** BE Statistical Hydrology **

FREQUENCY ANALYSIS WITH R

(Worked on February Flows)

GROUP MEMBERS

Hammad Khalid

Umar Alfa

Zohaib Saleem

Md Golam Sarwar

M1-HCE
2022-2023

Table of Contents

1. PART I – DESCRIPTIVE STATISTICS AND DISTRIBUTION FITTING 5

Objectives of Part I 5

Description of the Loire station at Blois 5

Hydrological regime 6

1.1. ANALYSIS ON ANNUAL STREAM FLOW 7

Stream flow data structure: 7

Descriptive statistics 9

Empirical distribution 12

Distribution fitting on annual stream flow 16

1.2. ANALYSIS ON MONTHLY STREAM FLOW 23

2. PART II – STATISTICAL TESTS 30

Objectives of Part II 30

Short reminder on statistical tests 30

2.1. CONFORMITY TEST ON ANNUAL DATA 32

2.2. HOMOGENEITY TESTS ON ANNUAL DATA 33

2.3. GOODNESS-OF-FIT TESTS ON MONTHLY DATA 36

2.4. WORK ON OTHER MONTHLY DATA AND WRITE A REPORT 40

References 51

TABLE OF FIGURES

FIGURE 1 : LOCATION MAP OF THE LOIRE AT BLOIS (CHARLES LEMARCHAND, 2014) 6

FIGURE 2: ANNUAL AVERAGE MONTHLY RAINFALL REGIME HISTOGRAM (HYDROPORTAIL, 2015) 7

FIGURE 3 : ANNUAL STREAM FLOW SCATTERPLOT 9

FIGURE 4 : OCTOBER STREAM FLOWS 10

FIGURE 5 AVERAGE ANNUAL OCTOBER STREAM FLOW AND 1:1 LINE 11

FIGURE 6 : BOX PLOT FOR ANNUAL DATA 12

FIGURE 7 : EMPIRICAL DENSITY AND CUMULATIVE DENSITY 15

FIGURE 8 : THEORETICAL DISTRIBUTIONS NORMAL, GAMMA AND LOG NORMAL 17

FIGURE 9: CDF FOR THEORETICAL DISTRIBUTIONS 17

FIGURE 10 : Q-Q PLOT FOR THEORETICAL DISTRIBUTIONS 18

FIGURE 11 : P-P PLOT FOR THEORETICAL DISTRIBUTIONS 18

FIGURE 12 : FEBRUARY STREAM FLOW VS YEARS 23

FIGURE 13 : EMPERICAL DISTRIBUTION AND CDF FOR FEBRUARY FLOWS 25

FIGURE 14 : THEORETICAL DISTRIBUTIONS FOR FEBRUARY FLOWS 26

FIGURE 15: EMPIRICAL AND THEORETICAL CDFS FOR FEBRUARY FLOWS 26

FIGURE 16 : Q-Q PLOT FOR THEORETICAL DISTRIBUTION OF FEBRUARY FLOWS. 27

FIGURE 17 : P-P PLOT FOR THEORETICAL DISTRIBUTION OF FEBRUARY FLOWS 27

Work Organization:
Before starting, register yourself in a working group (work by groups of 3) within the Engineering Hydrology working space in Chamilo.

Data
Mean monthly and mean annual discharge [m^3/s] observed on the Loire River at Blois from 1863 to 2019 – cf. Loire_Blois_R.csv (the data separator is ";"). Download the file Loire_Blois_R.csvfrom Chamilo into a chosen working directory.

Report instructions
A comprehensive report on the full Statistical Hydrology BE (Part I and Part II) will be due by November 4th. It should be dropped on Chamilo in Engineering Hydrology > Travaux > BE Statistical Hydrology. Your group number should appear in the file name. It will be graded.

The sections and subsections below provide the structure for your report. Once the analysis in R is ready, knit the document to Word and finalize its formatting. A comprehensive report should include:

- a table of content
- numbered figures and tables (if applicable)
- a table of figures
- a table of tables
- references to figures and tables within the text
- a list of the sources and references you used in your report (e.g. figures you downloaded, values/information you extracted from webpages/books/articles) – these references should also appear in the text, next to the information taken from the reference.

Reminder on how to include R code in your report
In your reports, all commands run in R should appear (echo=TRUE) as well as results (eval=TRUE), unless you are (a) loading or installing a package, or (b) setting the path to a directory. All the text that does not correspond to R commands should be added outside chunks.

Setup of the R environment

To set up your R environment for both parts of the BE, you will need, in R, to:

1. Install the 'fitdistrplus' library if it is not already installed
2. Load the 'fitdistrplus' library library('fitdistrplus')
3. Set the working directory to the location of your Loire_Blois_R.csv file setwd
4. Check the working directory getwd.

