**Analysing Time Series Data using R**

***Abstract:*** *In this project we are analysing the median heights of the Rio Negro river at Manaus which is a time series data with 1080 observations. For this analysis we are using the analytical methods such as empirical CDF, time series bootstrapping, Maximum Likelihood Estimator and Wald test.*

**Introduction:**

The dataset we are using is a monthly averages of the daily stages (heights) of the Rio Negro river at Manaus. Manaus is 18km upstream from the confluence of the Rio Negro with the Amazon. The height is observed using Manaus gauge which is tied in with an arbitrary benchmark of 100m set in the steps of the Municipal Prefecture. The data here cover 90 years from January 1903 until December 1992. The Manaus time series is of class "ts" and has 1080 observations of the river height.

First, we are observing the distribution of the observations using empirical CDF which shows that the observations are normally distributed. Then we are using the time series bootstrap function to find the median height. Here the challenge is the normal bootstrapping is for independent and identically distributed observations. But with time series data using this bootstrap method will destroy the time - dependence structure. Thus, here we are using block bootstrapping for time series data. There are different techniques like fixed block sampling, Stationary block sampling and model-based resampling. In our analysis we are using the technique of fixed block sampling where we are calculating the statistic for a block of fixed length B times. Then we are estimating the mean parameter of the distribution using Maximum Likelihood estimator method. With this estimated mean parameter, the distribution can give values that are near to actual observation with maximum likelihood. Finally, we are splitting the data into two – one set with first 45 years (1903 to 1948) observations and second set with next 45 years (1949 to 1992) observations. We are using Wald test to test if median of both the samples are same. From this hypothesis testing we got, at 0.05 level of significance the median of both the samples are same.

**Project Description:**

**True Distribution**

The true distribution of the variable under analysis was found to be normal.

Fig 1 represents the true distribution of the height variable.

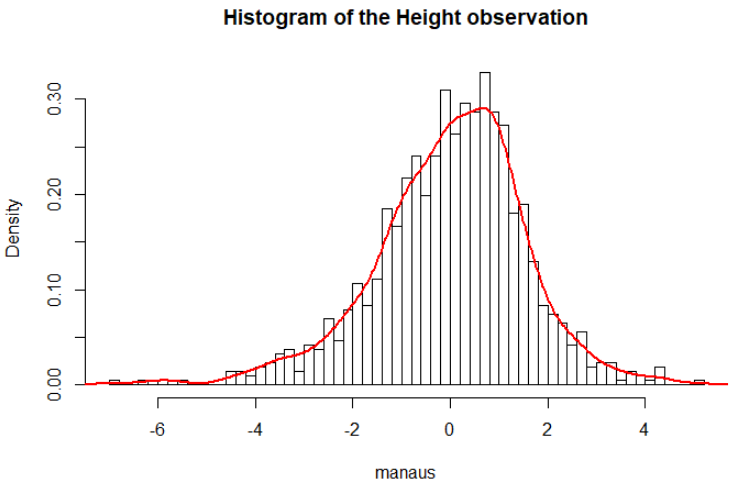


Fig 1. True distribution of height variable

**Empirical Cumulative Distribution Function (ECDF)**

The empirical cumulative distribution function is an estimator(non-parametric) of the true CDF of the random variable. The random variable used is the index of average height of Rio Negro river at Manaus (compared to the reference level).

***Calculating ECDF for the average height of Rio Negro river at Manaus***

The ECDF of the sample data was obtained using the ecdf() function in R.

