# Applied Probability II

## Section 8: The Bootstrap

Professor Caroline Brophy

Semester 2, 2020-21

# Applied Probability II. Section 8: The Bootstrap

# Section 8.1: The concept of bootstrapping

# Introduction

So far, we have generally assumed a sample $X_1, ..., X_n$ to arise from some known distribution. That means that outside our data observations, we make additional assumptions about the shape of the underlying distribution, for example:

$$X_1, ..., X_n \sim \text{IID Normal}(\mu, \sigma^2)$$

or

$$X_1, ..., X_n \sim \text{IID Poisson}(\lambda)$$

The bootstrap is based on the idea that, without further information about the underlying distribution, the observed sample $x_1, ..., x_n$ contains all available information about $F_X(t)$, and hence the actual distribution of the data.

Essentially with bootstrapping, we resample the sample and consider it to be a sample from the population of interest.

# Sampling distributions

Suppose we use a sample $X_1, ..., X_n$ to estimate some unknown $\theta$ from the true distribution of $X$. This could be the mean or the variance.

$$\hat{\theta} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \hat{\theta} = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

In both of these cases, we can write the estimate as a function of the sample:

$$\hat{\theta} = h(X_1, ..., X_n)$$

I.e., the definition of a statistic.

In earlier sections of the module, we used $\bar{X}$ as an estimator of $\mu$ the population mean and derived its sampling distribution:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

Using the sampling distribution, we constructed CIs and hypothesis tests for $\mu$.

With bootstrapping, we construct the sampling distribution by repeatedly resampling the observed data and evaluating $\hat{\theta}$ for each of these samples.

# Overview of the bootstrap approach

When the underlying population of $X_1, ..., X_n$, call it $F_X$, is not known, we can find the sampling distribution of $\hat{\theta}$ using bootstrapping.

The steps are:

1. Consider $X_1, ..., X_n$ from distribution $F_X$.
2. $\theta$ is the parameter of interest.
3. Sample $n$ values from $X_1, ..., X_n$ **with replacement**, giving $X_1^*, ..., X_n^*$, a bootstrap sample.
4. Compute the bootstrap estimator
$$\hat{\theta}^* = h(X_1^*, ..., X_n^*)$$
5. Repeat the previous two steps B times. Index the samples by $b = 1, ..., B$: $X_1^{*(b)}, ..., X_n^{*(b)}$ and similarly index the bootstrap statistic estimate: $\hat{\theta}^{*(b)}$.
6. We now have B bootstrapped samples and B $\hat{\theta}^*$ values.

The distribution of the $\hat{\theta}^{*(b)}$'s approximates the sampling of $\hat{\theta}$ under $F_X$.

The bootstrap is a very general method and relies on no assumptions about $F_X$, the actual distribution of $X$.

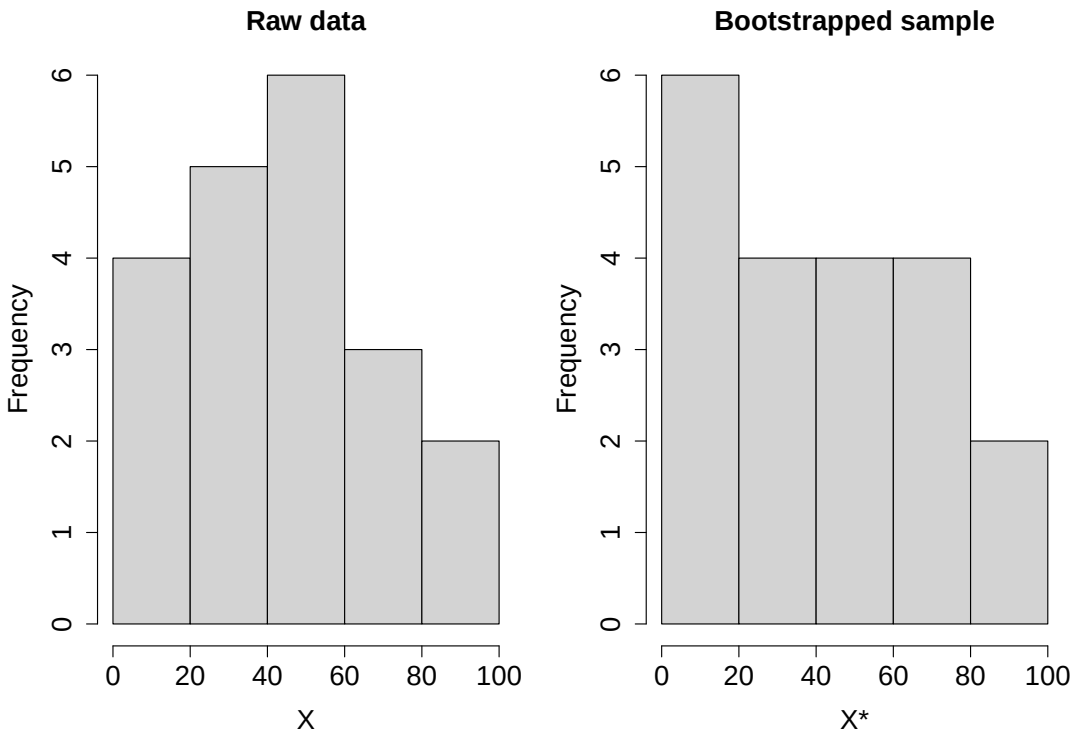
# Example of a bootstrap sample

```
Raw_data
```

```
##  [1]  8 10 15 19 28 35 36 38 39 41 43 44 49 54 57 62 69 75 85 95
```

```
Bootstrapped_sample
```

```
##  [1]  8 10 10 15 15 15 28 35 38 39 41 41 41 49 62 62 69 75 95 95
```

