Bootstrap confidence intervals
Class 24, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be able to construct and sample from the empirical distribution of data.
2. Be able to explain the bootstrap principle.
3. Be able to design and run an empirical bootstrap to compute confidence intervals.
4. Be able to design and run a parametric bootstrap to compute confidence intervals.

2 Introduction

The empirical bootstrap is a statistical technique popularized by Bradley Efron in 1979. Though remarkably simple to implement, the bootstrap would not be feasible without modern computing power. The key idea is to perform computations on the data itself to estimate the variation of statistics that are themselves computed from the same data. That is, the data is 'pulling itself up by its own bootstrap.' (A google search of 'by ones own bootstraps' will give you the etymology of this metaphor.) Such techniques existed before 1979, but Efron widened their applicability and demonstrated how to implement the bootstrap effectively using computers. He also coined the term 'bootstrap' [1].

Our main application of the bootstrap will be to estimate the variation of point estimates; that is, to estimate confidence intervals. An example will make our goal clear.

Example 1. Suppose we have data

$$x_1, x_2, \dots, x_n$$

If we knew the data was drawn from $\mathrm{N}(\mu, \sigma^2)$ with the unknown mean μ and known variance σ^2 then we have seen that

$$\left[\overline{x} - 1.96\frac{\sigma}{\sqrt{n}},\ \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

is a 95% confidence interval for μ.

Now suppose the data is drawn from some completely unknown distribution. To have a name we'll call this distribution F and its (unknown) mean μ. We can still use the sample mean $\overline{x}$ as a point estimate of μ. But how can we find a confidence interval for μ around $\overline{x}$? Our answer will be to use the bootstrap!

In fact, we'll see that the bootstrap handles other statistics as easily as it handles the mean. For example: the median, other percentiles or the trimmed mean. These are statistics where, even for normal distributions, it can be difficult to compute a confidence interval from theory alone.

[1] Paraphrased from Dekking et al. *A Modern Introduction to Probabilty and Statistics*, Springer, 2005, page 275.

3 Sampling

In statistics to sample from a set is to choose elements from that set. In a random sample the elements are chosen randomly. There are two common methods for random sampling.

Sampling without replacement
Suppose we draw 10 cards at random from a deck of 52 cards without putting any of the cards back into the deck between draws. This is called sampling without replacement or simple random sampling. With this method of sampling our 10 card sample will have no duplicate cards.

Sampling with replacement
Now suppose we draw 10 cards at random from the deck, but after each draw we put the card back in the deck and shuffle the cards. This is called sampling with replacement. With this method, the 10 card sample might have duplicates. It's even possible that we would draw the 6 of hearts all 10 times.

Think: What's the probability of drawing the 6 of hearts 10 times in a row?

Example 2. We can view rolling an 8-sided die repeatedly as sampling with replacement from the set {1,2,3,4,5,6,7,8}. Since each number is equally likely, we say we are sampling uniformly from the data. There is a subtlety here: each data point is equally probable, but if there are repeated values within the data those values will have a higher probability of being chosen. The next example illustrates this.

Note. In practice if we take a small sample from a very large set then it doesn't matter whether we sample with or without replacement. For example, if we randomly sample 400 out of 300 million people in the U.S., then it is so unlikely that the same person will be picked twice that there is no real difference between sampling with or without replacement.

4 The empirical distribution of data

The empirical distribution of data is simply the distribution that you see in the data. Let's illustrate this with an example.

Example 3. Suppose we roll an 8-sided die 10 times and get the following data, written in increasing order:

```
1, 1, 2, 3, 3, 3, 3, 4, 7, 7.
```

Imagine writing these values on 10 slips of paper, putting them in a hat and drawing one at random. Then, for example, the probability of drawing a 3 is 4/10 and the probability of drawing a 4 is 1/10. The full empirical distribution can be put in a probability table

value x	1	2	3	4	7
$p(x)$	2/10	1/10	4/10	1/10	2/10

Notation. If we label the true distribution from which the data is drawn from as F, then we'll label the empirical distribution of the data as F^*. If we have enough data then the law of large numbers tells us that F^* should be a good approximation of F.

