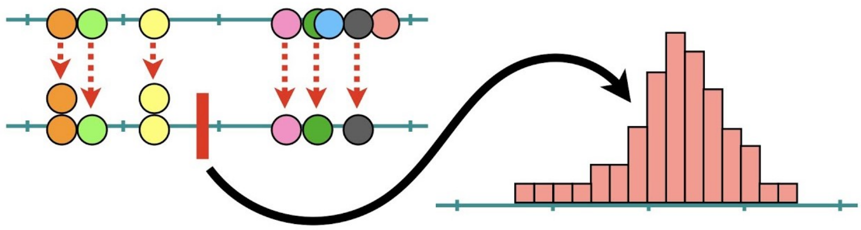
# Bootstrap Method : A Review




**Name :** Swarnajit Podder

**Roll No. :** 430

**Department :** Statistics

**Session :** 2019 - 2022
**Supervisor :** Dr. Durba Bhattacharya

St. Xavier's College (Autonomous) - Kolkata

## Declaration

***I affirm that I have identified all my sources and that no part of my dissertation paper uses unacknowledged materials.***

**Signature :**

Swarnajit Podder.

# Contents

# 1 Introduction

In applied statistics, estimation of a parameter from an unknown population comprises two steps -

- determination of an estimator for the parameter and,
- determination of the accuracy of that estimator. The accuracy can be evaluated either from the estimate of standard error of the estimator or finding an appropriate confidence interval.

Generally the above problem is handled by drawing random sample(s) from the population. The sample observations can be considered as a realisation of the i.i.d. random variables with a common distribution $F$. Based on them, we determine the estimate of the parameter under consideration, say, $\hat{\theta}$. We would like to study the sampling distribution of $\hat{\theta}$, but two situations may arise here :

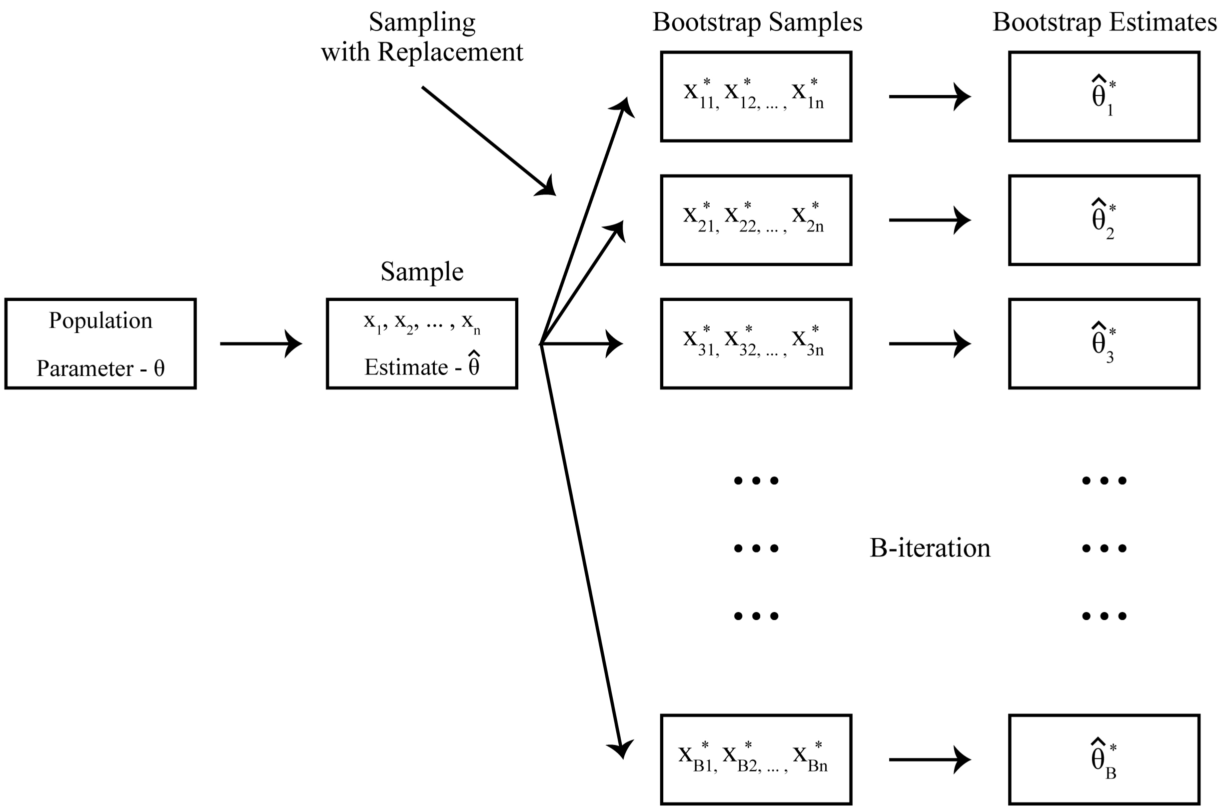
- The population distribution $F$ may be completely unknown.
- Even if we know $F$, $\hat{\theta}$ is such a function of sample units, that the sampling distribution is difficult to determine or not of a closed form.

This problem may be alleviated if large number of sample observations are available, in which case we can leverage the central limit theorem to approximate the sampling distribution as Normal distribution. But in real life, collecting samples of large size may not always be possible due to lack of resources or difficulties in sample survey.
The resampling technique provides a way to handle the situations without any parametric assumptions. However these techniques go back a long way (e.g. Jackknife (1949), Permutation testing (1930s)). It was **Efron** who unified and connected the ideas to introduce the **Bootstrap method** in 1979. The method is developed and extended in the later papers by monograph Efron (1982), Efron and Gong (1983), Efron and Tibshirani (1986) etc.

Here is an informal description of how the Bootstrap method works. Consider the **empirical distribution** of the sample in hand (The empirical distribution is the probability distribution that has probability $\frac{1}{n}$ assigned to each sample value). The idea is to replace the unknown population distribution with the empirical distribution and determine the properties of the estimator, such as standard error, based on it.
The general procedure is like this -

- A bootstrap sample, i.e. a sample with replacement is generated from the empirical distribution.
- Value of the parameter under consideration (denoted by $\hat{\theta}$) is estimated using the bootstrap sample instead of the original one.
- Repeat the above two steps a large number of times, say $B$ (e.g., for estimation of standard error, B is generally to be taken as at least 100).

Here is a pictorial representation of the method.




It should be noted that almost any statistic based on the bootstrap distribution can be used as a bootstrap estimate of the corresponding population parameter. Though ignorant use of bootstrapping may lead to inconsistency and the result obtained may not be valid. This will be discussed in a later section.

## 2 Point Estimation using bootstrap

Let us check accuracy of Bootstrap method in point estimation using some examples .

### 2.1 Example 1 : Distribution of sample mean of $N(\mu, \sigma^2)$ distribution

**Result :** Let $X_1, X_2, ..., X_n$ be a sample of size n from $N(\mu, \sigma^2)$ distribution. Then the sample mean $\bar{X}$ follows $N(\mu, \frac{\sigma^2}{n})$ distribution.

**Algorithm**

**Step 1 :** A random sample of size 30 is drawn from $N(2,4)$ distribution. The sample is named as $x$.

**Step 2 :** $B$ independent bootstrap samples each consists of 30 data values are drawn **with replacement** from $x$. [Take $B$ = 1000]

**Step 3 :** Sample means are computed for each bootstrap replications as $\bar{x}_1^*$, $\bar{x}_2^*$, ... , $\bar{x}_B^*$.

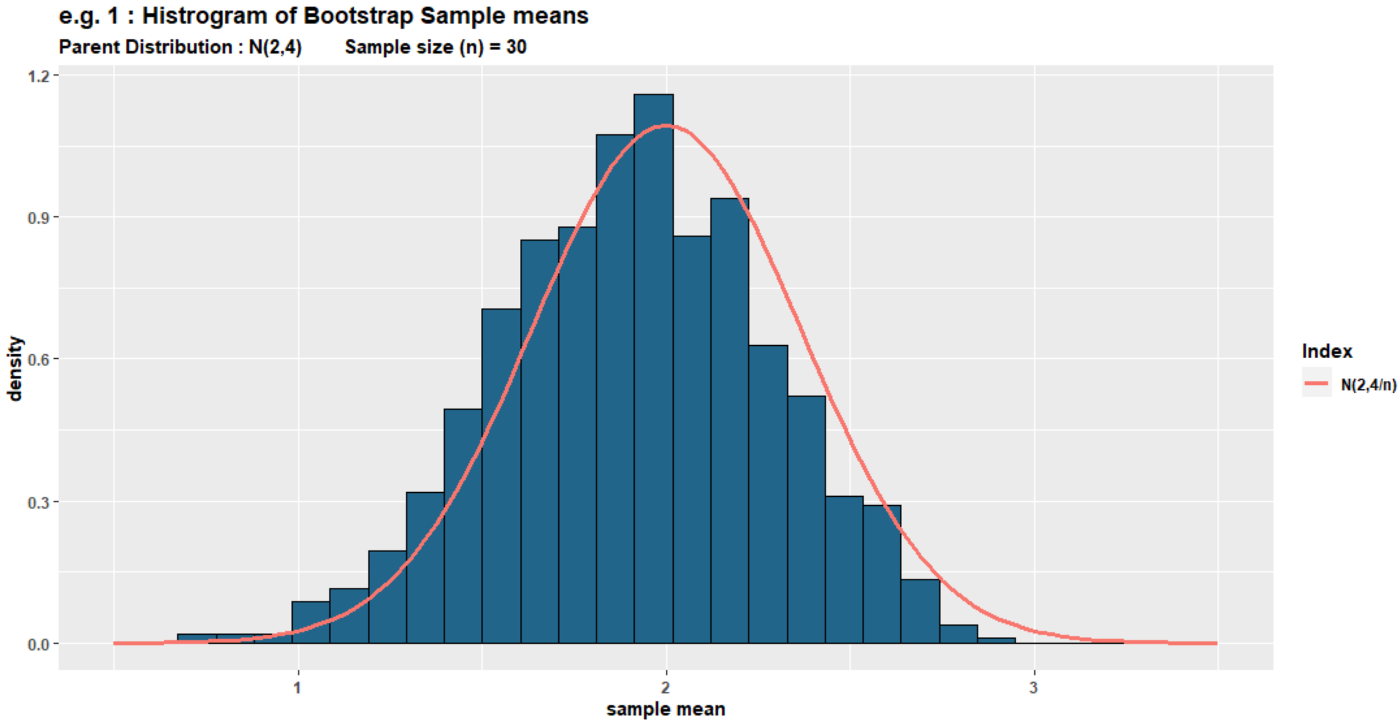
The estimate of sample mean is computed as -

$$\hat{\mu}_B = \sum_{n=1}^{B} \bar{x}_n^* / B \tag{1}$$

and estimate of standard error is computed as -

$$s(\hat{\mu}_B) = [\sum_{n=1}^{B} (\bar{x}_n^* - \hat{\mu}_B)^2 / (B-1)]^{\frac{1}{2}} \tag{2}$$

A histogram is drawn based on the bootstrap estimates obtained from each resamples and the original distribution is fitted on it.