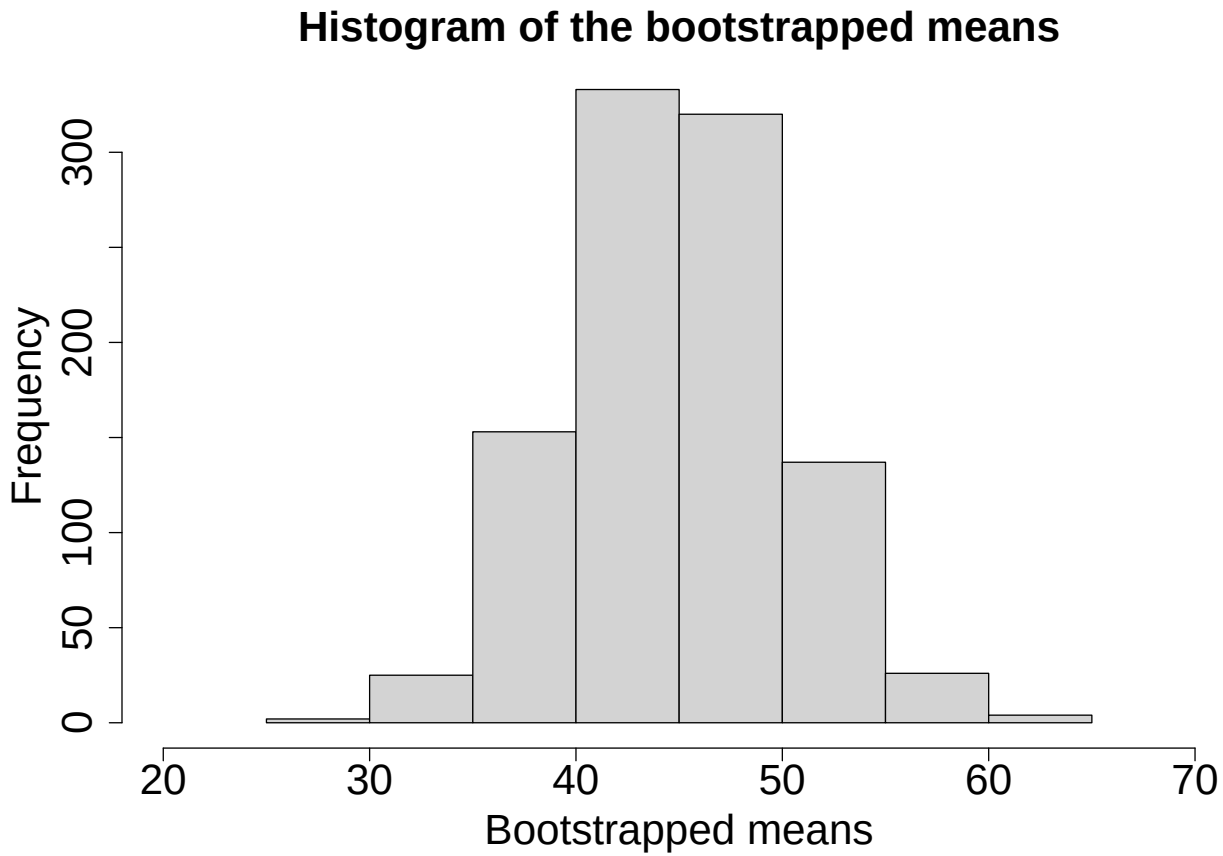



The mean of the raw data is 45.1, and the median of the bootstrapped sample is 42.15.

# Sampling distribution

Continuing with the previous example, the mean from the raw data is 45.1 and the mean of the bootstrapped sample is 42.15.

Let's take 1000 bootstrap samples and compute the mean each time.



# Fun fact about bootstrapping

The method is called the 'bootstrap', to suggest pulling oneself up by the bootstraps (as an example of an impossible task!).

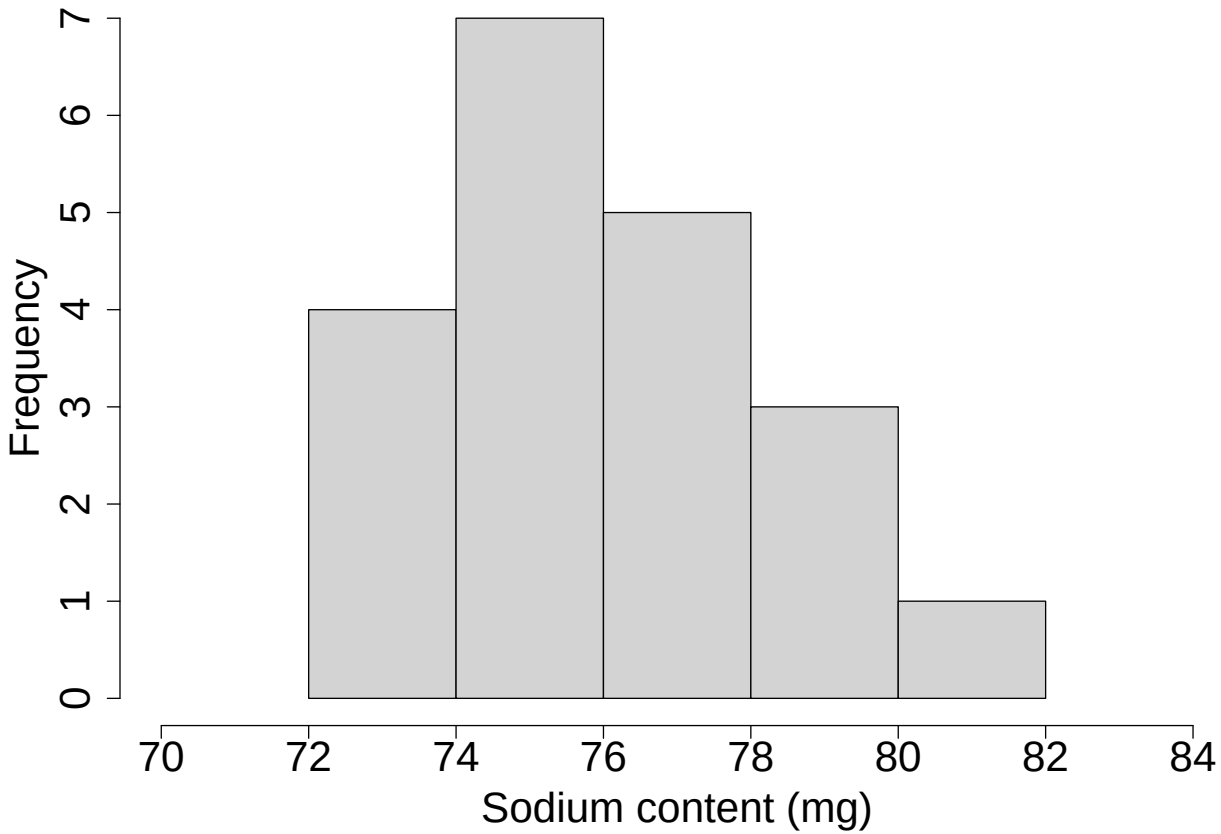
# Section 8.2: Bootstrap standard error for the median