ECDF function =

ECDF Plot

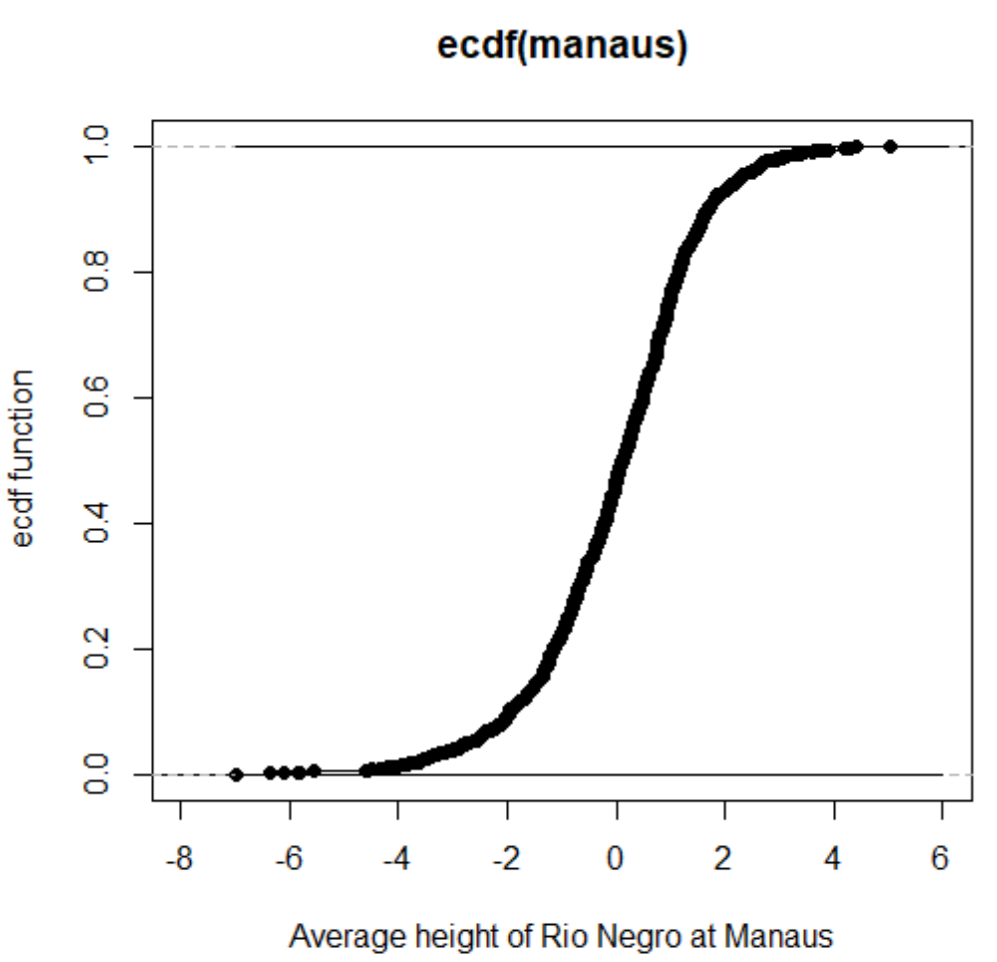


Fig 2 ECDF of the ‘average height of Rio Negro at Manaus’

**Block Bootstrapping for Time Series Data**

Block bootstrapping is a bootstrapping method used for time series data (data correlated with time). This method of bootstrapping resamples blocks of data (and not individual values).

Finding the bootstrap standard errors and confidence interval in R

Block bootstrapping method was used to find the standard error of the median height index `of Rio Negro river. The value of this standard error was then used to calculate the normal, pivotal and quantile confidence intervals for the median height index.

The standard error of the median height index of the river from bootstrapping: **0.1136248**

At 95% confidence, the normal confidence interval for median height index of Rio Negro river at Manaus was found to lie within -**0.1090345** and **0.3363745.**

At 95% confidence, the pivotal confidence interval for median height index of Rio Negro river at Manaus was found to lie within -**0.1017267** and **0.** **3136839.**

At 95% confidence, the quantile confidence interval for median height index of Rio Negro river at Manaus was found to lie within **-0.08634388** and` **0.32906675.**

**Maximum Likelihood Estimator**

Maximum likelihood function estimates the value of the parameter such that the observed data is most probable under the assumed statistical model. The MLE of the mean (of average height index of the River) was found out to be **1.018519e-06.**

The asymptotic distribution of the data was found to be a normal distribution.

**Hypothesis test for the difference in medians**

The objective of the test was to see if the median height index during the years 1903- 1947 is significantly different from the median height index during the years 1948-1992. A Wald test was used for this purpose.