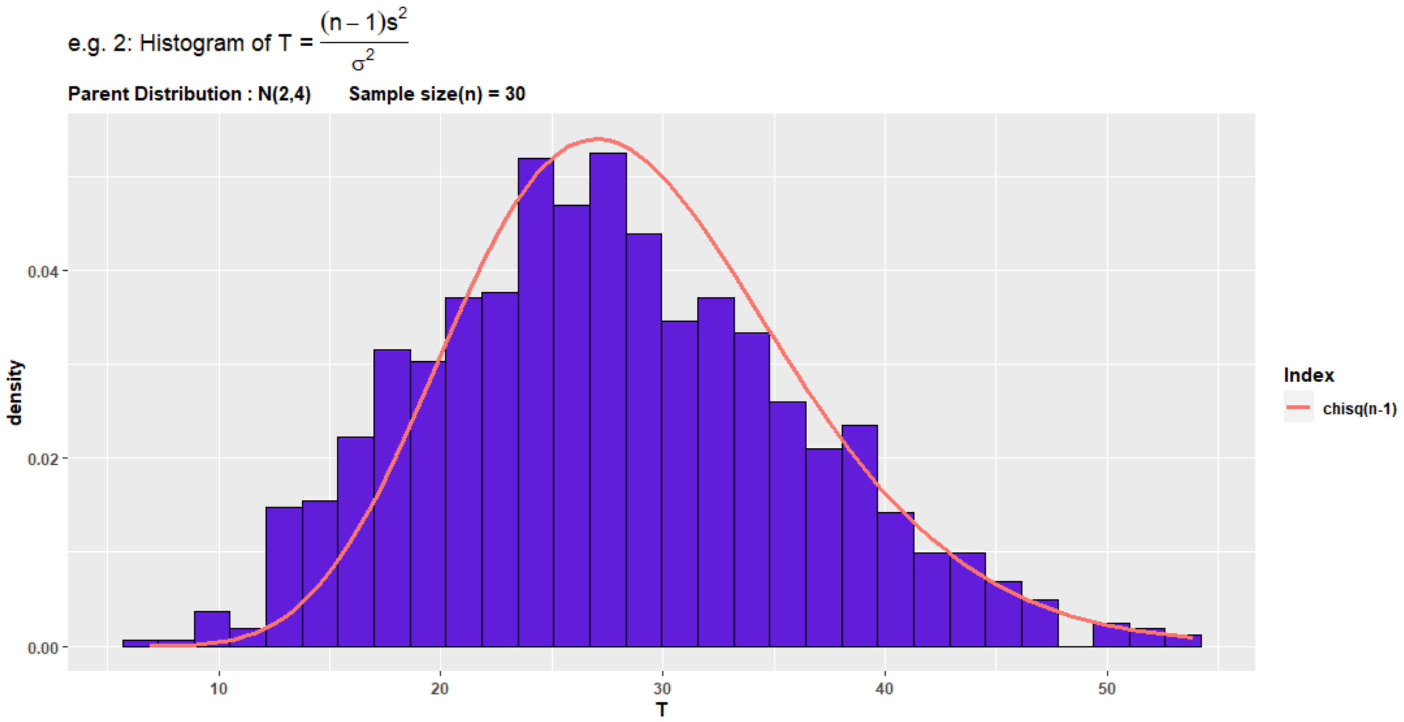

Here, we can see the fit is quite good. Also, the estimate of $\mu$ (sample mean) is obtained as 1.913 and estimate of standard error of sample mean is obtained as 0.3676 which are very close to the theoretically obtained values 2 and 0.3652 respectively.

We, hereafter, follow the same algorithm to check some more theoretically obtained sampling distributions.

## 2.2 Example 2 : Distribution of sample variance of $N(\mu, \sigma^2)$ distribution

**Result :** Let $X_1, X_2, ..., X_n$ be a sample of size n from $N(\mu, \sigma^2)$ distribution. Then the statistic $T = (n-1) * \frac{s^2}{\sigma^2}$ follows $\chi^2_{n-1}$ distribution [s : sample standard deviation].

A random sample of size 30 is drawn from $N(2,4)$ distribution, named as $x$. $B$ independent bootstrap samples each consists of 30 data values are drawn **with replacement** from $x$ [Take $B$ = 1000]. Then, sample standard variances are obtained from each resamples and the statistic $T$ is computed.

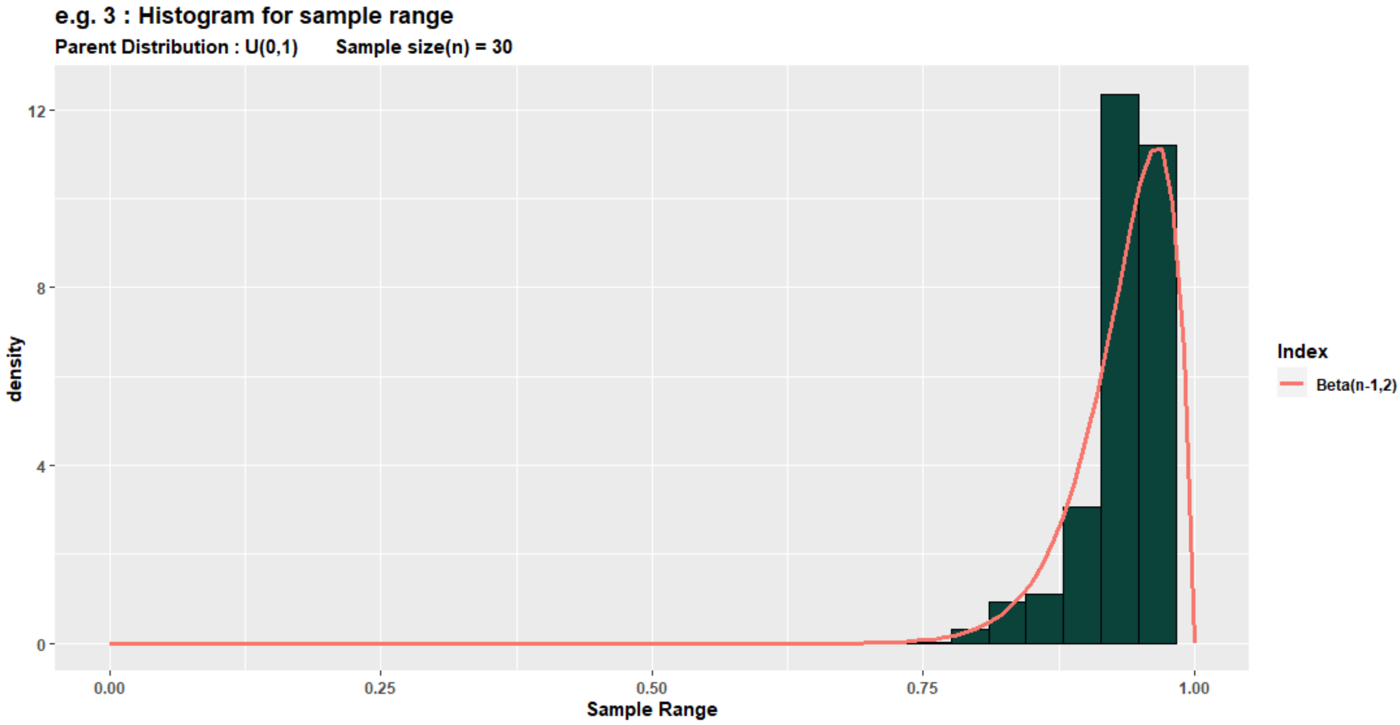



Here, we can see the fit is quite good. Also, the estimate of $\sigma^2$ (sample variance) is obtained as 3.916 and estimate of standard error of sample variance is obtained as 1.3082 which are very close to the theoretically obtained values 4 and 1.103 respectively.

## 2.3 Example 3 : Distribution of sample range of $U(0,1)$ distribution

**Result :** Let $X_1, X_2, \ldots, X_n$ be a sample of size n from U(0,1) distribution.Then the sample range follows $Beta(n-1,2)$ distribution of first kind.

A random sample of size 30 is drawn from $U(0,1)$ distribution, named as $x$. $B$ independent bootstrap samples each consists of 30 data values are drawn **with replacement** from $x$ [Take $B = 1000$]. Then, sample ranges are obtained from each resamples.




Here, we can see the fit is quite good. Also, the bootsrap estimate of Expectation of the distribution is obtained as 0.9378 which is very close to the theoretically obtained values 0.9355.