Example 4. In the dice example just above, the true and empirical distributions are:

value x	1	2	3	4	5	5	7	8
true $p(x)$	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8
empirical $p(x)$	2/10	1/10	4/10	1/10	0	0	2/10	0

The true distribution F and the empirical distribution F^* of the 8-sided die.

Because F^* is derived strictly from data we call it the empirical distribution of the data. We will also call it the resampling distribution. Notice that we always know F^* explicitly. In particular the expected value of F^* is just the sample mean $\overline{x}$.

5 Resampling

The empirical bootstrap proceeds by resampling from the data. We continue the dice example above.

Example 5. Suppose we have 10 data points, given in increasing order:

```
1, 1, 2, 3, 3, 3, 3, 4, 7, 7
```

We view this as a sample taken from some underlying distribution. To resample is to sample with replacement from the empirical distribution, e.g. put these 10 numbers in a hat and draw one at random. Then put the number back in the hat and draw again. Draw as many numbers as the desired size of the resample.

To get us a little closer to implementing this on a computer we rephrase this in the following way. Label the 10 data points x_1, $x_2, \ldots, x_{10}$. To resample is to draw a number j from the uniform distribution on $\{1, 2, \ldots, 10\}$ and take x_j as our resampled value. In this case we could do so by rolling a 10-sided die. For example, if we roll a 6 then our resampled value is 3, the 6^{th} element in our list.

If we want a resampled data set of size 5, then we roll the 10-sided die 5 times and choose the corresponding elements from the list of data. If the 5 rolls are

```
5, 3, 6, 6, 1
```

then the resample is

```
3, 2, 3, 3, 1.
```

Notes: 1. Because we are sampling with replacement, the same data point can appear multiple times when we resample.

2. Also because we are sampling with replacement, we can have a resample data set of any size we want, e.g. we could resample 1000 times.

Of course, in practice one uses a software package like R to do the resampling.

5.1 Star notation

If we have sample data of size n

$$x_1, x_2, \ldots, x_n$$

then we denote a resample of size m by adding a star to the symbols

$$x_1^*, x_2^*, \ldots, x_m^*$$

Similarly, just as $\overline{x}$ is the mean of the original data, we write $\overline{x}^*$ for the mean of the resampled data.

6 The empirical bootstrap

Suppose we have n data points

$$x_1, x_2, \ldots, x_n$$

drawn from a distribution F. An empirical bootstrap sample is a resample of the same size n:

$$x_1^*, x_2^*, \ldots, x_n^*.$$

You should think of the latter as a sample of size n drawn from the empirical distribution F^*. For any statistic v computed from the original sample data, we can define a statistic v^* by the same formula but computed instead using the resampled data. With this notation we can state the bootstrap principle.

6.1 The bootstrap principle

The bootstrap setup is as follows:

1. $x_1, x_2, \ldots, x_n$ is a data sample drawn from a distribution F.
2. u is a statistic computed from the sample.
3. F^* is the empirical distribution of the data (the resampling distribution).
4. $x_1^*, x_2^*, \ldots, x_n^*$ is a resample of the data of the same size as the original sample
5. u^* is the statistic computed from the resample.

Then the bootstrap principle says that

1. F^* is approximately equal to F.

2. The statistic u is well approximated by u^*.

3. The variation of u is well approximated by the variation of u^*.

Since we have the data in hand and can compute the statistic u exactly, point 2 is not interesting in itself. Our real interest is in point 3: we can approximate the variation of u by that of u^*. We will exploit this to estimate the size of confidence intervals.

6.2 Why the resample is the same size as the original sample

This is straightforward: the variation of the statistic u will depend on the size of the sample. If we want to approximate this variation we need to use resamples of the same size.

6.3 Toy example of an empirical bootstrap confidence interval

Example 6. Toy example. We start with a made-up set of data that is small enough to show each step explicitly. The sample data is

```
30, 37, 36, 43, 42, 43, 43, 46, 41, 42
```

Problem: Estimate the mean μ of the underlying distribution and give an 80% bootstrap confidence interval.