Setup of the RMarkdown package

To set up RMarkdown:

1. Install the 'stringi' package if it is not already installed. You can check in the package list on the righ panel of Rstudio. If a pop-up window opens, choose not to install from source (click No).
2. Install the 'rmarkdown' package if it is not already installed.
3. Load the 'rmarkdown' library
4. Open the file BE1.Rmd in Rstudio
5. Test the compilation by knitting to HTML
6. Make sure that the line 'pdf_document: default' does not appear in your header (this will compile the document to pdf by default, which will fail on school computers)
7. Pick the RMarkdown display in the top left corner of the text editor (Source/Visual)

---

1. PART I – DESCRIPTIVE STATISTICS AND DISTRIBUTION FITTING

Objectives of Part I

- To become familiar with descriptive statistics and plotting with R/RStudio
- To become familiar with distribution fitting with R/RStudio
- To become familiar with Q-Q and P-P plots

Description of the Loire station at Blois

Add a brief description of the Loire station at Blois including at least:

- who operates it
  The administrative responsibilities are fulfilled by **DREAL.**
- its area (km^2)
  Watershed area of the basin is **38,320** km^2.

- a map of its location

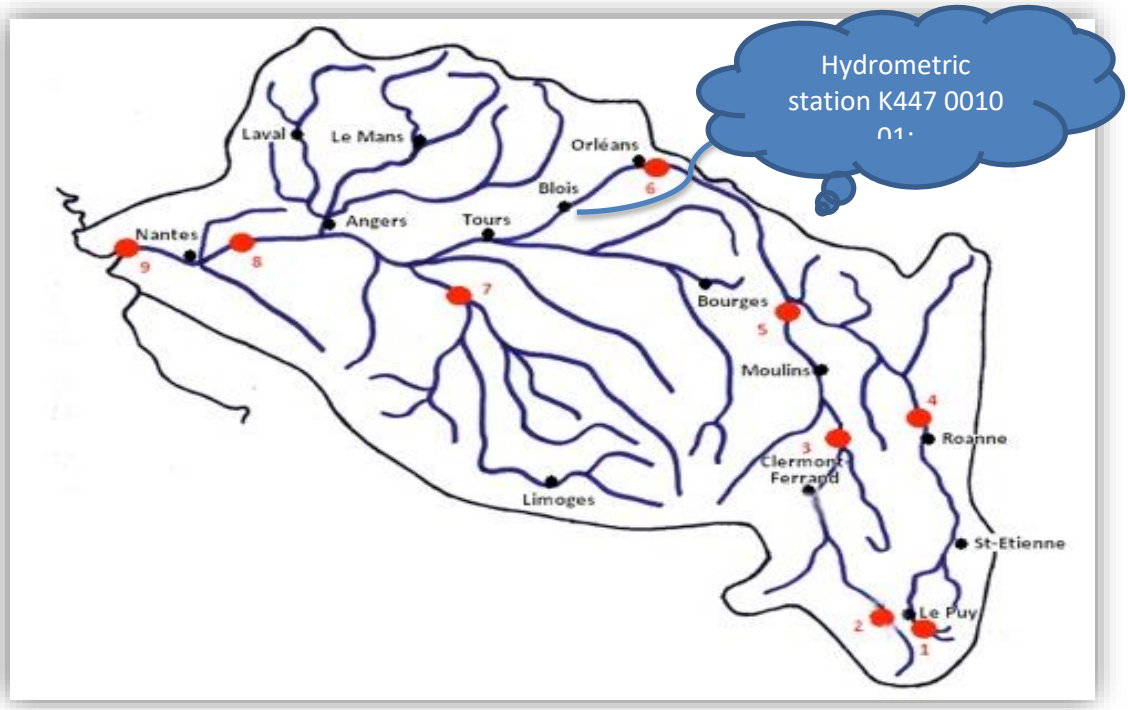


Figure 1 : Location Map of the Loire at Blois (Charles Lemarchand, 2014)

- its mean annual discharge (m^3/s) and annual runoff (mm)
  The mean annual discharge is **355 m3/sec** and annual runoff is **292 mm.** (HYDROPORTAIL, 2015)

- its mean annual precipitation (mm)
  The mean average precipitation is approximately **530mm** (calculated over a period of 20 years Data) (Tutiempo Network)

Hydrological regime

Add information about the hydrological regime at the station including at least:

- the lowest monthly discharge (value, time of year)
  As per **Figure 2**, the lowest monthly discharge is in **August** of **117m^3/sec**

- the highest monthly discharge (value, time of year)
  As per **Figure 2**, the lowest monthly discharge is in **February** of **583m^3/sec**

- the timing of high and low waters and your conclusions on the type of flow regime
  From **Figure 2** it could be observed that High waters are observed in winter, and low waters in summer, therefore the Loire river corresponds to a flow regime of a ***Perennial river system.***

- A figure of the annual flow regime.

