# 3 Bootstrapping Dependent Data

Suppose we have dependent data $\boldsymbol{y} = (y_1, \dots, y_n)$ generated from some unknown distribution $F = F_{\boldsymbol{Y}} = F_{(Y_1,\dots,Y_n)}$.

**Goal:**

**Challenge:**

We will consider 2 approaches

## 3.1 Model-based approach

**Example 3.1** Suppose we observe a time series $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ which we assume is generated by an AR(1) process, i.e.,

If we assume an AR(1) model for the data, we can consider a method similar to bootstrapping residuals for linear regression.

**Model-based** – the performance of this approach depends on the model being appropriate for the data.

## 3.2 Nonparametric approach

To deal with dependence in the data, we will employ a nonparametric *block* bootstrap.

**Idea:**

### 3.2.1 Nonoverlapping Blocks (NBB)

Consider splitting $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ in $b$ consecutive blocks of length $\ell$.

We can then rewrite the data as $\boldsymbol{Y} = (\boldsymbol{B}_1, \ldots, \boldsymbol{B}_b)$ with $\boldsymbol{B}_k = (Y_{(k-1)\ell+1}, \ldots, Y_{k\ell})$, $k = 1, \ldots, b$.

Note, the order of data within the blocks must be maintained, but the order of the blocks that are resampled does not matter.

### 3.2.2 Moving Blocks (MBB)

Now consider splitting $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ into overlapping blocks of adjacent data points of length $\ell$.

We can then write the blocks as $\boldsymbol{B}_k = (Y_k, \ldots, Y_{k+\ell-1})$, $k = 1, \ldots, n - \ell + 1$.

### 3.2.3 Choosing Block Size

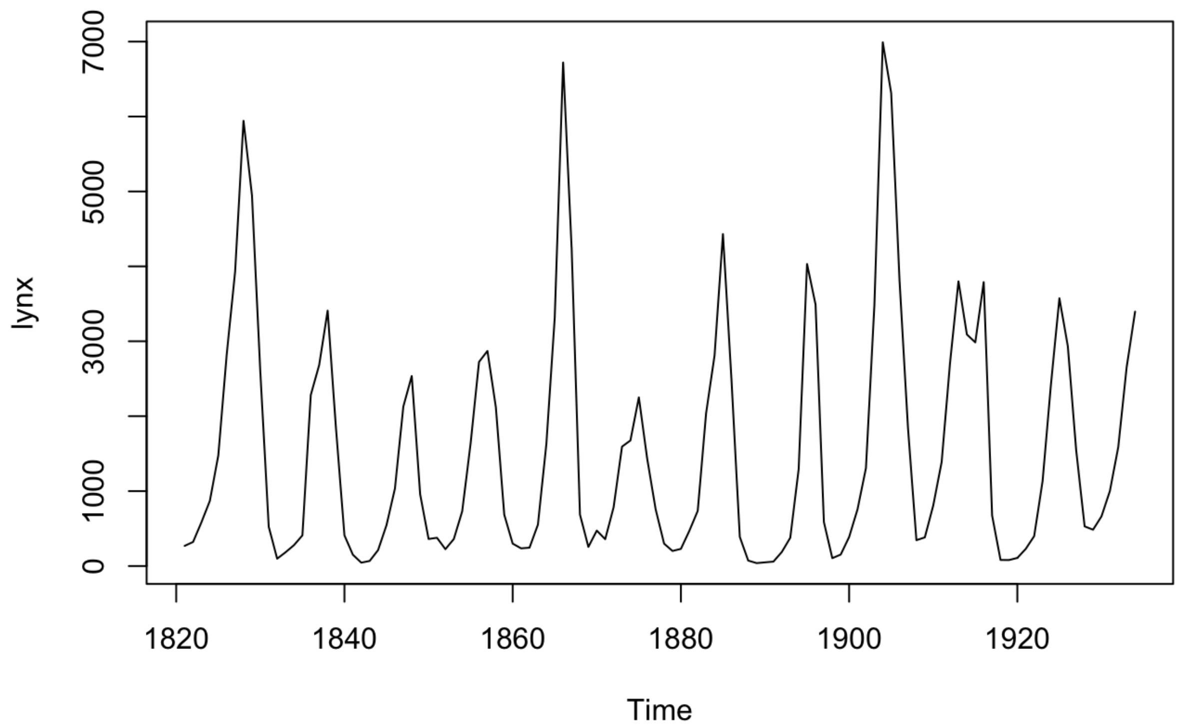
If the block length is too short,

If the block length is too long,

# Your Turn

We will look at the annual numbers of lynx trappings for 1821–1934 in Canada. Taken from Brockwell & Davis (1991).

```
data(lynx)
plot(lynx)
```




**Goal:** Estimate the sample distribution of the mean

```
theta_hat <- mean(lynx)
theta_hat
```

```
## [1] 1538.018
```