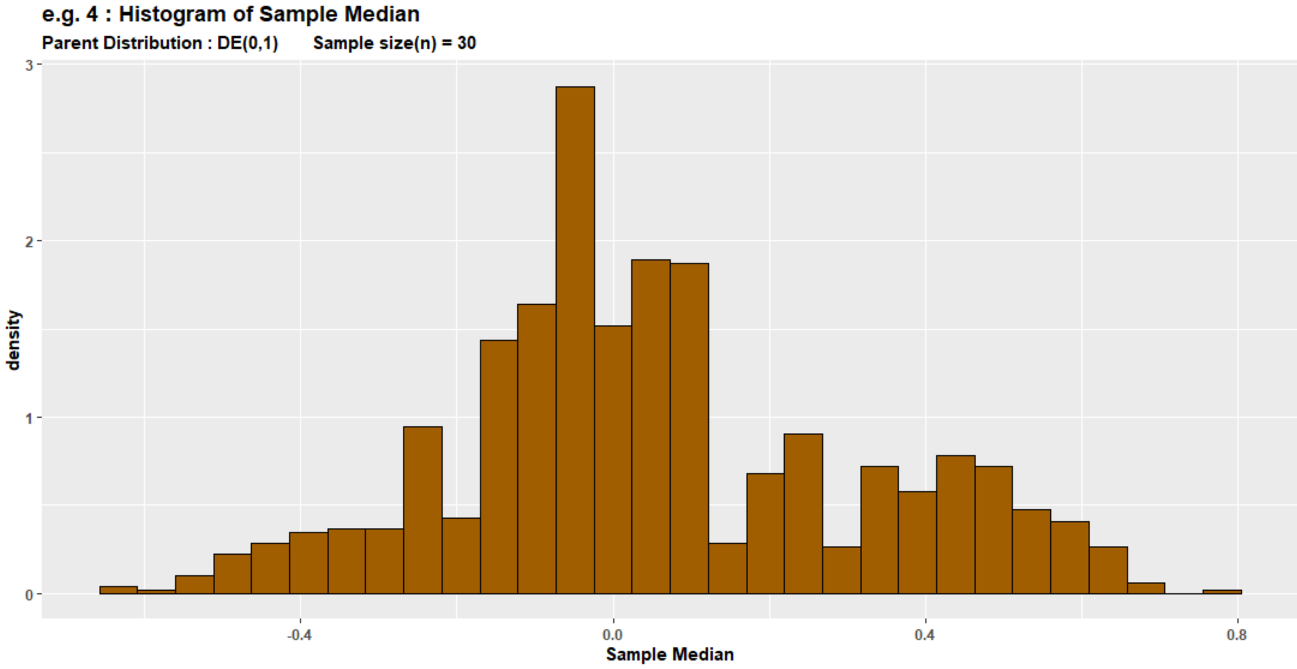
## 2.4 Example 4 : Distribution of sample median of DE(0,1) distribution

Let $X_1, X_2, \ldots, X_n$ be a sample of size n from $DE(0,1)$ distribution. We wish to obtain estimate of standard error of sample median. But it is quite difficult to obtain sampling distribution of sample median for Double Exponential distribution.

We use bootstrap technique here to obtain an estimate.

A random sample of size 30 is drawn from $DE(0,1)$ distribution, named as $x$ . $B$ independent bootstrap samples each consists of 30 data values are drawn **with replacement** from $x$ [Take $B$ = 1000]. Then, sample medians are obtained from each resamples.

A histogram is drawn based on the bootstrap estimates to get an idea of the nature of sampling distribution of sample median.




From the histogram, ignoring sampling fluctuations, it seems that the distribution is symmetric about 0.

The estimate of standard error is obtained as 0.2624.

# 3 Types of Bootstrap

Previously it was said that, bootstrap samples are generated with replacement from the empirical distribution. Here, no parameter is involved in drawing repeated samples. That's why, the method is commonly known as **non-parametric bootstrap**.

However, bootstrap sampling can be carried out parametrically also. Let, F (the population distribution) is unknown upto an unknown parameter $\theta$. An reasonable estimator of $\theta$ is to be chosen, say $\hat{\theta}$.

The idea of the **parametric bootstrap** is to simulate data from $F(x|\hat{\theta})$.

After generating the bootstrap samples, rest of the steps are exactly as same as the non-parametric setup.

Let's compare the two methods with some examples.

## 3.1 Example 1 : Distribution of sample mean of $N(\mu, \sigma^2)$

We know that, if $X_1, X_2, \ldots, X_n$ be a sample of size n from $N(\mu, \sigma^2)$ distribution ($\sigma$ is known), the sample mean $\bar{X}$ follows $N(\mu, \frac{\sigma^2}{n})$ distribution.

A random sample of size 30 is drawn from N(2,4) distribution. The sample mean is obtained as 1.9027.

**Step 1 :** $B$ independent bootstrap samples each consists of 30 data values are drawn from N(1.9027,4) distribution. [Take $B$ = 1000]

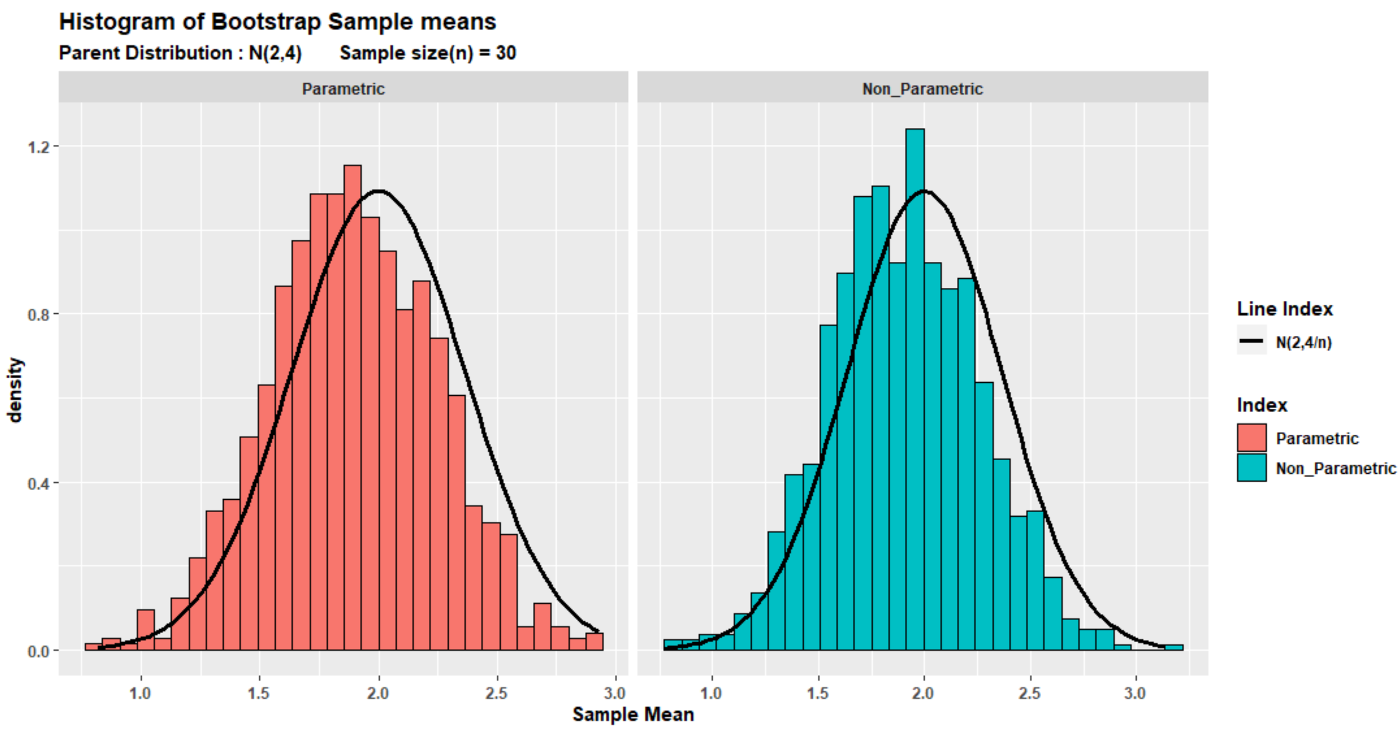
**Step 2 :** Sample means are computed for each bootstrap replications as $\bar{x}_1^*, \bar{x}_2^*, \dots, \bar{x}_B^*$.

Bootstrap estimate of sample mean and its SE are computed using eq. 1 and 2 respectively.

| | Population | Parametric Estimate | Non-parametric Estimate |
|---|---|---|---|
| Mean | 2 | 1.8936 | 1.9032 |
| Std. Error of estimator | 0.3652 | 0.3525 | 0.3521 |

Table 1: Estimates of population mean (sample mean) and its standard error using (a) Parametric Bootstrap and (b) Non-parametric Bootstrap

For both the cases, histograms are drawn and a Normal Curve (N(2,4/n) - obtained from the theoretical result) is fitted on both of them. The graphs are shown below :



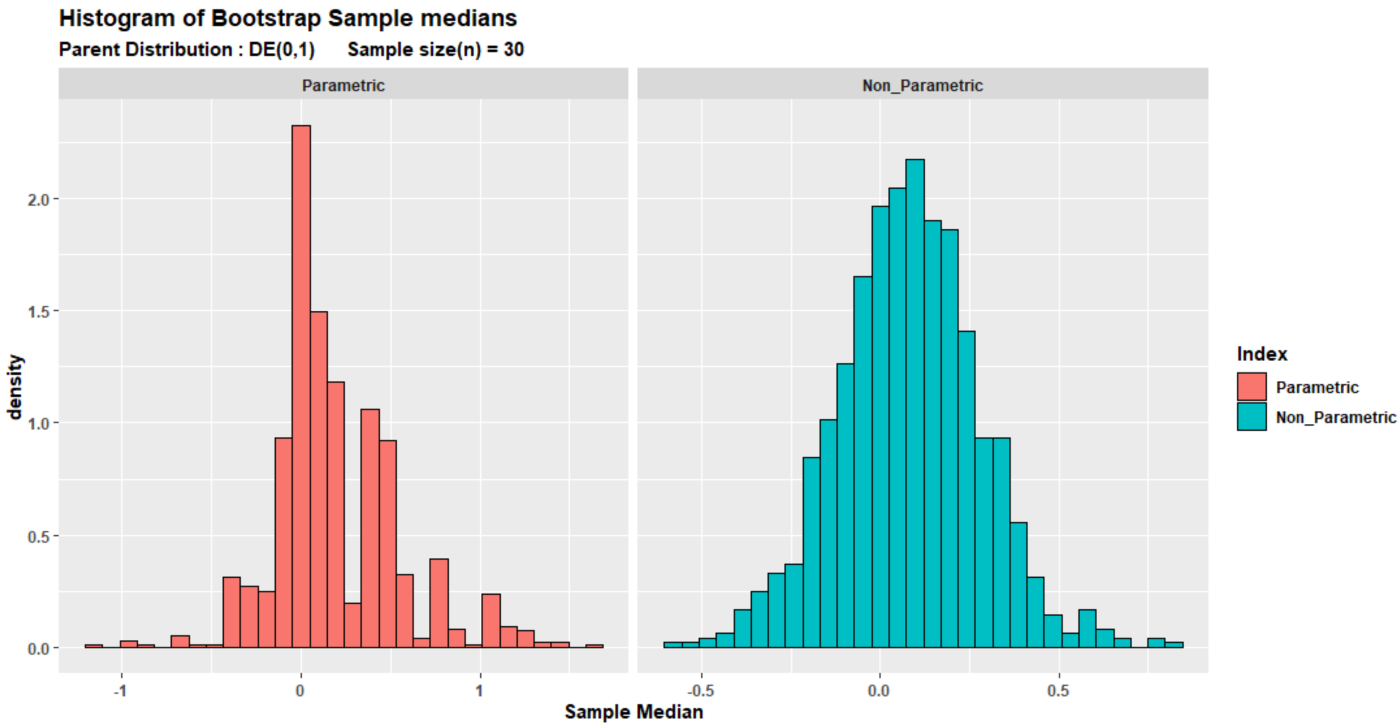

We can see that, using both approaches, the fittings are quite good and the results obtained are also quite accurate.