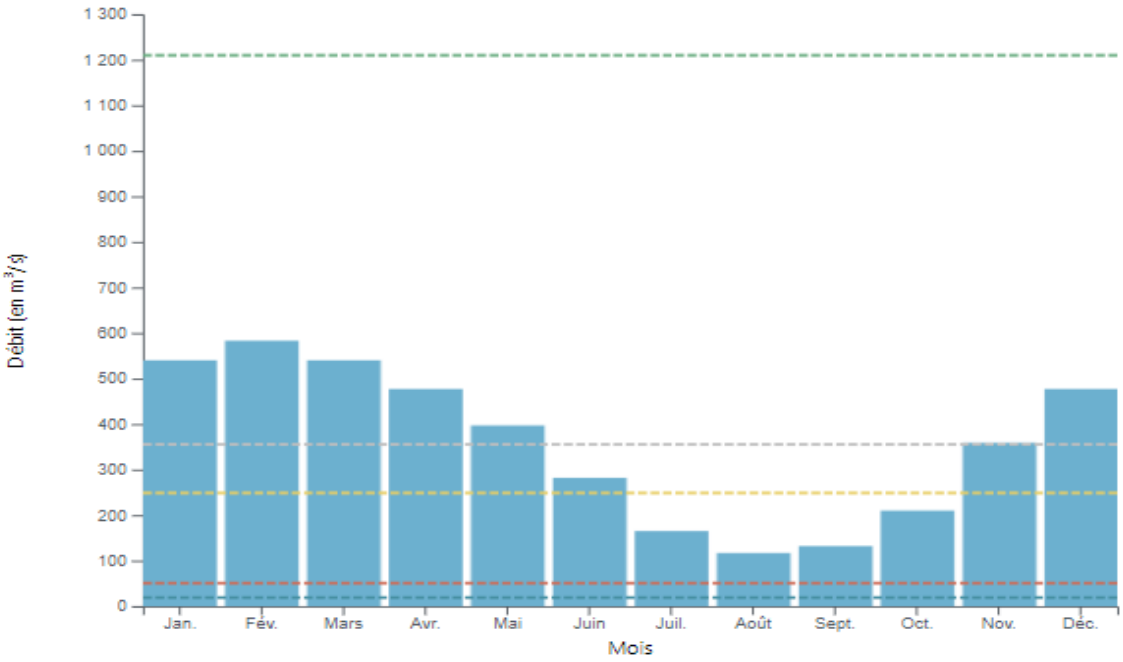


Figure 2: Annual Average Monthly Rainfall Regime Histogram (HYDROPORTAIL, 2015)

1.1. ANALYSIS ON ANNUAL STREAM FLOW

Stream flow data structure:

- Read the Loire_Blois_R.csvfile (using the function read.csv) and store the data set into a variable called Data. Choose the sepand headerarguments carefully. You can find more information on the read.csvfunction by typing ? read.csv.

```
data <- read.csv("Loire_Blois_R.csv", header = TRUE, sep = ";" )
```

- Display the internal structure of the data you just loaded using the function str

```
str(data)

## 'data frame':                154 obs. of          14 variables:
##     $ Year       : int       1863 1864 1865 1866 1867 1868 1869 1870 1871
1872 ...
##     $ Jan        :   num     821   308   557   335   961   ...
##     $ Feb        :   num     361   303   850   600   867   ...
##     $ Mar        :   num     262   592   678   863   899   ...
##     $ Apr        :   num     398   293   525   876   704   ...
##     $ May        :   num     217   135   199   317   438   ...
##     $ Jun        :   num     132   168   102   342   181   ...
```

```
##  $ Jul      :  num  115.5 97.9 80.1 192.3 199 ...
##  $ Aug      :  num  56.2 62.6 78.7 155.7 166.8 ...
##  $ Sep      :  num  150.9 69.4 55.1 909.8 93.7 ...
##  $ Oct      :  num  543.4 135.7 64.9 598.1 371.4 ...
##  $ Nov      :  num  417 575 268 335 262 ...
##  $ Dec      :  num  328 480 446 760 314 ...
##  $ Annual:     num  317 268 322 522 453 ...
```

Write a paragraph to answer the following questions:

- How many variables are there in the Data data.frame?
  There are **14 variables** in the data frame.
- What is the name of each variable?

```
variable.names(data)
```

```
##  [1] "Year"   "Jan"    "Feb"    "Mar"    "Apr"    "May"    "Jun"
"Jul"
##  [9] "Aug"    "Sep"    "Oct"    "Nov"    "Dec"    "Annual"
```

- Which is the type of each variable?

```
lapply(data,class)
```

```
## $Year
## [1] "integer"
##
## $Jan
## [1] "numeric"
##
## $Feb
## [1] "numeric"
##
## $Mar
## [1] "numeric"
##
## $Apr
## [1] "numeric"
##
## $May
## [1] "numeric"
##
## $Jun
## [1] "numeric"
##
## $Jul
## [1] "numeric"
##
## $Aug
```

```
##
## $Sep
## [1] "numeric"
##
## $Oct
## [1] "numeric"
##
## $Nov
## [1] "numeric"
##
## $Dec
## [1] "numeric"
##
## $Annual
```

- How many observations are there of each variable?
  There are **154** observations in the each variable.