# Bootstrap application

Let's have a look at a practical way that bootstrapping can be used.

**Example**

The sodium content (mg) of 20 fast food products was recorded. The dataset:



# Sodium example contd.

We can easily find the median of the Sodium data: 75.35.

We will now use bootstrapping to find the standard error of the median.

Here are the steps:

- Take the original sample of data. Note the sample size $n$.
- Sample $n$ values from it **with replacement** B times. These are called the bootstrapped samples.
- Compute the median for each bootstrapped sample and denote $M^{*(i)}$ for $i = 1, ..., B$.
- Create a histogram of the B $M^{*(i)}$ values. (Useful but not required.)
- Compute the standard deviation of $M^{*(1)}, \ldots, M^{*(B)}$. This gives us our standard error.

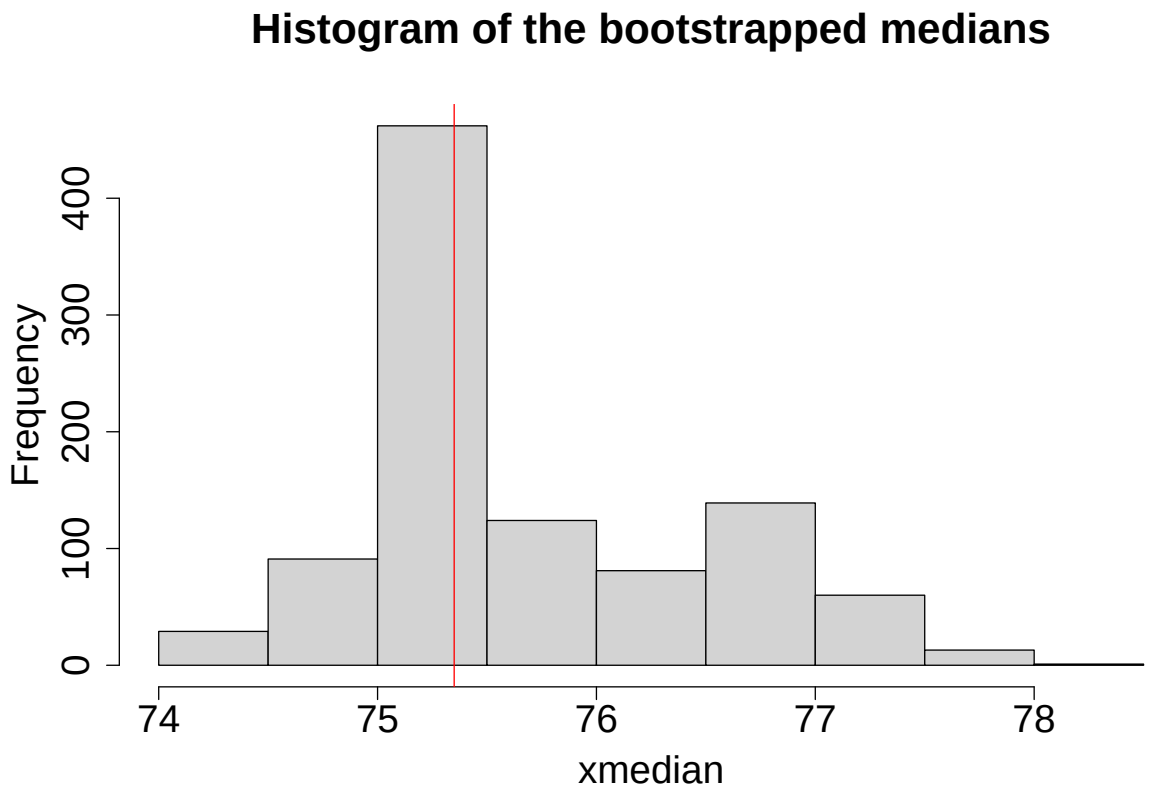
This has been implemented using R.

Here are the first 20 medians:

```
##  [1] 75.95 75.20 74.95 75.15 75.15 75.30 75.80 75.10 75.30 75.35 77.00 75.35
## [13] 74.20 77.10 75.20 75.95 75.10 77.25 75.95 75.15
```

# Sodium example contd.

Here are all the bootstrapped medians in a histogram, with the median of the Sodium dataset (= 75.35) highlighted in red:




The standard deviation across the bootstrapped medians = 0.7922.