### 3.2 Example 2 : Distribution of sample median of $DE(\mu, b)$ distribution

Let's get back to section 2.4. It is already mentioned that sampling distribution of sample median is difficult to determine in case of double exponential distribution and so is to find estimate of standard error of it. Therefore, we used bootstrap method to simulate the distribution and find an estimate of standard error. It was done in Non-parametric technique. Now, we follow the same algorithm of the last example to perform parametric bootstrap method in this case. The results and the histograms are given below [The sample median is 0.0855746].

| | Population | Parametric Estimate | Non-parametric Estimate |
|---|---|---|---|
| Median | 0 | 0.0785 | 0.1942 |
| Std. Error of estimator | - | 0.1985 | 0.3535 |

Table 2: Estimates of population mean (sample mean) and its standard error using (a) Parametric Bootstrap and (b) Non-parametric Bootstrap




**Comment :** Here, we can see from the values that the parametric estimate is closer to the population value than the non-parametric estimate. Besides, the estimate of standard error of sample median is less for the parametric method than the non-parametric method.

On the other hand, though the histogram obtained using the non-parametric method looks more compact than the parametric method, the latter is more symmetric about the population value 0 while the former one is slightly skewed than the latter. However, overall we can conclude that distribution of sample median seems to be symmetric about the population median.

# 4 Interval Estimation

Till now, the problems have been done on point estimation. Here, different techniques for constructing confidence intervals using the bootstrap will be discussed.

Suppose, we have a sample obtained from an unknown distribution $F$. Let, $\hat{\theta}$ be the plug-in estimate of the unknown parameter $\theta$ and $s$ be some reasonable estimate of standard error for $\hat{\theta}$. Under most circumstances, as sample size $n$ goes large, the distribution of $\hat{\theta}$ can be written as -

$$\frac{\hat{\theta} - \theta}{s} \overset{a}{\sim} t_{n-1}$$

Then the **Student's t-interval** is given by -

$$(\hat{\theta} - t_{n-1}^{1-\alpha} * s, \hat{\theta} + t_{n-1}^{\alpha} * s)$$

where, $t_{n-1}^{\alpha}$ denotes the upper $\alpha$ point of the t distribution with (n-1) degrees of freedom.
This is the most general confidence interval. Now, we discuss different methods of interval estimation using bootstrap.

## 4.1 Percentile Confidence Interval

The percentile method is the most obvious way to construct a confidence interval for a parameter based on bootstrap estimates. Suppose that $\hat{\theta_i^*}$ is the bootstrap estimate from the i-th bootstrap sample where each bootstrap sample is of size $n$ . It is expected that if the observations are ordered in ascending order, an interval that contains 95% of the values of $\hat{\theta_i^*}$ to be a 95% confidence interval for $\theta$ . The most sensible way to choose the interval is to exclude the lowest 2.5% and the highest 2.5%. A bootstrap confidence interval generated this way is called a percentile method confidence interval or, more specifically, **Efron ' s percentile confidence interval**.
Let $\theta^{*(\alpha)}$ indicate the $100 \cdot \alpha$-th percentile of B bootstrap replications. The percentile interval of intended coverage 1 - $2\alpha$, is obtained directly from these percentiles as - $(\hat{\theta}^{*(\alpha)}, \hat{\theta}^{*(1-\alpha)})$.

Unfortunately, in the case of small samples (especially **for asymmetric distributions**) the percentile method does not work well. There are modifications that will get around these difficulties.

## 4.2 Bias-corrected and Accelarated (BCa) Bootstrap Interval

To overcome the overcoverage issues in percentile bootstrap CIs , the BCa method corrects both bias and skewness of the bootstrap parameter estimates by incorporating a bias-correction factor and an acceleration factor (Efron, 1987; Efron and Tibshirani, 1993). The bias-correction factor $\hat{z}_0$ is estimated as the proportion of the bootstrap estimates less than the original parameter estimate $\hat{\theta}$ -

$$\hat{z}_0 = \Phi^{-1}(\frac{\#(\hat{\theta}^* < \hat{\theta})}{B})$$

where $\Phi^{-1}$ is the inverse function of a standard normal CDF. The acceleration factor â is estimated through jackknife resampling, as -

$$\hat{a} = \frac{\Sigma_{i=1}^{n}(\hat{\theta}_{(.)} - \hat{\theta}_{(-i)})^3}{6 * [\Sigma_{i=1}^{n}(\hat{\theta}_{(.)} - \hat{\theta}_{(-i)})^2]^{3/2}}$$

where $\hat{\theta}_{(-i)}$ is estimate of $\theta$ based on the sample leaving out the i-th observation and $\hat{\theta}_{(.)}$ is the average of this estimates.

Based on the values of $\hat{z}_0$ and $\hat{a}$, the BCa Confidence Interval is given by -

$$(G^{-1}(\Phi[\hat{z}_0 + \frac{\hat{z}_0 + z^{\frac{\alpha}{2}}}{1 - \hat{a} * (\hat{z}_0 + z^{\frac{\alpha}{2}})}]), G^{-1}(\Phi[\hat{z}_0 + \frac{\hat{z}_0 + z^{1-\frac{\alpha}{2}}}{1 - \hat{a} * (\hat{z}_0 + z^{1-\frac{\alpha}{2}})}]))$$

Here, $z^{\frac{\alpha}{2}}$ is the $100.\frac{\alpha}{2}$ th percentile point of a standard normal distribution and $G$ is the empirical distribution.

## 4.3 Example 1 : Confidence interval of $\mu$ in $N(\mu, \sigma^2)$ distribution

A random sample of size 30 is drawn from N(2,4) distribution. The sample mean is 2.178788.

| Method | Confidence Interval |
|---|---|
| Student's t | (1.506300, 2.851276) |
| Percentile | (1.568250, 2.788583) |
| BCa | (1.567355, 2.785226) |

Table 3: Different Confidence Intervals for $\mu$ in N($\mu, \sigma^2$) distribution

**Comment :** Here we can see that all the confidence intervals contain the true value of the parameter $\mu$ (which is 2). However the bootstrap confidence intervals are slightly shorter in length than the classical Student's t-interval. Also, the two types of bootstrap confidence intervals (Percentile and BCa confidence intervals) are found to be almost identical.

## 4.4 Example 2 : Confidence interval of $\mu$ in DE($\mu$,b) distribution

In the above example, the distribution being normal, we didn't have to make any large sample approximations to determine classical confidence interval (Student's t-interval) for the population mean. However, if the distribution is non-normal, sometimes it is difficult to find sampling distributions of some statistics (See **example 4** of non-parametric point estimation), and so is to find the confidence intervals for population parameters involving those statistics. In these case we either can use large sample approximations or find different bootstrap confidence intervals.
Let's see an example.
Suppose, we want to find an appropriate confidence interval of the location parameter of double exponential distribution using sample median. But here, the sampling distribution of sample median is not very easy to determine and so is its standard error. So, we can take samples of large size to apply large sample approximations or use methods to find confidence intervals using bootstrap.

Here, sample median is a reasonable estimator of population median (location parameter). Also. the large sample variance of sample median is given by-

$$Var(Z_{\frac{1}{2}}) = \frac{1}{4n} f(\xi_{\frac{1}{2}})^{-2} = \frac{b^2}{n}$$

where,
$Z_{\frac{1}{2}}$ and $\xi_{\frac{1}{2}}$ are the sample median and population median respectively. $b$ is the population scale parameter.
Let us take a sample of size 30 from DE(0,1) distribution. The sample median is found to be 0.03672406.

| **Method** | **Confidence Interval** |
|---|---|
| Student's t | (-0.2783183, 0.3517664) |
| Percentile | (-0.1477735 0.2688582 ) |
| BCa | (-0.1740336, 0.2584333) |

Table 4: Different Confidence Intervals for $\mu$ in DE($\mu, b$) distribution

**Comment :** Here also, we can see that all the confidence intervals contain the value of the true parameter $\mu$ (which is 0). However the bootstrap confidence intervals are shorter in length than the classical Student's t-interval. Also, the two types of bootstrap confidence intervals (Percentile and BCa confidence intervals) are close to each other.

## 4.5 Why BCa over Percentile Confidence interval : An Example

Previously, it was mentioned that for an asymmetric distribution, percentile confidence interval doesn't work well, as this method doesn't take the factor of skewness into consideration. So, there might be an issue of overcoverage. However, BCa method helps to overcome this issue by incorporating a bias-correction factor and an acceleration factor. Let us visualize these through an example.

### 4.5.1 Confidence Interval of $\theta$ in Exp($\theta$) distribution

Suppose, we want to find a confidence interval of $\theta$ in Exp($\theta$) distribution. The estimator of $\theta$ is considered to be $\hat{\theta} = \frac{Me}{ln2}$, where $Me$ is the sample median (As the population median is $\theta ln2$).