Descriptive statistics

- Fill in the command below to display the annual stream flow as a function of the year. Add labels to your axes, indicating units when necessary. Set the y-axis limits between 0 and 1224 m^3/s.

```
library(viridis)
plot(data$Year, data$Annual, xlab = 'Year', ylab = 'Annual Flows  (m3/sec)', main = "Annual Stream Flow", col.lab = "Black", pch = 19,  col = viridis(n=256, option = "H"), bg = "Yellow", font.lab = 2, bty= "n", cex.axis = 0.9)
```

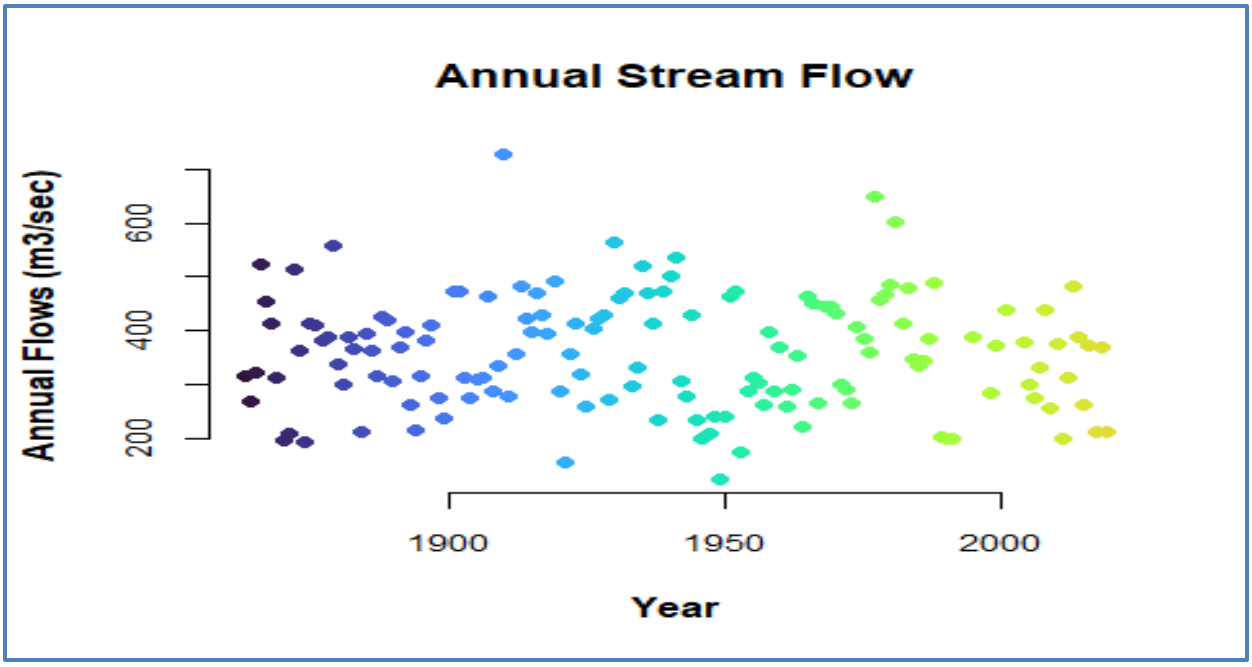


Figure 3 : Annual Stream Flow Scatterplot

- Using the function plot, display the October stream flow as a function of the year. Add labels to your axes and set the y-axis limits between 0 and 1224 m^3/s.

```
plot(data$Year, data$Oct, xlab = 'Year', ylab = 'October Flows (m3/sec)', main = "October Stream Flows", col.lab = "Black", pch = 18, cex= 1,col = viridis(n=256, option = "D"), bg = "Yellow", font.lab = 2, bty= "n", cex.axis = 0.9, ylim = c(0,1224), xlim = c(1863,2019))
```

