could also have been created using...

```
R> hist(mo@BSsamples)
```

References

J. P. Barrett and L. Goldsmith. When is n sufficiently large? *The American Statistician*, 30:67–70, 1976. 1

J. P. Barrett and M. E. Nutt. *Survey sampling in the environmental sciences: A computer approach.* COMPress, Inc., 1979. 1