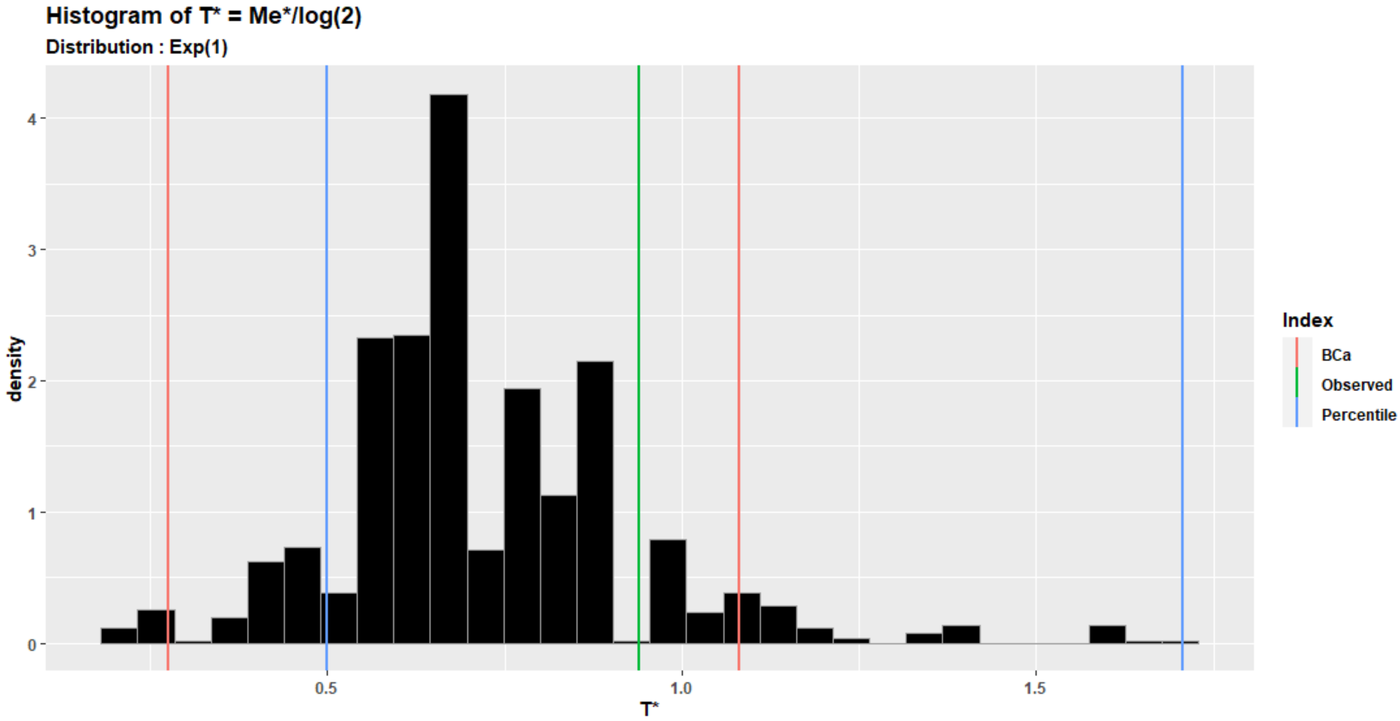
We draw a random sample of size 20 from $Exp(1)$ distribution. The sample median is 0.6510186. Therefore, the observed value of the statistic is $T_{obs} = 0.9392213$.

Now, to find the percentile and BCa confidence intervals, we generate 1000 bootstrap samples from the parent sample drawn previously to get 1000 replicates of $\hat{\theta}$, name them $T^*$, i.e., $T^* = \frac{Me^*}{ln2}$, $Me^*$ be the median of each bootstrap samples. Based on these $T^*$ values the confidence intervals are given by -

| **Method** | **Confidence Interval** |
|---|---|
| Percentile | (0.4987378, 1.7062339 ) |
| BCa | (0.2764285, 1.0817174) |

Table 5: Different Confidence Intervals for $\theta$ in Exp($\theta$) distribution

From the values, it is to be observed that the confidence limits obtained by BCa method is slightly on the left of those obtained by percentile method. Now, we draw an histogram of $T^*$ and try to establish this fact graphically.




As exponential distribution is a positively skewed distribution, the confidence interval obtained by percentile method shifts itself towards right abnormally. However, in the above picture, we can see that BCa confidence interval was able to overcome the issue.

# 5 Testing of Hypothesis

A bootstrap test, like the classical methods of testing of hypothesis, is based on a statistic, but not necessarily an estimate of the corresponding parameter. To empathise this, we denote the statistic here by $\boldsymbol{t(x)}$. We want to collect evidence whether or not we can reject the null hypothesis at certain level of significance.

Having observed the value of test statistic we want to determine the **achieved significance level (ASL)** of the test. ASL of the test is defined to be the probability of observing at least that large a value when the null hypothesis is true, i.e.,

$$ASL = P_{H_0}(\hat{\theta}^* \geq \hat{\theta})$$

The quantity $\hat{\theta}$ is fixed at its observed value and the random variable $\hat{\theta}^*$ follows the null distribution (distribution if $H_0$ is true, denoted by $F_0$).The smaller the value of ASL, the stronger the evidence against $H_0$.

In bootstrap hypothesis testing, we use a 'plug-in' style estimate for $F_0$. If we have a sample of size $n$ (denoted by $x$), the empirical distribution $\hat{F}_0$ is obtained by putting probability $\frac{1}{n}$ to each of the member of $x$.

Bootstrap method is introduced in hypothesis testing when the population null distribution is not specified completely, or there is not enough reason to assume normality (e.g. small sample size).

## 5.1 Example : Testing for equality of means

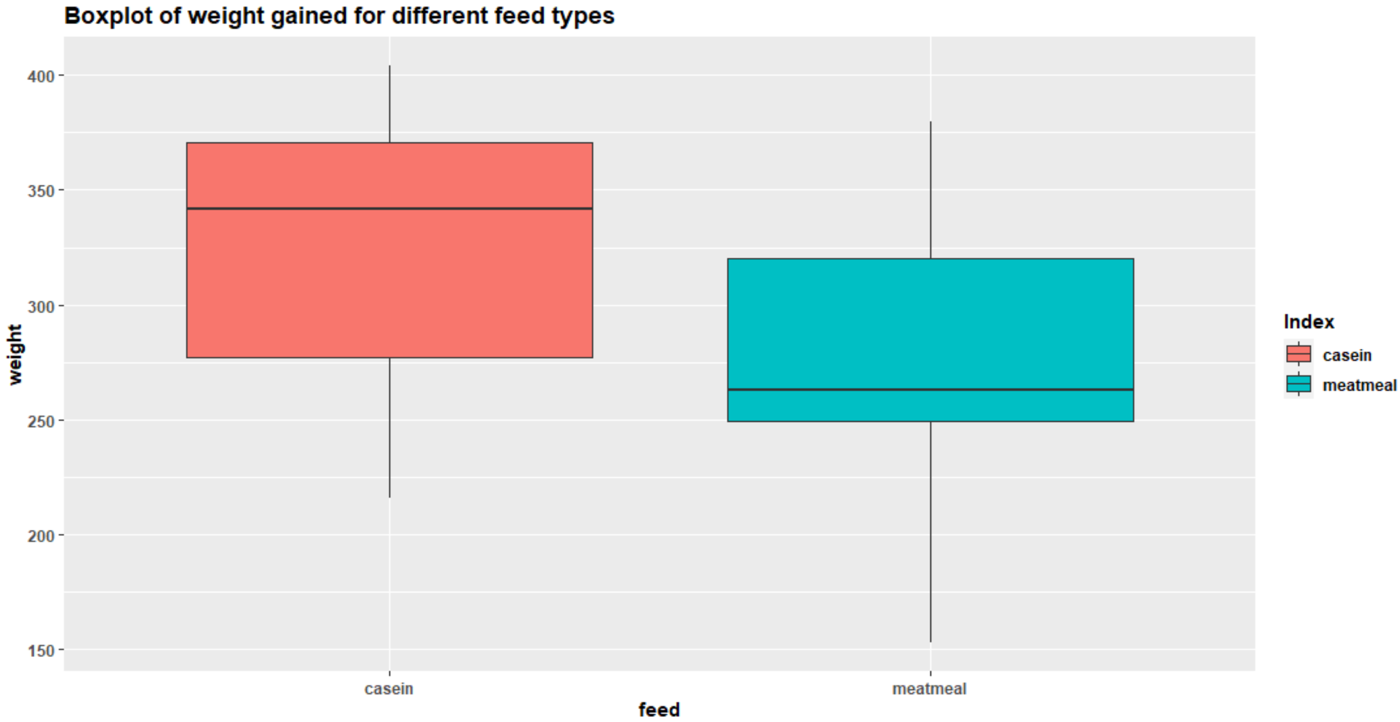
**Data :** 23 Newly hatched chicks were randomly allocated into two groups, and each group was given a different feed supplement - meatmeal and casein. Their weights in grams after six weeks are given along with feed types.

The data is given below -

| | Weight | Feed |
|---|---|---|
| 1 | 325 | meatmeal |
| 2 | 257 | meatmeal |
| 3 | 303 | meatmeal |
| 4 | 315 | meatmeal |
| 5 | 380 | meatmeal |
| 6 | 153 | meatmeal |
| 7 | 263 | meatmeal |
| 8 | 242 | meatmeal |
| 9 | 206 | meatmeal |
| 10 | 344 | meatmeal |
| 11 | 258 | meatmeal |
| 12 | 368 | casein |
| 13 | 390 | casein |
| 14 | 379 | casein |
| 15 | 260 | casein |
| 16 | 404 | casein |
| 17 | 318 | casein |
| 18 | 352 | casein |
| 19 | 359 | casein |
| 20 | 216 | casein |
| 21 | 222 | casein |
| 22 | 283 | casein |
| 23 | 332 | casein |

Table 6: Weight (in grams) of the chicks after six weeks of feeding

We want to know whether there is any significant difference in gaining weight due to different feed types. Let us first draw a boxplot of weight gain by two feed types.