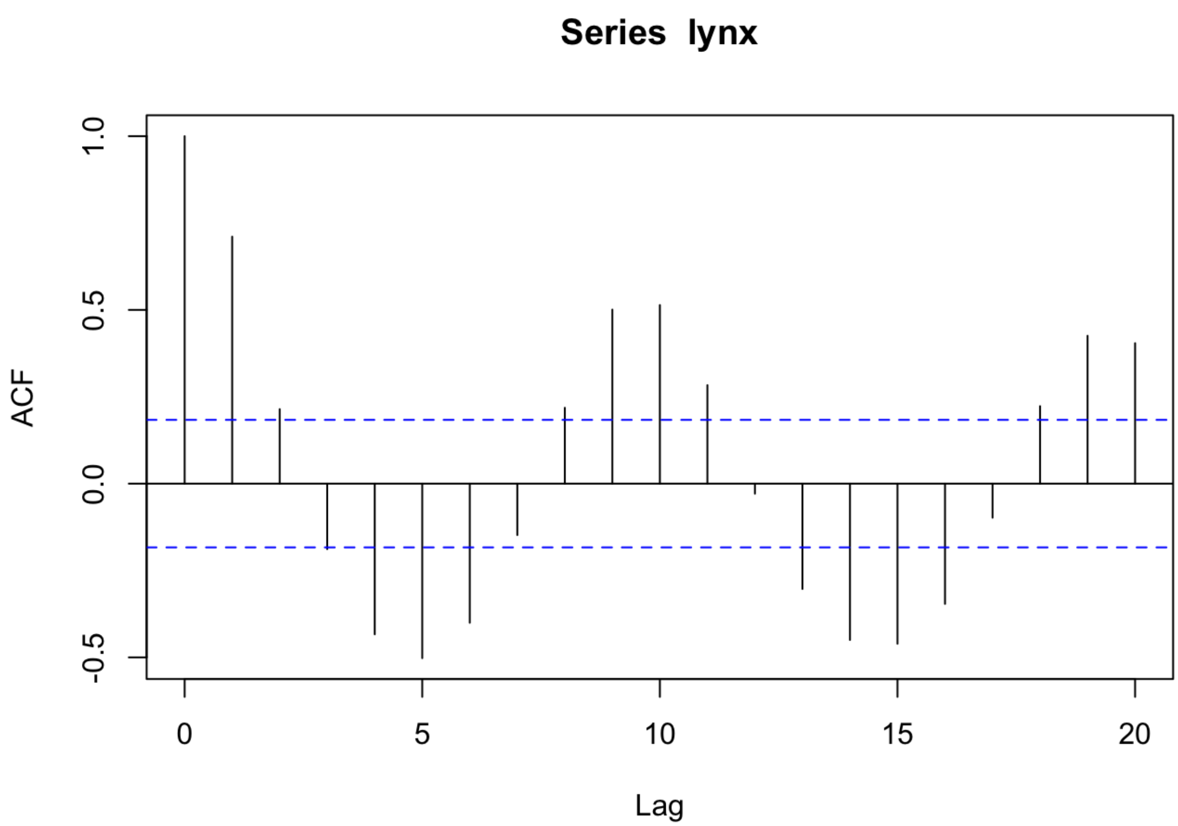
### 3.2.4 Independent Bootstrap

```r
library(simpleboot)
B <- 10000

## Your turn: perform the independent bootstap
## what is the bootstrap estimate se?
```

We must account for the dependence to obtain a correct estimate of the variance!

```r
acf(lynx)
```




The acf (autocorrelation) in the dominant terms is positive, so we are *underestimating* the standard error.

### 3.2.5 Non-overlapping Block Bootstrap

```
# function to create non-overlapping blocks
nb <- function(x, b) {
  n <- length(x)
  l <- n %/% b

  blocks <- matrix(NA, nrow = b, ncol = l)
  for(i in 1:b) {
    blocks[i, ] <- x[((i - 1)*l + 1):(i*l)]
  }
  blocks
}

# Your turn: perform the NBB with b = 10 and l = 11
theta_hat_star_nbb <- rep(NA, B)
nb_blocks <- nb(lynx, 10)
for(i in 1:B) {
  # sample blocks
  # get theta_hat^*
}

# Plot your results to inspect the distribution
# What is the estimated standard error of theta hat? The Bias?
```

### 3.2.6 Moving Block Bootstrap

```
# function to create overlapping blocks
mb <- function(x, l) {
  n <- length(x)
  blocks <- matrix(NA, nrow = n - l + 1, ncol = l)
  for(i in 1:(n - l + 1)) {
    blocks[i, ] <- x[i:(i + l - 1)]
  }
  blocks
}

# Your turn: perform the MBB with l = 11
mb_blocks <- mb(lynx, 11)
theta_hat_star_mbb <- rep(NA, B)
for(i in 1:B) {
  # sample blocks
  # get theta_hat^*
```

```
}

# Plot your results to inspect the distribution
# What is the estimated standard error of theta hat? The Bias?
```

### 3.2.7 Choosing the Block size

```
# Your turn: Perform the mbb for multiple block sizes l = 1:12
# Create a plot of the se vs the block size. What do you notice?
```

# 4 Summary

Bootstrap methods are simulation methods for frequentist inference.

Bootstrap methods are useful for

Bootstrap methods can fail when