Note: R code for this example is shown in the section 'R annotated transcripts' below. The code is also implemented in the R script `class24-empiricalbootstrap.r` which is posted with our other R code.

answer: The sample mean is $\overline{x} = 40.3$. We use this as an estimate of the true mean μ of the underlying distribution. As in Example 1, to make the confidence interval we need to know how much the distribution of $\overline{x}$ varies around μ. That is, we'd like to know the distribution of

$$\delta = \overline{x} - \mu.$$

If we knew this distribution we could find $\delta_{.1}$ and $\delta_{.9}$, the 0.1 and 0.9 critical values of δ. Then we'd have

$$P(\delta_{.9} \le \overline{x} - \mu \le \delta_{.1} \,|\, \mu) = 0.8 \;\Leftrightarrow\; P(\overline{x} - \delta_{.9} \ge \mu \ge \overline{x} - \delta_{.1} \,|\, \mu) = 0.8$$

which gives an 80% confidence interval of

$$[\overline{x} - \delta_{.1},\ \overline{x} - \delta_{.9}]\,.$$

As always with confidence intervals, we hasten to point out that the probabilities computed above are probabilities concerning the statistic $\overline{x}$ given that the true mean is μ.

The bootstrap principle offers a practical approach to estimating the distribution of $\delta = \overline{x} - \mu$. It says that we can approximate it by the distribution of

$$\delta^* = \overline{x}^* - \overline{x}$$

where $\overline{x}^*$ is the mean of an empirical bootstrap sample.

Here's the beautiful key to this: since δ^* is computed by resampling the original data, we we can have a computer simulate δ^* as many times as we'd like. Hence, by the law of large numbers, we can estimate the distribution of δ^* with high precision.

Now let's return to the sample data with 10 points. We used R to generate 20 bootstrap samples, each of size 10. Each of the 20 columns in the following array is one bootstrap sample.

```
43 36 46 30 43 43 43 37 42 42 43 37 36 42 43 43 42 43 42 43
43 41 37 37 43 43 46 36 41 43 43 42 41 43 46 36 43 43 43 42
42 43 37 43 46 37 36 41 36 43 41 36 37 30 46 46 42 36 36 43
37 42 43 41 41 42 36 42 42 43 42 43 41 43 36 43 43 41 42 46
42 36 43 43 42 37 42 42 42 46 30 43 36 43 43 42 37 36 42 30
36 36 42 42 36 36 43 41 30 42 37 43 41 41 43 43 42 46 43 37
43 37 41 43 41 42 43 46 46 36 43 42 43 30 41 46 43 46 30 43
41 42 30 42 37 43 43 42 43 43 46 43 30 42 30 42 30 43 43 42
46 42 42 43 41 42 30 37 30 42 43 42 43 37 37 37 42 43 43 46
42 43 43 41 42 36 43 30 37 43 42 43 41 36 37 41 43 42 43 43
```

Next we compute $\delta^* = \overline{x}^* - \overline{x}$ for each bootstrap sample (i.e. each column) and sort them from smallest to biggest:

-1.6, -1.4, -1.4, -0.9, -0.5, -0.2, -0.1, 0.1, 0.2, 0.2, 0.4, 0.4, 0.7, 0.9, 1.1, 1.2, 1.2, 1.6, 1.6, 2.0

We will approximate the critical values $\delta_{.1}$ and $\delta_{.9}$ by $\delta^*_{.1}$ and $\delta^*_{.9}$. Since $\delta^*_{.1}$ is at the 90th percentile we choose the 18th element in the list, i.e. 1.6. Likewise, since $\delta^*_{.9}$ is at the 10th percentile we choose the 2nd element in the list, i.e. -1.4.

Therefore our bootstrap 80% confidence interval for μ is

$$[\overline{x} - \delta^*_{.1},\ \overline{x} - \delta^*_{.9}] \ = \ [40.3 - 1.6,\ 40.3 + 1.4] \ = \ [38.7,\ 41.7]$$

In this example we only generated 20 bootstrap samples so they would fit on the page. Using R, we would generate 10000 or more bootstrap samples in order to obtain a very accurate estimate of $\delta^*_{.1}$ and $\delta^*_{.9}$.