# Sodium example contd.

We have found the standard error for the median for the Sodium dataset using the bootstrap.

Summary of our process:

- The original Sodium dataset had sample size $n = 20$.
- We sampled 20 values with replacement from the Sodium dataset 1000 times to create 1000 bootstrapped datasets.
- We computed the median for each of the 1000 bootstrapped datasets.
- We examined a histogram of the 1000 bootstrapped medians.
- We computed the standard deviation of the 1000 bootstrapped medians.

Our median for the Sodium dataset was 75.35.

Our bootstrapped standard error for the median was 0.7922.

# Section 8.3: Bootstrap confidence intervals

# Introduction

We can use bootstrapping to construct confidence intervals.

Here we will illustrate for the population median, but other statistics of interest could also be used.

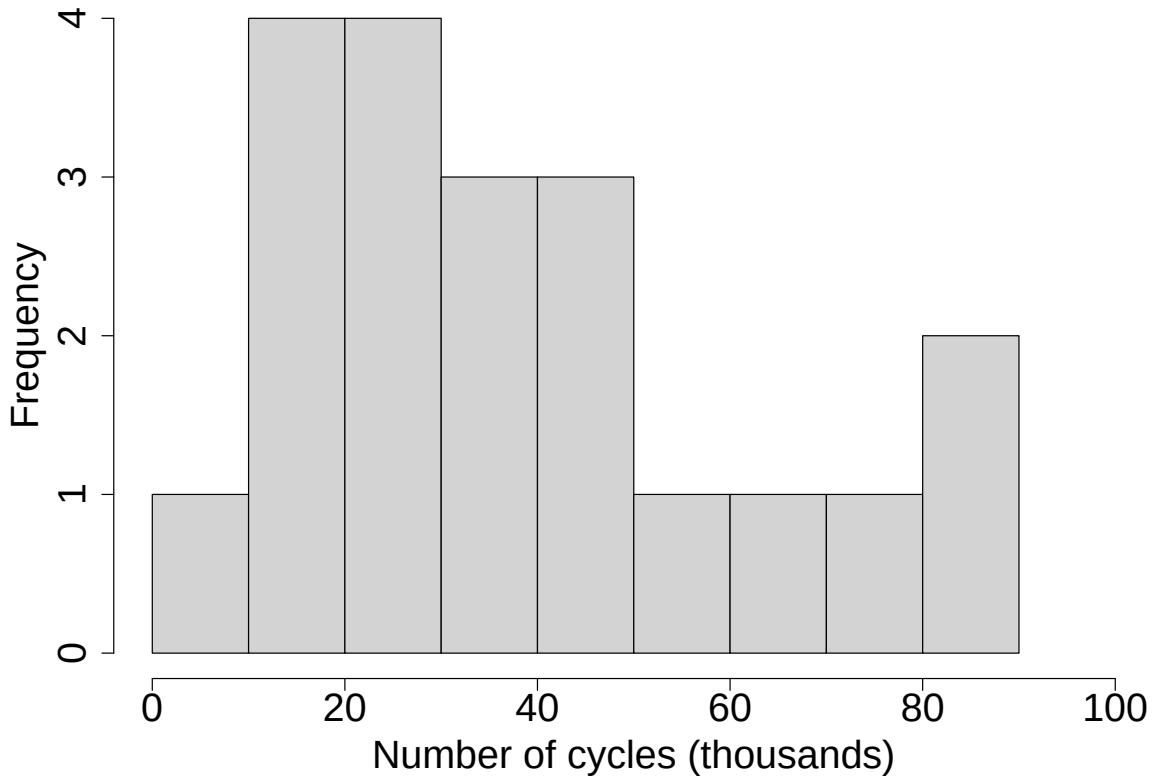
Let $\theta$ be the population median. Suppose we have a sample $x_1, ..., x_n$ from the population. We can find the estimate of the median $\hat{\theta}$ from this sample. We can then use bootstrap methods to find a confidence interval for the median.

# Example

Testing of electrical and mechanical devices often involves an action such as turning a device on and off or opening and closing a device many times. The number of open-close cycles (in thousands) that it took 20 door latches to fail was recorded.

Here is the dataset and a histogram of it.

```
##   [1]   7  11  15  16  20  22  24  25  29  33  34  37  41  42  49  57  66  71  84  90
```



# Steps to construct a bootstrap confidence interval

Let $\theta$ be the population median. Let $\hat{\theta}$ be the sample median for a sample of size $n$.

Here are the steps to construct a bootstrap confidence interval for $\theta$:

- From the original sample of size $n$, sample values with replacement B times. These are called the bootstrap samples.
- Compute the median for each bootstrap sample and denote $M^{*i}$.
- Create a histogram of the $M^{*i}$ values. (Useful but not required.)
- Order the $M^{*i}$ values and denote each ordered value by $M^{*(i)}$. So for example, $M^{*(1)}$ is the lowest and $M^{*(B)}$ the highest of the bootstrap sample median estimates.
- Let $a = (\alpha/2) * B$. The ordered bootstrap median estimates are $M^{*(1)}, M^{*(2)}, ..., M^{*(B)}$. Then $(M^{*(a)}, M^{*(B-a)})$ is an approximate $(1 - \alpha) * 100\%$ confidence interval for $\theta$.

# Example Step 1 and 2

Here is the raw data again:

7 11 15 16 20 22 24 25 29 33 34 37 41 42 49 57 66 71 84 90

The estimate of the median is: 33.5. Let's now construct a confidence interval.