From the boxplot we can see that, the distributions of weight gain by the two feed types do not seem very similar. Let's proceed with the testing procedure to check.

The null hypothesis can be given by $\boldsymbol{H_0 : \mu_1 = \mu_2}$ and the alternative hypothesis is $\boldsymbol{H_1 : \mu_1 \neq \mu_2}$, where $\mu_1$ and $\mu_2$ are average weights of the chicken after six weeks due to application of meatmeal and casein respectively. The test statistic is given by -

$$T = |\bar{z} - \bar{y}|$$

where $\bar{z}$ and $\bar{y}$ denote the sample means of weight gain by the feed types meatmeal and casein respectively. The observed value of the statistic is 46.67424.

Now, here we have only 23 observations, of which 11 are from one feed type and the rest are from the other one. The sample size is thereby not so large to assume normality as the population distribution. Let's introduce bootstrap method to do the testing procedure.

**Algorithm :**

**Step 1 :** Draw $B$ (we take $B$=10000) samples of size 23 with replacement from the data. Call the first 11 observations $z^*$ and the remaining 12 observations $y^*$.

**Step 2 :** Evaluate $t(.)$ on each sample,

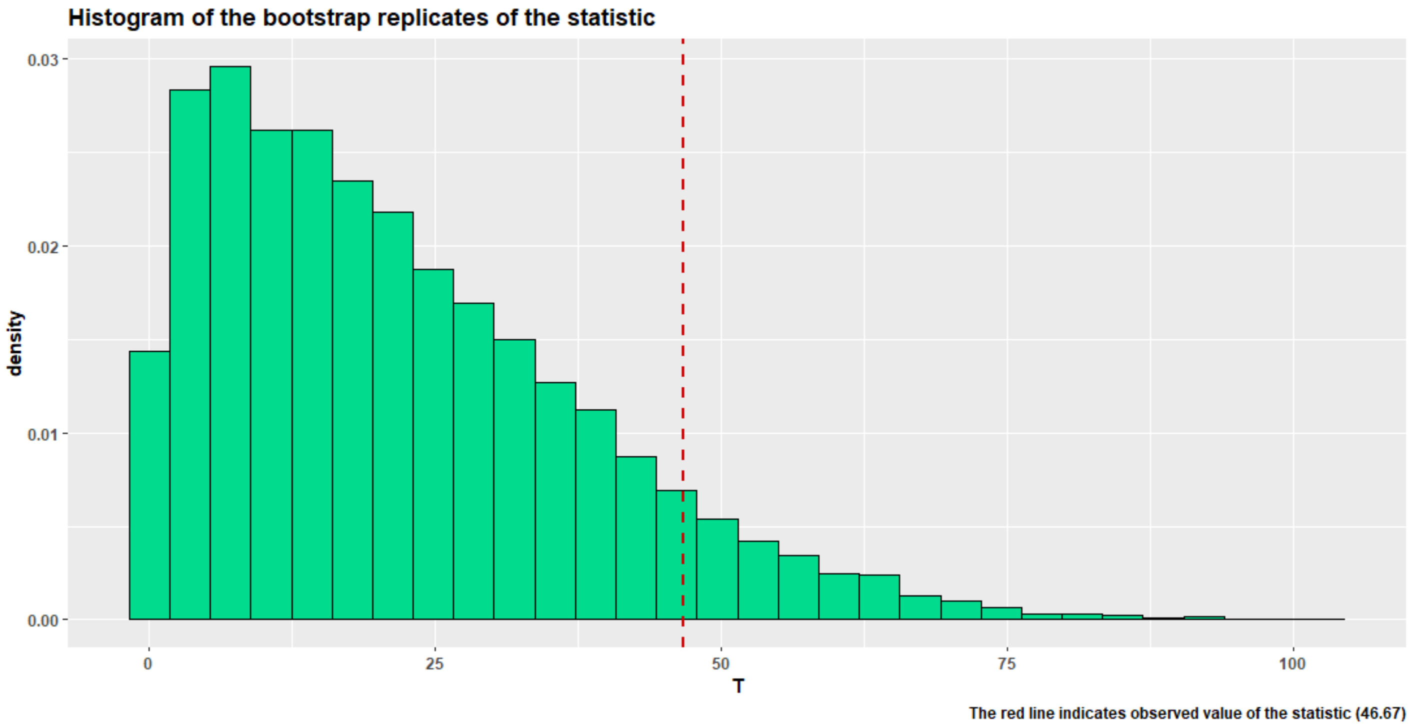
$$t(x^{*b}) = |\bar{z}^* - \bar{y}^*|, b = 1,2,...,B$$

where $\bar{z}^*$ and $\bar{y}^*$ are the means of the bootstrap samples $z^*$ and $y^*$ generated in Step 1.

**Step 3 :** The achieved significance level (ASL) is approximated as -

$$\widehat{ASL}_{boot} = \#(t(x^{*b}) \geq t_{obs})/B$$

Now, we draw a histogram of the values of the bootstrap statistic obtained from the second step of the algorithm.




From the shape of the histogram, we can see that the distribution is highly positively skewed. So, our initial decision of not assuming normality is well justified here.

Now, the ASL is obtained as 0.0862, greater than 0.05. So, our decision is to accept $H_0$ at 5% level of significance.

**Conclusion :** In the light of the given data, it seems that there is no significant difference in weight gain of the chickens by the two feed types - meatmeal and casein.

# 6 Limitations of Bootstrap

Bootstrapping depends heavily on the estimator used. So, simple, ignorant use of bootstrapping will not always yield asymptotically valid results and can lead to inconsistency. Broadly speaking, the bootstrap method is not about the actual estimation of the parameter, the purpose is to construct an approximate sampling distribution of the for the statistic under consideration. If the statistic of interest $\hat{X}_n \to X_\infty$ in distribution, as $n \to \infty$, we would assume that the bootstrap distribution to converge to the same. If not, the inferences made based on bootstrap method may not be trusted at all.

## 6.1 Example : Distribution of maximum ordered statistic of $U(0,\theta)$ distribution

An example of when the bootstrap can fail, even in an i.i.d. framework, is when trying to approximate the sampling distribution of an extreme order statistic. Here is a brief discussion on the above fact.

Let $X_1, X_2, \ldots, X_n$ be a sequence of random variables that follows $U(0,\theta)$ distribution. Consider the statistic $X_{(n)} = max_{1\leq k\leq n} X_k$. The distribution fucntion of $X_{(n)}$ is given by -

$$F_{X_n}(x) = P(X_n \leq x) = (\frac{x}{\theta})^n$$

Now, consider a function of $X_{(n)}$ as $T = n(\theta - X_{(n)})$. The distribution function is given by -

$$\begin{aligned}
F_T(x) &= P(n(\theta - X_{(n)}) \leq x) \\
&= P(X_{(n)} \geq \theta - \frac{x}{n}) \\
&= 1 - P(X_{(n)} \leq \theta - \frac{x}{n}) \\
&= 1 - (\frac{\theta - \frac{x}{n}}{\theta})^n \\
&= 1 - (1 - \frac{x}{n\theta})^n \\
&\to 1 - e^{\frac{x}{\theta}}
\end{aligned}$$

i.e. $T = n(\theta - X_{(n)})$ converges in distribution to an exponential random variable with mean $\theta$.

Now, we form a bootstrap estimate of $T$ by resampling $X_1, X_2, \ldots, X_n$ with replacement to get

$X_1^*, X_2^*, \ldots, X_n^*$ and using the distribution of $T^* = n(X_{(n)} - X_{(n)}^*)$ conditional on $X_1, X_2, \ldots, X_n$. But,

$$\begin{aligned} P(X_{(n)} = X_{(n)}^*) &= 1 - P(X_{(n)} \neq X_{(n)}^*) \\ &= 1 - P(X_{(n)}^* \leq X_{(n-1)}) \\ &= 1 - (\frac{n-1}{n})^n \\ &= 1 - (1 - \frac{1}{n})^n \\ &\to 1 - e^{-1} \end{aligned}$$

So, we can see that, though the true limiting distribution is exponential with mean $\theta$, the limiting bootstrap distribution places a point mass at zero with probability $1 - e^{-1} \approx 0.632$ independent of the actual value of $\theta$. By taking $\theta$ sufficiently large, we can make the probability of the true limiting distribution arbitrary small for any fixed interval $[0, \varepsilon)$, yet the bootstrap will still report that there is at least probability 0.632 in this interval! Let's represent the above fact graphically.

**Algorithm :**

**Step 1 :** Draw a sample of size $n$ (consider a large value of $n$, say 1000) from $U(0, \theta)$ distribution, for increasing values of $\theta$. The values of $\theta$ are taken to be 5, 10, 20, 50, 100 and 500.