6.4 Justification for the bootstrap principle

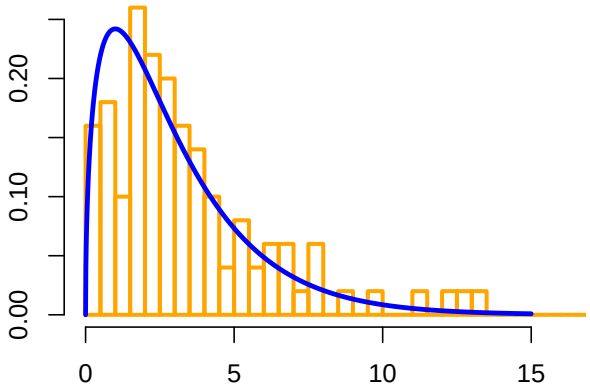
The bootstrap is remarkable because resampling gives us a decent estimate on how the point estimate might vary. We can only give you a 'hand-waving' explanation of this, but it's worth a try. The bootstrap is based roughly on the law of large numbers: for large enough data sets, the empirical distribution will be a good approximation of the true distribution. Visually it says that the histogram of the data should approximate the density of the true distribution.

First let's note what resampling can't do for us: it can't improve our point estimate. For example, if we estimate the mean μ by $\overline{x}$ then in the bootstrap we would compute $\overline{x}^*$ for many resamples of the data. If we took the average of all the $\overline{x}^*$ we would expect it to be very close to $\overline{x}$. This wouldn't tell us anything new about the true value of μ.

Even with a fair amount of data the match between the true and empirical distributions is not perfect, so there will be error in our estimates for the mean (or any other statistic). But the amount of variation in the estimates is much less sensitive to differences between the true density and the data histogram: as long as they are reasonably close, the empirical and true distributions will exhibit the similar amounts of variation. So, in general the bootstrap principle is more robust when approximating the distribution of relative variation than when approximating absolute distributions.

What we have in mind is the scenario of our examples. The distribution (over different sets x of experimental data) of $\overline{x}$ is 'centered' at μ. The distribution (over different samples x^* of x) of $\overline{x}^*$ is centered at $\overline{x}$. If there is a significant separation between $\overline{x}$ and μ then these two distributions will also differ significantly. On the other hand the distribution of $\delta = \overline{x} - \mu$ describes the variation of $\overline{x}$ about its center. Likewise the distribution of $\delta^* = \overline{x}^* - \overline{x}$ describes the variation of $\overline{x}^*$ about its center. So even if the centers are quite different the two variations about the centers can be approximately equal.

The figure below illustrates how the empirical distribution approximates the true distribution. To make the figure we generate 100 random values from a chi-square distribution with 3 degrees of freedom. The figure shows the pdf of the true distribution as a blue line and a histogram of the empirical distribution in orange.



The true and empirical distributions are approximately equal.

7 Other statistics

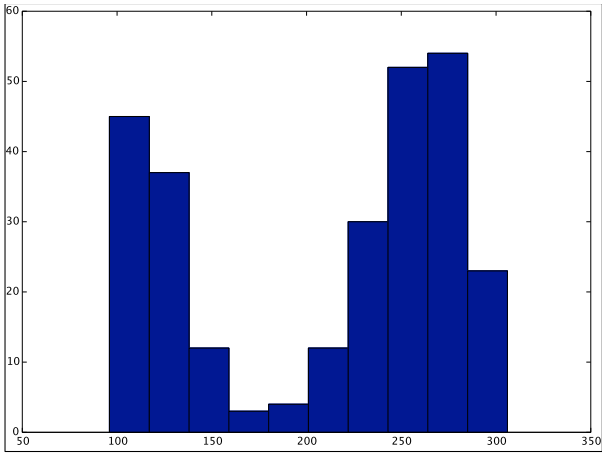
So far in this class we've avoided confidence intervals for the median and other statistics because their sample distributions are hard to describe theoretically. The bootstrap has no such problem. In fact, to handle the median all we have to do is change 'mean' to 'median' in the R code from Example 6.

Example 7. Old Faithful: confidence intervals for the median

Old Faithful is a geyser in Yellowstone National Park in Wyoming:

`http://en.wikipedia.org/wiki/Old_Faithful`

There is a publicly available data set which gives the durations of 272 consecutive eruptions. Here is a histogram of the data.



Question: Estimate the median length of an eruption and give a 90% confidence interval for the median.

answer: The full answer to this question is in the R file `oldfaithful_simple.r` and the Old Faithful data set. Both are posted on the class R code page. (Look under 'Other R code' for the old faithful script and data.)

Note: the code in `oldfaithful_simple.r` assumes that the data `oldfaithful.txt` is in the current working directory.