**Step 1**: From the original sample of size $n = 20$, sample values with replacement B = 1000 times.

Here are the first five bootstrap samples down to the 1000th:

37 66 16 33 42 7 41 29 20 41 16 66 29 71 24 37 16 42 29 15
20 42 90 25 71 15 90 20 84 20 15 7 90 37 25 37 15 33 22 66
16 49 71 22 66 16 49 16 22 16 25 16 20 41 25 57 15 84 22 22
29 20 37 22 15 37 66 42 42 11 42 7 29 15 25 20 25 66 34 42
84 42 71 33 66 20 15 34 66 41 90 29 15 42 37 24 49 84 71 22

⋮

33 33 71 25 33 25 57 29 20 29 37 71 29 16 42 15 66 37 42 71

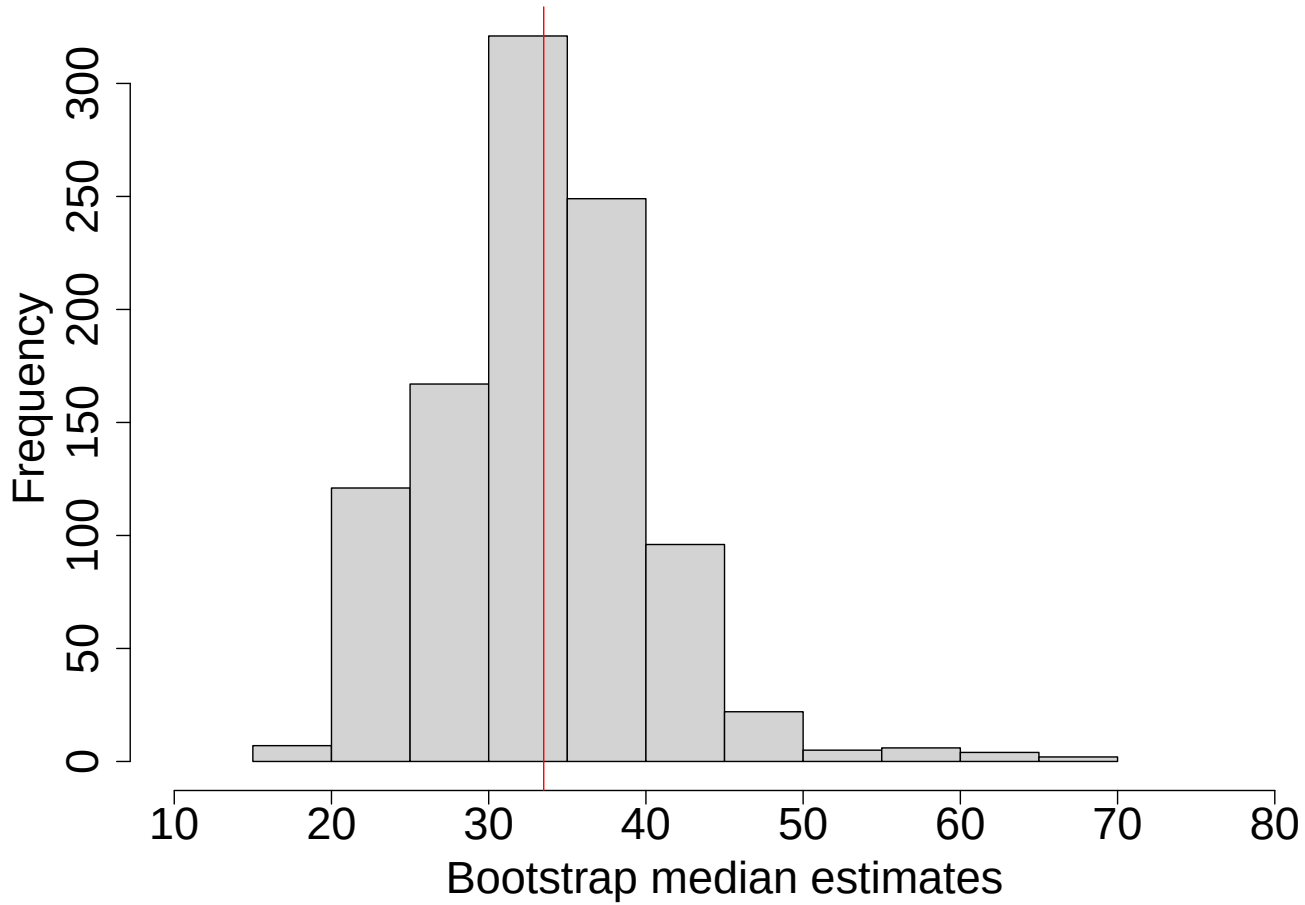
**Step 2**: Compute the median for each bootstrap sample and denote $M^{*i}$.

The medians $M^{*1}, M^{*2}, M^{*3}, M^{*4}, M^{*5}, ..., M^{*1000}$ are:

31, 29, 22, 29, 41.5, . . . , 33.

# Example Step 3

**Step 3**: Create a histogram of the 1000 $M^{*i}$ bootstrap median estimates. The red lines shows the sample median.



# Example Step 4

**Step 4**: Order the $M^{*i}$ values and denote each ordered value by $M^{*(i)}$. So for example, $M^{*(1)}$ is the lowest and $M^{*(B)}$ the highest of the bootstrap sample median estimates.