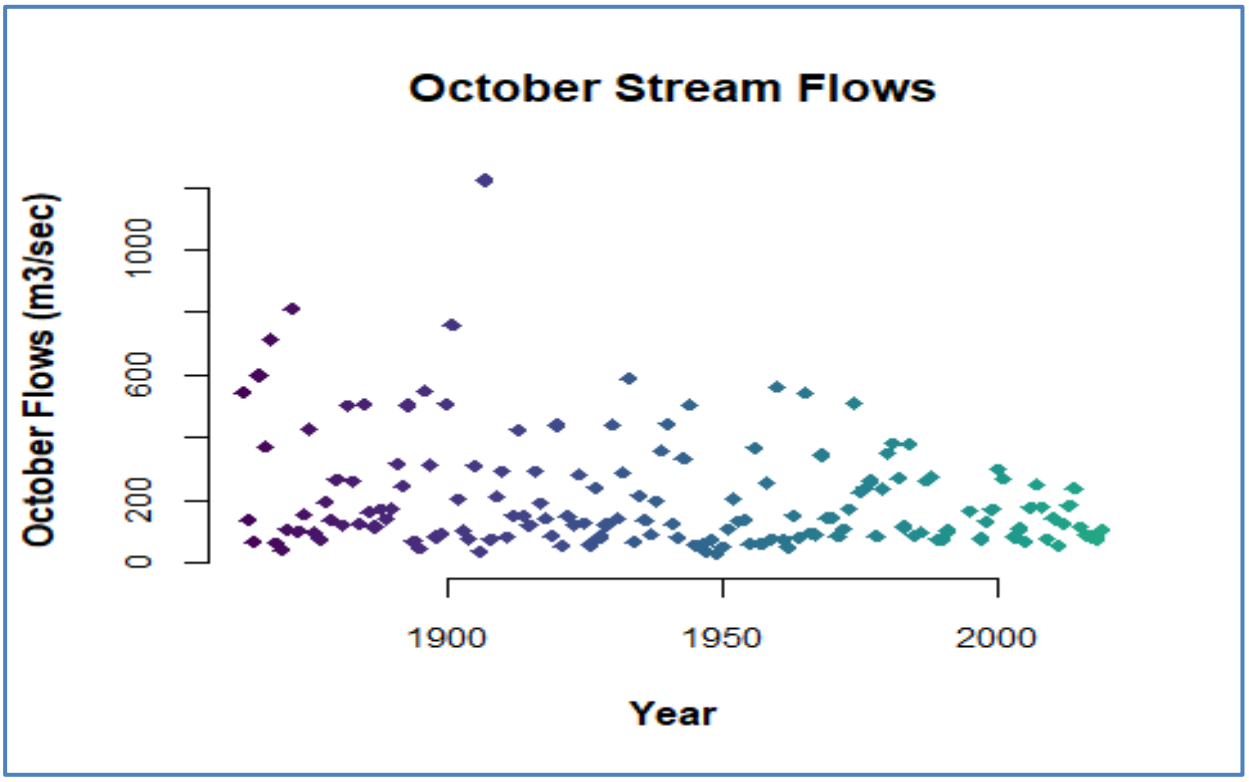


Figure 4 : October Stream Flows

- Display the annual stream flow against the October stream flow. Add labels to your axes. Add a 1:1 line.

```
plot(data$Annual, data$Oct, xlab = 'Annual Flow', ylab = 'October Flow', main = "Average Annual October Stream Flow", col.lab = "Black", pch = 19, col = viridis(n=256, option = "D"), bg = "Yellow", font.lab
= 2, bty= "n", cex.axis = 0.9, ylim = c(0,1224)) abline(a = 0, b = 1, col =
"Purple", lwd = 2.5)
```

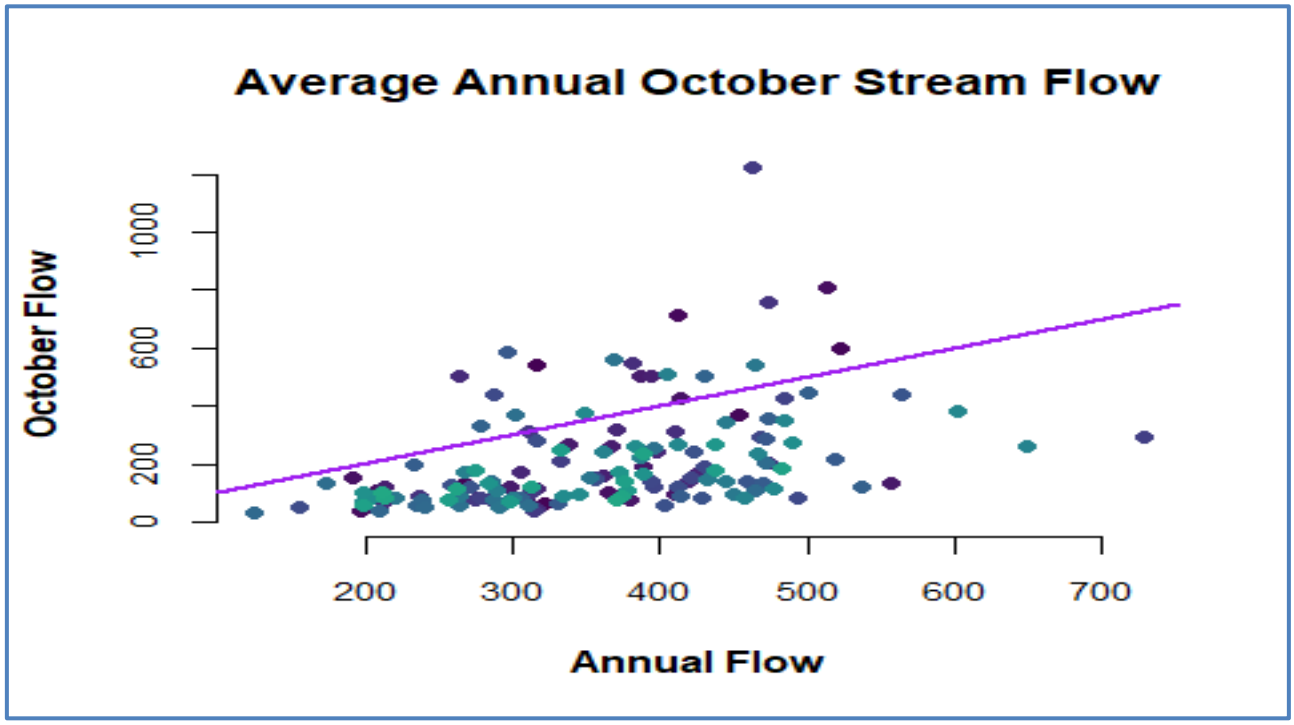


Figure 5 Average Annual October Stream Flow and 1:1 Line

- Calculate the main statistical descriptors of the annual stream flow using the function summary.

```
summary(data$Annual)
##      Min. 1st Qu.      Median      Mean 3rd Qu.      Max.      NA's
##     124.2     282.8     362.8     359.0     430.4     727.9         6
```

- Calculate the standard deviation in annual stream flow. Use ?sdto see how to handle missing values.

```
sd(data$Annual, na.rm = T)

## [1] 104.4074
```

Write a paragraph to summarize your results and answer the following questions:

- How many missing values (NA) are there?

  There are **6** missing values in the data.