***Hypothesis Formulation***

Let X1 be the median height index of the river during the years 1903- 1947 and X2 be the median height index of the river during the years 1948-1992

**Null Hypothesis (Ho):** X1 = X2

**Alternate Hypothesis (HA):** X1 ≠ X2

***Obtaining the test statistic***

Z test statistic = **0.1444021**

**p- value** corresponding to the above test statistic is **0.5574085**

At 95% confidence, α = 0.05

Since **p-value > α /2, we do not reject the null hypothesis**

***Hypothesis test conclusion***

At 95% confidence, we do not have enough evidence to reject the null hypothesis.

**Conclusion:**

In this project, we analysed the distribution of height over the years using the actual observations using the method of empirical CDF. Without losing the time – dependence structure of the dataset we calculated median height using block bootstrapping. We also calculated the normal confidence interval, pivotal confidence interval and quantile confidence interval for the median height. We calculated the maximum likelihood estimate of mean parameter. Finally, we split the observation into two and using Wald test we concluded the median of two samples are same at 0.05 level of significance.

**Bibliography:**

https://cran.r-project.org/web/packages/boot/boot.pdf

https://sjfox.github.io/post/2015-07-23-r-rmarkdown/

Davison, A.C. and Hinkley, D.V. (1997) Bootstrap Methods and Their Application. Cambridge University Press.

Sternberg, H. O'R. (1987) Aggravation of floods in the Amazon river as a consequence of deforestation? Geografiska Annaler, 69A, 201-219.

Sternberg, H. O'R. (1995) Waters and wetlands of Brazilian Amazonia: An uncertain future. In The Fragile Tropics of Latin America: Sustainable Management of Changing Environments, Nishizawa, T. and Uitto, J.I. (editors), United Nations University Press, 113-179.

**Appendix:**

*R code*

install.packages("boot")

library(boot)

data("manaus")

?manaus

hist(as.numeric(manaus))

#Empirical CDF

manaus.emp <- ecdf(manaus)

n<- length(manaus)

min(manaus)

max(manaus)

grid=seq(-7,6,length.out=1000)

manaus.lower <- (manaus.emp(grid) - sqrt((1/2\*n)\*log(2/.05)))

manaus.lower<-pmax(manaus.lower, 0)

manaus.upper <- (manaus.emp(grid) + sqrt((1/2\*n)\*log(2/.05)))

manaus.upper<-pmin(manaus.upper, 1)

plot(ecdf(manaus),xlab='X',ylab='ecdf function')

lines(grid, manaus.lower)

lines(grid, manaus.upper)

#Bootstrap standard errors and confidence intervals

num\_resamples<- 1000

block\_length<- round(length(manaus)^(1/3))

manaus.tsboot<- tsboot(manaus, median, R = num\_resamples, l = block\_length, sim = "fixed")

manaus.median<-median(as.numeric(manaus))

#Normal confidence interval

manaus.tsboot.se<-sd(manaus.tsboot$t)

CI.normal<-c(manaus.median-1.96\*manaus.tsboot.se, manaus.median+1.96\*manaus.tsboot.se)

#Pivotal confidence interval

CI.pivotal<-2\*manaus.median-quantile(manaus.tsboot$t,probs = c(0.975, 0.025))

#Quantile confidence inteval

CI.quantile<-quantile(manaus.tsboot$t,probs = c(0.025, 0.975) )

​

#MLE

mu<- mean(manaus)

# Wald test for Hypothesis Testing  
## First 45 years data  
manaus\_num <- as.data.frame(as.numeric(manaus))  
manaus\_first <-manaus\_num[1:540,1]  
manaus\_first\_median <- median(manaus\_first)  
manaus\_first\_variance <- sd(manaus\_first)\*sd(manaus\_first)

##Last 45 years data  
manaus\_second <-manaus\_num[541:1080,1]  
manaus\_second\_median <- median(manaus\_second)  
manaus\_second\_variance <- sd(manaus\_second)\*sd(manaus\_second)

z\_statistic <- (manaus\_first\_variance - manaus\_second\_variance)/sqrt(manaus\_first\_variance+manaus\_second\_variance)

pnorm(z\_statistic)