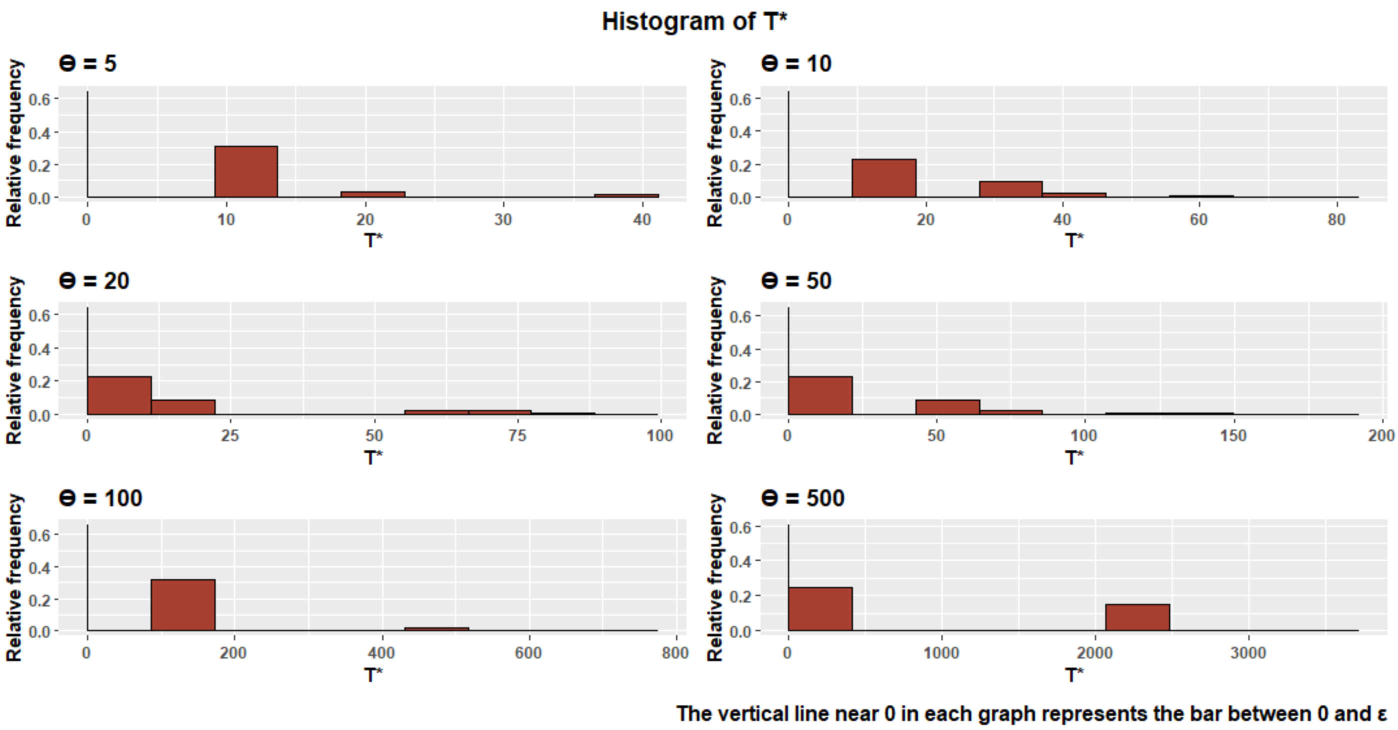
**Step 2 :** Draw $B$ (we take $B$=1000) samples of size $n$ with replacement from the original sample.

**Step 3 :** Evaluate $T^{*b} = n(X_{(n)} - X_{(n)}^*)$, $b = 1, 2, \ldots, B$ on each sample.

**Step 4 :** Draw a histogram of $T^{*b}$ for each $\theta$.

**Step 5 :** Consider $\varepsilon = 0.001$. Compute the empirical probability that $T^{*b} \in (0, \varepsilon)$ as -

$$\widehat{P}(T^{*b} \in (0, \varepsilon)) = \#(0 \leq T^{*b} \leq \varepsilon)/B$$




From the graph, it can be easily observed that the relative frequency in the interval $(0,\varepsilon)$ remains almost same with increase in the value of $\theta$. The empirical probabilities, as mentioned in the fifth step, are found to be 0.637, 0.643, 0.64, 0.658, 0.625 and 0.641 respectively for the values of $\theta$ mentioned above, i.e. the corresponding probability remains almost constant and almost equal to the value obtained mathematically.

From this it is clear that the bootstrap method behaves **arbitrarily badly** in this setting.

---

# 7 Appendix

## R Codes



Code 1: Use of Bootstrap in Point Estimation

```

#Example 1 : Distribution of sample mean of N(mu,sigma^2) distribution
    =======


library(ggplot2)

n=30 #sample size
s1=rnorm(n,2,2) #parent sample

mean(s1) #sample mean

#bootstrap samples
#-------------------
B=1000 #bootstrap number

M=matrix(0,nrow = B,ncol = n)

for(i in 1:B){
  M[i,]=sample(s1,size=n,replace = T)
}

m=apply(M,1,mean)

mean(m) #bootstrap estimate
sd(m)  #bootstrap standard error

#histogram
#----------

f=function(x){dnorm(x,2,sqrt(4/n))}
ggplot(as.data.frame(m),aes(m))+
  geom_histogram(aes(y=..density..),fill='#2596be',col='black')+
  stat_function(fun=f,aes(col='N(2,4/n)'),size=1.1)+
  labs(title="e.g. 1: Histogram of Bootstrap Sample means",
       subtitle = 'Parent Distribution : N(2,4)       Sample size(n) =
           30',
       x='sample mean',
```

```
      col='Index')+
  theme(text = element_text(face='bold'))

#=========================================
#=========================================


#Example 2 : Distribution of sample variance of N(mu,sigma^2) distribution =======


s1=rnorm(30,2,2) #parent sample

#bootstrap samples
#==================

M=matrix(0,nrow = 1000,ncol = 30)

for(i in 1:1000){
  M[i,]=sample(s1,size=30,replace = T)
}

v=apply(M,1,var)

mean(v) #bootstrap estimate
sd(v) #bootstrap standard error

T_original=2*2^4/29 #original value


#Histogram
#----------

f=function(x){dchisq(x,29)}

T=29*v/4 #statistic

ggplot(as.data.frame(T),aes(T))+
  geom_histogram(aes(y=..density..),fill='#621ddb',col='black')+
  stat_function(fun=f,aes(col='chisq(n-1)'),size=1.1)+
  labs(title=bquote("e.g. 2: Histogram of T ="~frac((n-1)*s^2, sigma^2)),
      subtitle = 'Parent Distribution : N(2,4)       Sample size(n) = 30',
      x='T',
      col='Index')+
  theme(text = element_text(face='bold'))

#==========================================
#==========================================


#Example 3 : Distribution of sample range of U(0,1) distribution =======
```

```
s1=runif(30) #parent sample

#bootstrap samples
#------------------

M=matrix(0,nrow = 1000,ncol = 30)

for(i in 1:1000){
  M[i,]=sample(s1,size=30,replace = T)
}

rn=apply(M,1,max)-apply(M,1,min)

mean(rn) #bootstrap estimate

R_original=29/31 #original value

#histogram
#----------

f=function(x){dbeta(x,29,2)}


ggplot(as.data.frame(rn),aes(rn))+
  geom_histogram(aes(rn,y=..density..),fill='#0b423a',col='black')+
  stat_function(fun=f,aes(col='Beta(n-1,2)'),size=1.1)+
  xlim(0,1)+
  labs(title='e.g. 3 : Histogram for sample range',
       subtitle = 'Parent Distribution : U(0,1)       Sample size(n) =
           30',
       x='Sample Range',
       col='Index')+
  theme(text = element_text(face='bold'))

#========================================
#========================================


#Example 4 : Distribution of sample median of DE(mu,b) distribution
   =======


library(nimble)
s1=rdexp(50) #parent sample

#bootstrap samples
#-------------------

M=matrix(0,nrow = 1000,ncol = 30)

for(i in 1:1000){
  M[i,]=sample(s1,size=30,replace = T)
```

```
}

med=apply(M,1,median)

mean(med) #bootstrap estimate
sd(med) #bootstrap standard error

#Histogram
#-----------

ggplot(as.data.frame(med),aes(med))+
  geom_histogram(aes(y=..density..),fill='#a15e00',col='black')+
  labs(title='e.g. 4 : Histogram of Sample Median',
       subtitle = 'Parent Distribution : DE(0,1)       Sample size(n)
          = 30',
       x='Sample Median')+
  theme(text = element_text(face='bold'))

#=========================================
#=========================================
```

Code 2: Types of Bootstrap (Parametric and Non-parametric)

```

#Example 1 : Distribution of sample mean of N(mu,sigma^2) distribution
    =======


library(ggplot2)

n=30 #sample size
B=1000 #bootstrap no.

s1=rnorm(n,2,2)
s1 #parent sample
t=mean(s1);t #sample mean

#parametric bootstrap
#---------------------

M=matrix(0,nrow = B,ncol = 30)

for(i in 1:B){
  M[i,]=rnorm(n,t,2)
}

m=rowMeans(M)
m

mu.hat=mean(m); mu.hat

se.hat = sd(m); se.hat

#original values
#----------------
```

```

mu = 2

se.mean = 2/sqrt(n); se.mean


#Histogram based on parametric bootstrap estimates
#-----------------------------------------------------

f=function(x){dnorm(x,mu,se.mean)}


g1=ggplot(as.data.frame(m),aes(m))+
  geom_histogram(aes(y=..density..),fill='#621ddb',col='black')+
  stat_function(fun=f,aes(col='N(2,4/n)'),size=1.1)+
  labs(title= 'Histogram of Bootstrap Sample means (Parametric
     approach)',
       subtitle = 'Parent Distribution : N(2,4)       Sample size(n) =
           30',
       x='Sample Mean',
       col='Index')+
  theme(text = element_text(face='bold'))
g1


# non-parametric bootstrap
#---------------------------

#bootstrap samples
#------------------

M1=matrix(0,nrow = B,ncol = n)

for(i in 1:1000){
  M1[i,]=sample(s1,size=n,replace = T)
}

m1=rowMeans(M1)

mu.hat.1=mean(m1); mu.hat.1

se.hat.1 = sd(m1); se.hat.1

#histogram of non-parametric bootstrap estimates
#-------------------------------------------------

g2=ggplot(as.data.frame(m1),aes(m1))+
  geom_histogram(aes(y=..density..),fill='#a15e00',col='black')+
  stat_function(fun=f,aes(col='N(2,4/n)'),size=1.1)+
  labs(title= 'Histogram of Bootstrap Sample means (Non-Parametric
     approach)',
       subtitle = 'Parent Distribution : N(2,4)       Sample size(n) =
           30',
       x='Sample Mean',
```

```
        col='Index')+
  theme(text = element_text(face='bold'))
g2


#Combining graphs
#-----------------
Parametric=m
Non_Parametric=m1
df=data.frame(Parametric,Non_Parametric)
d=stack(df)
colnames(d)=c('Sample_mean','Index')

ggplot(d,aes(Sample_mean))+
  geom_histogram(aes(y=..density..,fill=Index),col=1)+
  stat_function(fun=f,aes(col='1'),size=1.1)+
  facet_wrap(~Index,scales='free_x')+
  labs(title= 'Histogram of Bootstrap Sample means',
        subtitle = 'Parent Distribution : N(2,4)       Sample size(n) =
            30',
        x='Sample Mean',
        col='Line Index')+
  theme(text = element_text(face='bold'))+
  scale_color_manual(values = 1, label = 'N(2,4/n)')


#========================================
#========================================


#Example 2 : Distribution of sample median from DE(mu,b) distribution
   =======


library(nimble)


s1=rdexp(30) #parent sample
s1

m=median(s1) #sample median
m

#non parametric bootstrap
#-------------------------

M=matrix(0,nrow = 1000,ncol = 30)

for(i in 1:1000){
  M[i,]=sample(s1,size=30,replace = T)
}

med=apply(M,1,median)