Here are the ordered values (1000 in total, showing first 50 and last 50):

```
 [1] 19.0 20.0 20.0 20.0 20.0 20.0 20.0 21.0 21.0 22.0 22.0 22.0 22.0 22.0 22.0
[16] 22.0 22.0 22.0 22.0 22.0 22.5 22.5 22.5 23.0 23.0 23.0 23.0 23.0 23.0 23.0
[31] 23.0 23.0 23.0 23.0 23.0 23.0 23.0 23.0 23.0 23.0 23.0 23.0 23.5 23.5 23.5
[46] 23.5 23.5 23.5 23.5 24.0
```

⋮

```
 [1] 42.0 42.0 42.0 42.0 42.0 42.0 42.0 42.0 43.0 45.0 45.0 45.0 45.5 45.5 45.5
[16] 45.5 45.5 45.5 45.5 45.5 45.5 45.5 45.5 45.5 45.5 49.0 49.0 49.0 49.0 49.0
[31] 49.0 49.5 49.5 49.5 53.0 53.0 53.0 53.0 53.0 57.0 57.0 57.0 57.0 57.0 57.5
[46] 61.5 61.5 61.5 61.5 66.0 66.0
```

# Step 5

**Step 5**: Let $a = (\alpha/2) * B$. The ordered bootstrap median estimates are $M^{*(1)}, M^{*(2)}, ..., M^{*(B)}$. Then $(M^{*(a)}, M^{*(B-a)})$ is an approximate $(1 - \alpha) * 100\%$ confidence interval for $\theta$.

From the ordered medians, we want to pull out the values $(M^{*(a)}, M^{*(B-a)})$, where $a = (\alpha/2) * B$.

Let's construct a 95% confidence interval.

Then, $a = (0.05/2) * 1000 = 25$ and $B - a = 1000 - 25 = 975$, so we want to pull out the 25th and the 975th medians from the ordered vector of bootstrap medians.

# Step 5

Let's go back to the ordered medians.

We want the 25th and the 975th values from the ordered vector of bootstrap medians:

```
 [1] 19.0 20.0 20.0 20.0 20.0 20.0 20.0 21.0 21.0 22.0 22.0 22.0 22.0 22.0 22.0
[16] 22.0 22.0 22.0 22.0 22.0 22.5 22.5 22.5 23.0 23.0 23.0 23.0 23.0 23.0 23.0
[31] 23.0 23.0 23.0 23.0 23.0 23.0 23.0 23.0 23.0 23.0 23.0 23.0 23.5 23.5 23.5
[46] 23.5 23.5 23.5 23.5 24.0
```

⋮

```
 [1] 42.0 42.0 42.0 42.0 42.0 42.0 42.0 42.0 43.0 45.0 45.0 45.0 45.5 45.5 45.5
[16] 45.5 45.5 45.5 45.5 45.5 45.5 45.5 45.5 45.5 45.5 49.0 49.0 49.0 49.0 49.0
[31] 49.0 49.5 49.5 49.5 53.0 53.0 53.0 53.0 53.0 57.0 57.0 57.0 57.0 57.0 57.5
[46] 61.5 61.5 61.5 61.5 66.0 66.0
```

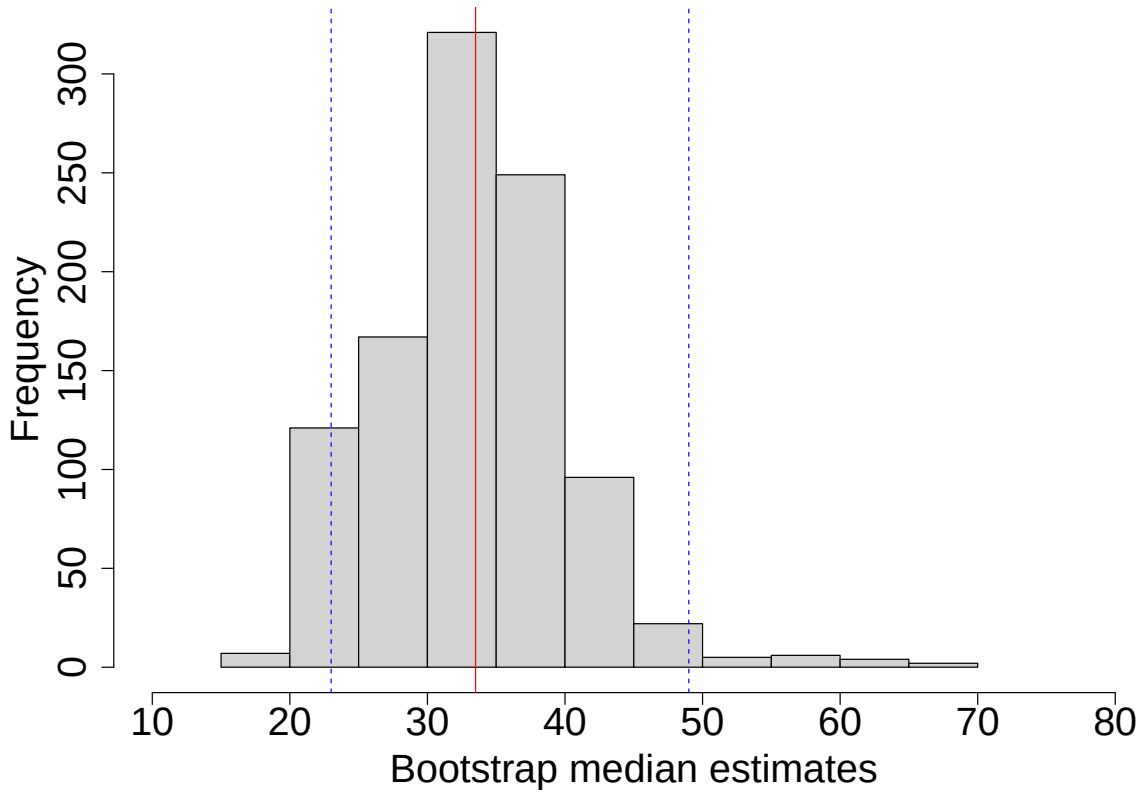
Our 95% CI is: (23, 49)

# The confidence interval graphically

The estimated median (for the original sample of data) is 33.5.