Let's walk through a summary of the steps needed to answer the question.

1. Data: $x_1, \ldots, x_{272}$
2. Data median: $x_{\text{median}} = 240$
3. Find the median x^*_{median} of a bootstrap sample $x^*_1, \ldots, x^*_{272}$. Repeat 1000 times.
4. Compute the bootstrap differences

$$\delta^* = x^*_{\text{median}} - x_{\text{median}}$$

Put these 1000 values in order and pick out the .95 and .05 critical values, i.e. the 50th and 950th biggest values. Call these $\delta^*_{.95}$ and $\delta^*_{.05}$.

5. The bootstrap principle says that we can use $\delta^*_{.95}$ and $\delta^*_{.05}$ as estimates of $\delta_{.95}$ and $\delta_{.05}$. So our estimated 90% bootstrap confidence interval for the median is

$$[x_{\text{median}} - \delta^*_{.05},\ x_{\text{median}} - \delta^*_{.95}]$$

The bootstrap 90% CI we found for the Old Faithful data was $[235, 250]$. Since we used 1000 bootstrap samples a new simulation starting from the same sample data should produce a similar interval. If in Step 3 we increase the number of bootstrap samples to 10000, then the intervals produced by simulation would vary even less. One common strategy is to increase the number of bootstrap samples until the resulting simulations produce intervals that vary less than some acceptable level.

Example 8. Using the Old Faithful data, estimate $P(|\overline{x} - \mu| > 5 \,|\, \mu)$.

answer: We proceed exactly as in the previous example except using the mean instead of the median.

1. Data: $x_1, \ldots, x_{272}$
2. Data mean: $\overline{x} = 209.27$
3. Find the mean $\overline{x}^*$ of 1000 empirical bootstrap samples: $x^*_1, \ldots, x^*_{272}$.
4. Compute the bootstrap differences

$$\delta^* = \overline{x}^* - \overline{x}$$

5. The bootstrap principle says that we can use the distribution of δ^* as an approximation for the distribution $\delta = \overline{x} - \mu$. Thus,

$$P(|\overline{x} - \mu| > 5 \,|\, \mu) = P(|\delta| > 5 \,|\, \mu) \approx P(|\delta^*| > 5)$$

Our bootstrap simulation for the Old Faithful data gave 0.225 for this probability.

8 Parametric bootstrap

The examples in the previous sections all used the empirical bootstrap, which makes no assumptions at all about the underlying distribution and draws bootstrap samples by resampling the data. In this section we will look at the parametric bootstrap. The only difference

between the parametric and empirical bootstrap is the source of the bootstrap sample. For the parametric bootstrap, we generate the bootstrap sample from a parametrized distribution.

Here are the elements of using the parametric bootstrap to estimate a confidence interval for a parameter.

0. Data: $x_1, \dots, x_n$ drawn from a distribution $F(\theta)$ with unknown parameter θ.

1. A statistic $\hat{\theta}$ that estimates θ.

2. Our bootstrap samples are drawn from $F(\hat{\theta})$.

3. For each bootstrap sample

$$x_1^*, \dots, x_n^*$$

we compute $\hat{\theta}^*$ and the bootstrap difference $\delta^* = \hat{\theta}^* - \hat{\theta}$.

4. The bootstrap principle says that the distribution of δ^* approximates the distribution of $\delta = \hat{\theta} - \theta$.

5. Use the bootstrap differences to make a bootstrap confidence interval for θ.

Example 9. Suppose the data $x_1, \dots, x_{300}$ is drawn from an $\exp(\lambda)$ distribution. Assume also that the data mean $\overline{x} = 2$. Estimate λ and give a 95% parametric bootstrap confidence interval for λ.

answer: This is implemented in the R script `class24-parametricbootstrap.r` which is posted with our other R code.