```

```
med_np=mean(med);med_np
se_np=sd(med);se_np


#parametric bootstrap
#---------------------

M1=matrix(0,nrow = 1000,ncol = 30)

for(i in 1:1000){
  M1[i,]=rdexp(30,m,1)
}

med1=apply(M1,1,median)

med_p=mean(med1);med_p
se_p=sd(med1);se_p


#Histograms
#-----------
Parametric=med
Non_Parametric=med1
df=data.frame(Parametric,Non_Parametric)
d=stack(df)
colnames(d)=c('Sample_median','Index')

ggplot(d,aes(Sample_median))+
  geom_histogram(aes(y=..density..,fill=Index),col=1)+
  facet_wrap(~Index,scales='free_x')+
  labs(title= 'Histogram of Bootstrap Sample medians',
       subtitle = 'Parent Distribution : DE(0,1)      Sample size(n) = 30',
       x='Sample Median',
       col='Index')+
  theme(text = element_text(face='bold'))

#=========================================
#=========================================
```

Code 3: Interval Estimation using Bootstrap

```

#Example 1 : Confidence interval for mu in N(mu,sigma^2 distribution)
#=====================================================================

n=30 #sample size
s1=rnorm(n,2,2) #parent sample

t=mean(s1);t #sample mean

#basic - t confidence interval
#-----------------------------

CI=mean(s1)+qt(0.975,n-1)*sd(s1)*c(-1,1)/sqrt(n)
```

```
CI


#bootstrap samples
#------------------

M=matrix(0,nrow = 1000,ncol = 30)

for(i in 1:1000){
  M[i,]=sample(s1,size=30,replace = T)
}

m=rowMeans(M)


#bootstrap percentile confidence interval
#----------------------------------------

CI2=quantile(m,c(0.025,0.975))
CI2

#bias corrected and accelarated (BCa) confidence interval
#--------------------------------------------------------

z0=qnorm(mean(m<t))
z0 #bias

m_i=0
for(i in 1:n)
m_i[i]=mean(s1[-i])


m.=mean(m_i)

a= sum((m.-m_i)^3)/(6*(sum((m.-m_i)^2))^(3/2))

p1=pnorm(z0+(z0+qnorm(0.025))/(1-a*(z0+qnorm(0.025))))
p2=pnorm(z0+(z0+qnorm(0.975))/(1-a*(z0+qnorm(0.975))))

CI3=quantile(m,c(p1,p2))
CI3


#=====================================
#=====================================


#Example 2 : Confidence Intervals for mu in DE(mu, b) distribution
#===============================================================

library(nimble)

n=30 #sample size
s1=rdexp(n) #parent sample
```

```
t=median(s1);t #sample median


#basic - t confidence interval
#------------------------------

CI=median(s1)+qt(0.975,n-1)*mad(s1)*c(-1,1)/sqrt(n)
CI


#Bootstrap Samples
#------------------

M=matrix(0,nrow = 1000,ncol = 30)

for(i in 1:1000){
  M[i,]=sample(s1,size=30,replace = T)
}

m=apply(M,1,median)


#bootstrap percentile confidence interval
#-----------------------------------------

CI2=quantile(m,c(0.025,0.975))
CI2

#bias corrected and accelarated (BCa) confidence interval
z0=qnorm(mean(m<t))
z0 #bias

m_i=0
for(i in 1:n)
  m_i[i]=median(s1[-i])


m.=mean(m_i)

a= sum((m.-m_i)^3)/(6*(sum((m.-m_i)^2))^(3/2))

p1=pnorm(z0+(z0+qnorm(0.025))/(1-a*(z0+qnorm(0.025))))
p2=pnorm(z0+(z0+qnorm(0.975))/(1-a*(z0+qnorm(0.975))))

CI3=quantile(m,c(p1,p2))
CI3

#======================================
#======================================


#Example 4 - Confidence interval of theta in exp(theta) distribution
   =======
```

```
n=20#sample size
s1=rexp(n)#parent sample
s1

t=median(s1)/log(2);t #sample mean

#bootstrap samples
#------------------
B=1000 #bootstrap number
M=matrix(0,nrow = B,ncol = n)

for(i in 1:B){
  M[i,]=sample(s1,size=n,replace = T)
}

T=apply(M,1,median)/log(2)
T

#bootstrap percentile confidence interval
#----------------------------------------

CI2=quantile(T,c(0.025,0.975))
CI2

#bias corrected and accelarated (BCa) confidence interval
#-------------------------------------------------------

z0=qnorm(mean(T<t))
z0 #bias

m_i=0
for(i in 1:n)
  m_i[i]=median(s1[-i])/log(2)


m.=mean(m_i)

a= sum((m.-m_i)^3)/(6*(sum((m.-m_i)^2))^(3/2))

p1=pnorm(z0+(z0+qnorm(0.025))/(1-a*(z0+qnorm(0.025))))
p2=pnorm(z0+(z0+qnorm(0.975))/(1-a*(z0+qnorm(0.975))))

CI3=quantile(m,c(p1,p2))
CI3

library(ggplot2)

ggplot(data.frame(m),aes(m))+
  geom_histogram(aes(y=..density..),col=8,fill=1)+
  geom_vline(aes(xintercept = CI2[1],col='Percentile'),size=1.05)+
  geom_vline(aes(xintercept = CI2[2],col='Percentile'),size=1.05)+
  geom_vline(aes(xintercept = CI3[1],col='BCa'),size=1.05)+
  geom_vline(aes(xintercept = CI3[2],col='BCa'),size=1.05)+
```

```
  geom_vline(aes(xintercept = t[1],col='Observed'),size=1.05)+
  geom_vline(aes(xintercept = t[2],col='Observed'),size=1.05)+
  labs(title='Histogram of T* = Me*/log(2)',
       subtitle='Distribution : Exp(1)',
       x='T*',
       col='Index')+
  theme(text = element_text(face='bold'))
```

Code 4: Testing of Hypothesis using Bootstrap

```

#Testing for equality of means of two population
#================================================

library(ggplot2)
library(knitr)


data=chickwts[chickwts$feed=='meatmeal'|chickwts$feed=='casein',]
View(data)

kable(data, "latex")


n=nrow(data[data$feed=='meatmeal',])
n
m=nrow(data[data$feed=='casein',])
m


#boxplot of weight for different feed types
#------------------------------------------

ggplot(data,aes(feed,weight))+
  geom_boxplot(aes(fill=feed))+
  labs(title = 'Boxplot of weight gained for different feed types',
        fill='Index')+
  theme(text = element_text(face='bold'))


m.feed=aggregate(data$weight,by=list(data$feed),FUN=mean)$x
m.feed        #average weight for different feed types

g.mean=mean(m.feed)
g.mean  #gropued mean

T=abs(m.feed[2]-m.feed[1])
T #test statistic

N=n+m
B=10000 #bootstrap number

w=data$weight

#bootstrap replicates of the statistic
```

```
#--------------------------------------
T1=0

bt=matrix(sample(w,size = N*B,replace = T),nrow=N,ncol=B)

for(i in 1:B){
  T1[i]=abs(mean(bt[1:11,i])-mean(bt[12:23,i]))
}
T1

#histogram of the test statistic
#-------------------------------
ggplot(data.frame(T1),aes(T1))+
  geom_histogram(aes(y=..density..),fill='#00db8e',col='black')+
  geom_vline(xintercept=T,col='red3',lwd=1,lty=2)+
  labs(title = 'Histogram of the bootstrap replicates of the statistic
     ',
       x='T',
       caption = 'The red line indicates observed value of the
          statistic (46.67)')+
  theme(text = element_text(face='bold'))

#achieved significance level
#----------------------------
ASL=mean(T1>=T);ASL
```

Code 5: Limitations of Bootstrap

```

#Distribution of maximum order statistic of U(0,theta) distribution
#==================================================================

library(ggplot2)
library(patchwork)

n=1000 #sample size
B=1000 #bootstrap number
eps=0.001


#function to draw histogram of T* and determine empiricial probability
#----------------------------------------------------------------------
f=function(theta){
  s1=runif(n,0,theta)#parent sample

  M=matrix(0,nrow = B,ncol = n)
  for(i in 1:B){
    M[i,]=sample(s1,size=n,replace = T)
  }#bootstrap samples

  T1=n*(max(s1)-apply(M,1,max))#statistic

  g=ggplot(data.frame(T1),aes(T1))+
    stat_bin(aes(T1,y=stat(count) / sum(count)),fill='#a84032',col='
      black',breaks = c(0,seq(eps,max(T1),len=10)))+
```

```
    labs(title=paste('\u03f4 =',theta),
         x='T*',
         y='Relative frequency')#histogram

  p=length(T1[T1>=0 & T1<=eps])/B #empirical probability

  return(list(p,g))
}


#histogram for different values of theta
#----------------------------------------
g1=f(5)[[2]]
g2=f(10)[[2]]
g3=f(20)[[2]]
g4=f(50)[[2]]
g5=f(100)[[2]]
g6=f(500)[[2]]


#final graph
#------------
(g1+g2)/(g3+g4)/(g5+g6)+plot_annotation(title = 'Histogram of T*',
       caption = 'The vertical line near 0 in each graph represents the bar between 0 and \u03b5',
       theme = theme(plot.title = element_text(size = 14,hjust=0.5),
       plot.caption=element_text(size=12)))&
  theme(text = element_text(face = 'bold'))
```

# 8 Acknowledgement

I would like to express my sincere gratitude to my supervisor Prof. Dr. Durba Bhattacharya for providing an opportunity to work on this project. Her guidance and invaluable advice helped me a lot during the course of the project.

# References

[1] Michael R Chernick. *Bootstrap methods: A guide for practitioners and researchers*. John Wiley & Sons, 2011.

[2] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.