The 95% confidence interval for the population median is (23, 49).

Graphically:



# Example summary

The 95% confidence interval for the population median is (23, 49).

We are 95% confident that the true population median number of cycles (in thousands) that it takes for the door latch to fail lies between 23 and 49.

There are alternative bootstrap methods for constructing the CI that will adjust for potential biases.

# Section 8.4: Bootstrap hypothesis tests

# Hypothesis testing

We can use bootstrapping to test hypotheses of interest.

**NB**

In bootstrap hypothesis testing, the resampling is conducted under the conditions that ensure the null hypothesis, $H_0$, is true.

We will consider the 1-sample example, where we may wish to test the hypothesis:

$$H_0\text{: } \theta = \theta_0 \text{ versus } H_A\text{: } \theta \neq \theta_0$$

Where the parameter of interest may be the median or the mean as a measure of location.

# Example

A set of data of size $n = 25$ was collected in a study. It was believed that the mean of the population from which it came from was $> 1$.

We will use bootstrapping to test the hypothesis:

$$\mathrm{H}_0\colon\ \mu = \mu_0 \text{ versus } \mathrm{H}_A\colon\ \mu > \mu_0$$

where $\mu_0 = 1$ in this case.

The steps in general to follow are:

- Let $x_1, \ldots, x_n$ denote the observed data. Let $\hat{\mu} = \bar{x}$ be an estimate of $\mu$.
- Sample with replacement from $x_1 - \bar{x} + \mu_0, \ldots, x_n - \bar{x} + \mu_0$. This will ensure that the null hypothesis is true.
- Compute $\bar{x}^{*i}$, the mean for each bootstrap sample. It is useful to generate a histogram of these bootstrap estimates.
- Compute the p-value:

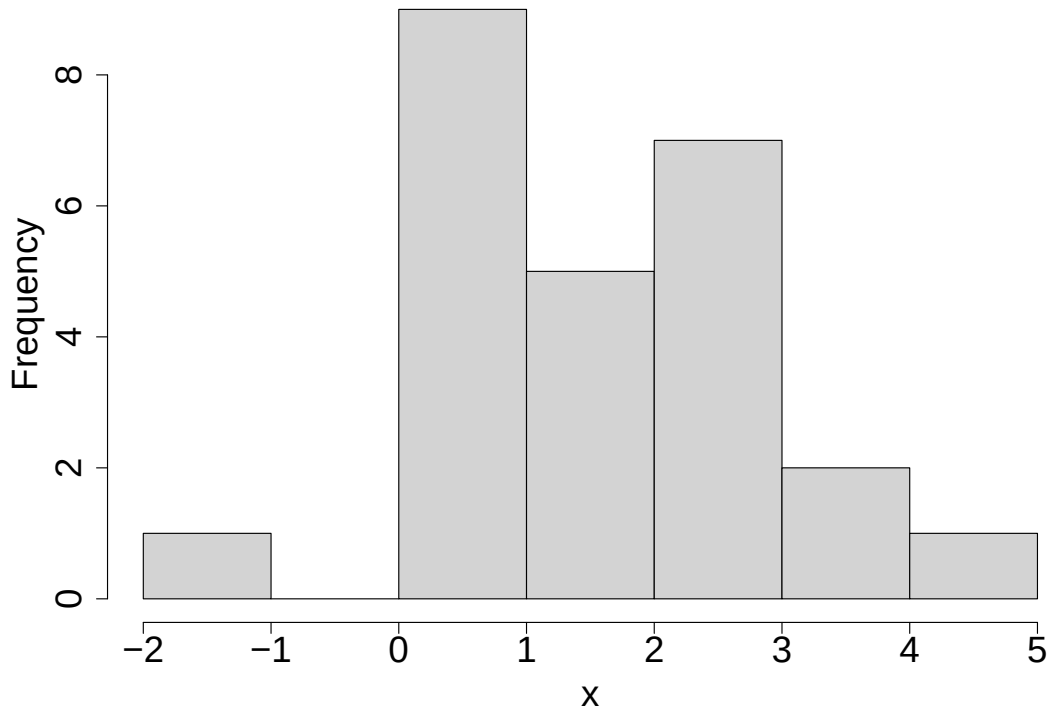
$$\text{p-value} = \frac{\#(\bar{x}^{*i} \geq \bar{x})}{B}$$

# Example step 1

**Step 1**: Let $x_1, \ldots, x_n$ denote the observed data. Let $\hat{\mu} = \bar{x}$ be an estimate of $\mu$.

Here is the raw data and a histogram of it:

```
 [1]  0.620  2.300  0.710  2.300  0.800  3.200  2.600  2.100  2.400  1.500
[11]  0.900  0.830  0.067  1.900  2.200  0.018  1.800  2.400  1.300 -1.100
[21]  4.600  1.600  0.380  3.800  0.500
```




The sample mean is: $\hat{\mu} = \bar{x} = 1.589$.

# Example step 2

**Step 2**: Sample with replacement from $x_1 - \bar{x} + \mu_0, \ldots, x_n - \bar{x} + \mu_0$. This will ensure that the null hypothesis is true.