Write a paragraph to summarize the obtained results.

The **Figure 3 : Annual Stream Flow Scatterplot** shows year wise annual flows, from the figure we see that there are two values above the **600 m3/sec** flows in the period of 156 years most of the flow values are concentrated between **200m3/sec** to **550 m3/sec.** From **Figure 4 : October Stream Flows**, it can be observed that in the late 19th Century month of October observed high flows – above **400m3/sec** as compared to recent years (late 20th century) where majority of flow values are concentrated below the **400m3/sec** mark. The statistical description of the data set shows that the data has skewness and will not perfectly follow the normal law and it could be deduced from the 1st quartiles that 25 % of the data set values are below 282 m3/sec and 75 % of the values are above it similarly from the 3rd quartile we can say that 75 % of the values are below 430 m3/sec and only 25 % are above it which gives better indication about the normal flow values. The percentage difference between the Standard deviation **(104.41 m3/sec)** and the mean **(359 m3/sec)** is calculated to be **70 %** which shows that majority of data values are closer to the mean values.

Empirical distribution

- Display a box plot of the annual stream flow. Display the statistics displayed in the box plot. See boxplotand boxplot.stats.

```
boxplot(data$Annual, col = "green", border = par("fg"))
```

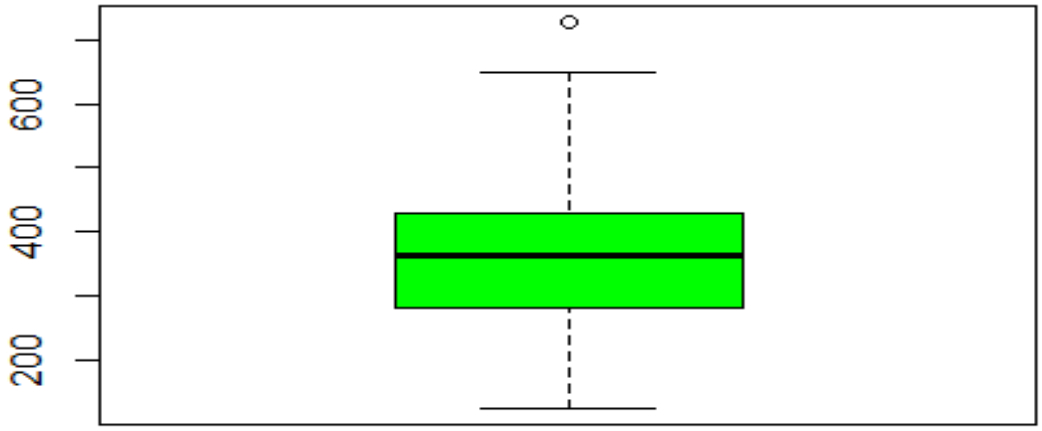


Figure 6 : Box Plot for Annual Data

```
boxplot.stats(data$Annual)

## $stats
## [1] 124.180 281.445 362.820 430.370 649.580
##
## $n
## [1] 148
##
## $conf
## [1] 343.4783 382.1617
##
## $out
## [1] 727.92
```

- Identify the indices of annual stream flow values that are missing. To that effect, you will look into the functions is.naand which.

```
is.na(data$Annual) # is.na function looks for na values and appears  true for that position.

##       [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  FALSE FALSE
##      [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  FALSE FALSE
##      [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  FALSE FALSE
##      [37] FALSE          TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  FALSE FALSE
##      [49] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  FALSE FALSE
##      [61] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  FALSE FALSE
##      [73] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  FALSE FALSE
##      [85] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  FALSE FALSE
##      [97] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  FALSE FALSE
## [109] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  FALSE FALSE
## [121] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE                                  TRUE
FALSE      TRUE
## [133] FALSE FALSE              TRUE FALSE       TRUE     TRUE FALSE FALSE FALSE FALSE  FALSE
FALSE
## [145] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

which(is.na(data$Annual)) # this gives the location of missing values

## [1]         38 130 132 135 137 138

!is.na(data$Annual) #! reverses the values from true to false: - so  this gives the location which are
not NA

##       [1]      TRUE     TRUE     TRUE     TRUE     TRUE     TRUE     TRUE     TRUE     TRUE     TRUE
TRUE     TRUE
##      [13]      TRUE     TRUE     TRUE     TRUE     TRUE     TRUE     TRUE     TRUE     TRUE     TRUE
TRUE     TRUE
##      [25]      TRUE     TRUE     TRUE     TRUE     TRUE     TRUE     TRUE     TRUE     TRUE     TRUE
TRUE     TRUE
##      [37]      TRUE    FALSE     TRUE     TRUE     TRUE     TRUE     TRUE     TRUE     TRUE     TRUE
TRUE     TRUE
##      [49]      TRUE     TRUE     TRUE     TRUE     TRUE     TRUE     TRUE     TRUE     TRUE     TRUE
TRUE     TRUE
```

```
##  [61]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
TRUE  TRUE
##  [73]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
TRUE  TRUE
##  [85]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
TRUE  TRUE
##  [97]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
TRUE  TRUE
## [109]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
TRUE  TRUE
## [121]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE
TRUE FALSE
## [133]  TRUE  TRUE FALSE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE
TRUE  TRUE
## [145]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

- Create a new vector annual_Q_without_NAcontaining only the annual stream flow values that are not missing.