It's will be easiest to explain the solution using commented code.

```
# Parametric bootstrap

# Given 300 data points with mean 2.
# Assume the data is exp(lambda)

# PROBLEM: Compute a 95% parametric bootstrap confidence interval for lambda

# We are given the number of data points and mean
n = 300
xbar = 2

# The MLE for lambda is 1/xbar
lambdahat = 1.0/xbar

# Generate the bootstrap samples
# Each column is one bootstrap sample (of 300 resampled values)
nboot = 1000

# Here's the key difference with the empirical bootstrap:
# We draw the bootstrap sample from Exponential(lambdahat)
x = rexp(n*nboot, lambdahat)
bootstrapsample = matrix(x, nrow=n, ncol=nboot)

# Compute the bootstrap lambdastar
lambdastar = 1.0/colMeans(bootstrapsample)

# Compute the differences
deltastar = lambdastar - lambdahat
```

```
# Find the 0.05 and 0.95 quantile for deltastar
d = quantile(deltastar, c(0.05,0.95))

# Calculate the 95% confidence interval for lambda.
ci = lambdahat - c(d[2], d[1])

# This line of code is just one way to format the output text.
# sprintf is an old C function for doing this.  R has many other
# ways to do the same thing.
s = sprintf("Confidence interval for lambda:  [%.3f, %.3f]", ci[1], ci[2])
cat(s)
```

9 The bootstrap percentile method (should not be used)

Instead of computing the differences δ^*, the bootstrap percentile method uses the distribution of the bootstrap sample statistic as a direct approximation of the data sample statistic.

Example 10. Let's redo Example 6 using the bootstrap percentile method.

We first compute $\overline{x}^*$ from the bootstrap samples given in Example 6. After sorting we get

35.7 37.4 38.0 39.5 39.7 39.8 39.8 40.1 40.1 40.6 40.7 40.8 41.1 41.1 41.7 42.0 42.1 42.4 42.4 42.4

The percentile method says to use the distribution of $\overline{x}^*$ as an approximation to the distribution of $\overline{x}$. The 0.9 and 0.1 critical values are given by the 2nd and 18th elements. Therefore the 80% confidence interval is [37.4, 42.4]. This is a bit wider than our answer to Example 6.

The bootstrap percentile method is appealing due to its simplicity. However it depends on the bootstrap distribution of $\overline{x}^*$ based on a particular sample being a good approximation to the true distribution of $\overline{x}$. Rice says of the percentile method, "Although this direct equation of quantiles of the bootstrap sampling distribution with confidence limits may seem initially appealing, it's rationale is somewhat obscure." [2] In short, **don't use the bootstrap percentile method**. Use the empirical bootstrap instead (we have explained both in the hopes that you won't confuse the empirical bootstrap for the percentile bootstrap).

10 R annotated transcripts

10.1 Using R to generate an empirical bootstrap confidence interval

This code only generates 20 bootstrap samples. In real practice we would generate many more bootstrap samples. It is making a bootstrap confidence interval for the mean. This code is implemented in the R script `class24-empiricalbootstrap.r` which is posted with our other R code.

```
# Data for the example 6
x = c(30,37,36,43,42,43,43,46,41,42)
n = length(x)
```

[2] John Rice, *Mathematical Statistics and Data Analysis*, 2nd edition, p. 272.

```
# sample mean
xbar = mean(x)

nboot = 20
# Generate 20 bootstrap samples, i.e.  an n x 20 array of
# random resamples from x
tmpdata = sample(x,n*nboot, replace=TRUE)
bootstrapsample = matrix(tmpdata, nrow=n, ncol=nboot)

# Compute the means x̄*
bsmeans = colMeans(bootstrapsample)

# Compute δ* for each bootstrap sample
deltastar = bsmeans - xbar

# Find the 0.1 and 0.9 quantile for deltastar
d = quantile(deltastar, c(0.1, 0.9))

# Calculate the 80% confidence interval for the mean.
ci = xbar - c(d[2], d[1])
cat('Confidence interval:  ',ci, '\n')

# ALTERNATIVE: the quantile() function is sophisticated about
# choosing a quantile between two data points.  A less sophisticated
# approach is to pick the quantiles by sorting deltastar and
# choosing the index that corresponds to the desired quantiles.
# We do this below.

# Sort the results
sorteddeltastar = sort(deltastar)

# Look at the sorted results
hist(sorteddeltastar, nclass=6)
print(sorteddeltastar)

# Find the .1 and .9 critical values of deltastar
d9alt = sorteddeltastar[2]
d1alt = sorteddeltastar[18]

# Find and print the 80% confidence interval for the mean
ciAlt = xbar - c(d1alt,d9alt)
cat('Alternative confidence interval:  ',ciAlt, '\n')
```