We are testing the hypothesis

$$\mathrm{H}_0\colon\ \mu = 1 \text{ versus } \mathrm{H}_A\colon\ \mu > 1$$

Therefore, we compute $x_i - \bar{x} + \mu_0 = x_i - 1.589 + 1$, which gives:

```
 [1]  0.031  1.711  0.121  1.711  0.211  2.611  2.011  1.511  1.811  0.911
[11]  0.311  0.241 -0.522  1.311  1.611 -0.571  1.211  1.811  0.711 -1.689
[21]  4.011  1.011 -0.209  3.211 -0.089
```

We sample from this with replacement B = 5000 times giving:

```
 [1] -1.689 -0.571 -0.209 -0.089  0.031  0.121  0.121  0.311  0.711  0.911
[11]  0.911  1.011  1.011  1.011  1.211  1.311  1.711  1.711  1.811  1.811
[21]  1.811  2.011  2.611  4.011  4.011
```

⋮

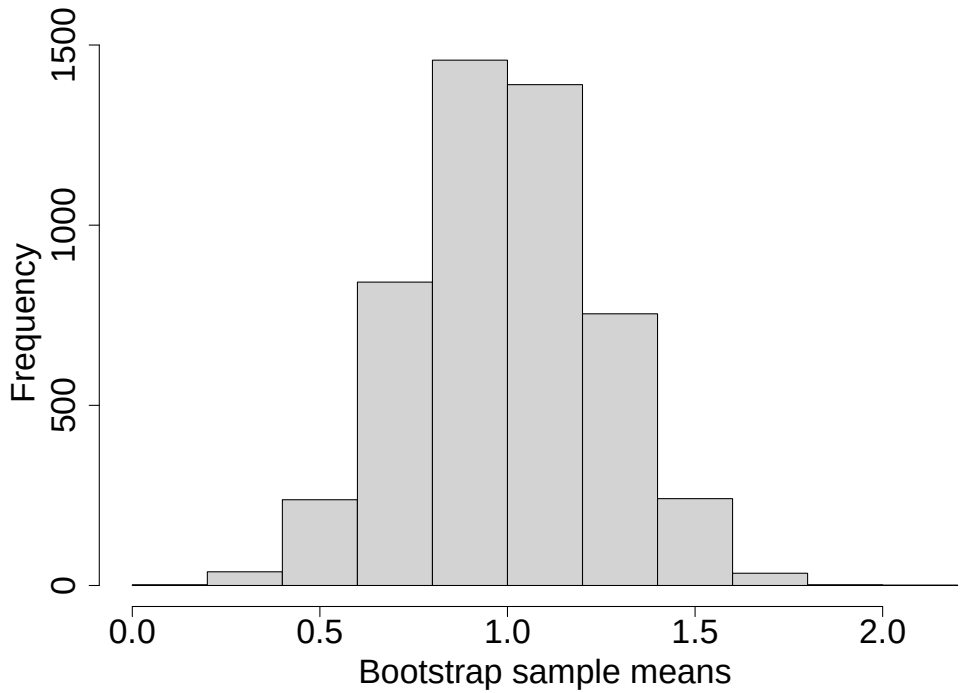
```
 [1] -1.689 -0.089  0.121  0.211  0.211  0.211  0.241  0.311  0.711  0.711
[11]  1.011  1.211  1.311  1.511  1.611  1.611  1.711  1.811  1.811  2.011
[21]  2.611  2.611  3.211  3.211  4.011
```

# Example step 3

**Step 3**: Compute $\bar{x}^{*i}$, the mean for each bootstrap sample. It is useful to generate a histogram of these bootstrap estimates.

The means for the first 15 bootstrap samples and a histogram of all bootstrap means:

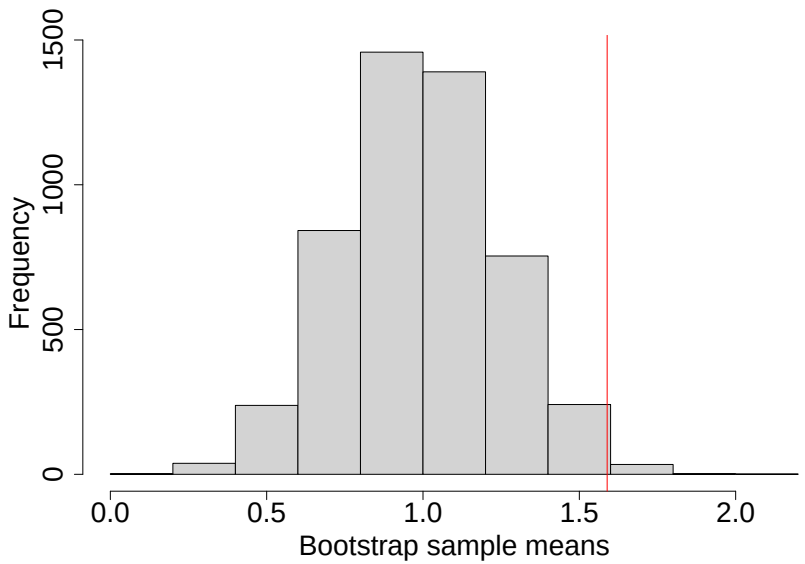
```
 [1] 1.10452 1.01788 1.27968 1.20768 1.31568 1.14280 1.04036 0.91676 0.86436
[10] 1.34540 1.20420 0.96940 0.80752 0.86052 1.03940 1.49180 1.33900 0.93840
[19] 1.06348 0.95316
```



# Example step 4

**Step 4**: Compute the p-value: p-value $= \frac{\#(\bar{x}^{*i} \geq \bar{x})}{B}$

Graphically (with the observed mean highlighted by a red line):




The number of $\bar{x}^{*i} > 1.589 = 46$. Therefore the p-value $= 46 \ / \ 5000 = 0.0092$.

Notes on the p-value calculation:

- Remember a p-value is the probability of getting a value as extreme or more extreme than what was observed assuming that the null hypothesis is true.
- We have used bootstrapping to approximate the sampling distribution of the mean under the null hypothesis.

# Example summary

$H_0$: $\mu = 1$ versus $H_A$: $\mu > 1$

$\hat{\mu} = 1.589$.

P-value $= 0.0092$.

We reject the $H_0$ and conclude that the true population mean is greater than 1.