```
annual_Q_without_NA <-data$Annual[which(!is.na(data$Annual))]
```

- Plot the empirical histogram, teh empirical density and the empirical cumulative distribution using the function plotdistof the fitdistrpluspackage. Look into the options histoand demp.

```
plotdist(annual_Q_without_NA, histo = TRUE, demp = TRUE, col =  viridis(n=154, option = "E"), pch =
20, type = "b" , lty = "solid",  font.lab = 2, bty= "n", cex.axis = 0.9,)
```

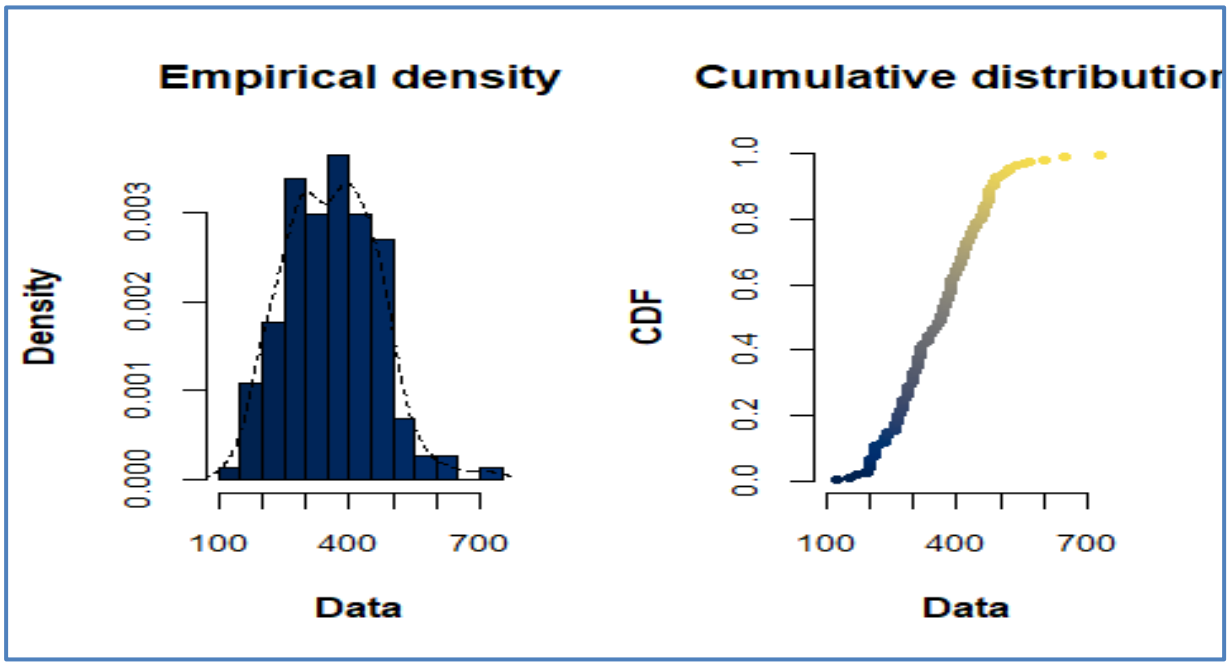


Figure 7 : Empirical Density and Cumulative Density

Write a paragraph to summarize your results and answer the following questions:

- How are the lower and the inner fence of the box plot calculated? Look at the help by typing? box plot.

The two fences are calculated as follows

Upper fence = Q3 + (1.5 * IQR)

**Lower fence = Q1 – (1.5 *

IQR) where,**

Q3 = Upper Quantile (75%)

Q1= Lower Quantile (25%)

- Comment the obtained results.

Box plot is a graphical representation of the data and box plot.stat function gives the $stats which define the statistical parameters of the box plot which are **124.180 281.445 362.820 430.370 649.580** (min, 1st quartile, median, 3rd quartile and Max respectively), **$n** represents the non-zero observations which were counted as **148**, **$conf** defines the two values from the notch (median) and $out defines the outlier value which is given as **727.92** These results are similar to the ones produced by the summary function in descriptive statistic sub-section under Analysis on Annual Stream Flow. However, from the box plot we cannot deduce the pattern of distribution, for which histogram was plotted and density plot was superimposed on it which demonstrates that the histogram follows a nearly normal law pattern, but to confirm this hypothesis we will fit different distributions on the data.

Distribution fitting on annual stream flow

- Use the function fitdist of the fitdistrplus package to fit successively a normal, log normal and gamma distribution to the annual streamflow (from which missing values were removed). You will use the moment matching estimation method for the fitting.

```
norm_fit <-          fitdist(annual_Q_without_NA, distr = "norm", method =  "mme")

gamma_fit <- fitdist(annual_Q_without_NA, distr = "gamma", method =  c("mme"))

lognormal <- fitdist(annual_Q_without_NA, distr = "lnorm", method =  c("mme"))
```

- Create a list containing all three fits.

```
all_fits <- list(norm_fit, gamma_fit, lognormal)
```

- Plot and compare the fitted density functions using denscomp and the list you created. Add a legend to identify each fit.

```
denscomp(all_fits, datacol = viridis(7, option = "H"), fitlwd = 2)
```

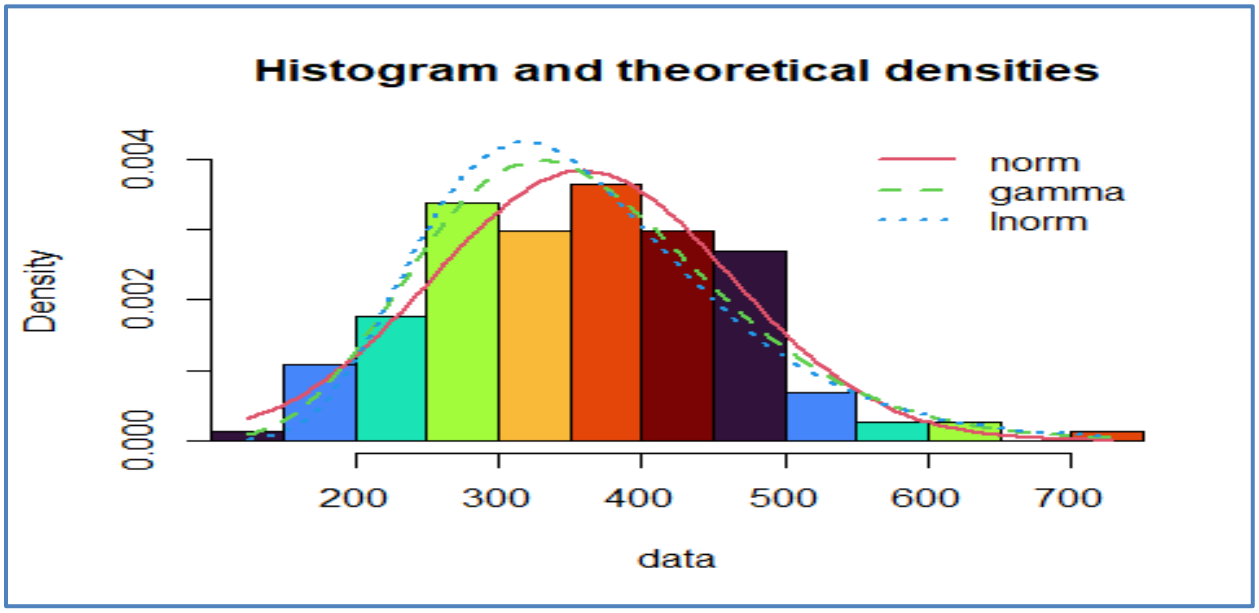


Figure 8 : Theoretical Distributions Normal, Gamma and Log Normal

- Plot and compare the fitted distributions using cdfcompand the list you created. Add a legend to identify each fit.

```
cdfcomp(all_fits, datapch = 20, datacol = viridis(10, option = "H"),  fitlwd = 1)
```

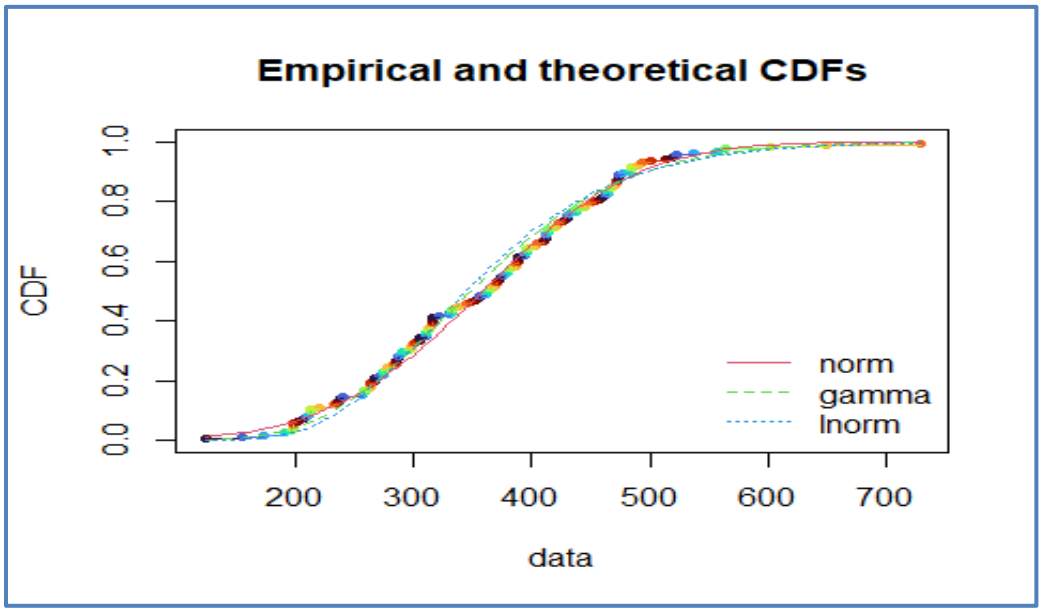


Figure 9: CDF For Theoretical distributions

- Create a quantile-quantile plot and a probability-probability plot using the list you created and the functions qqcompand ppcomp. Add legends to identify the fits.

```
qqcomp(all_fits, fitpch = 19, fitcol = viridis(5, option = "H"))
```

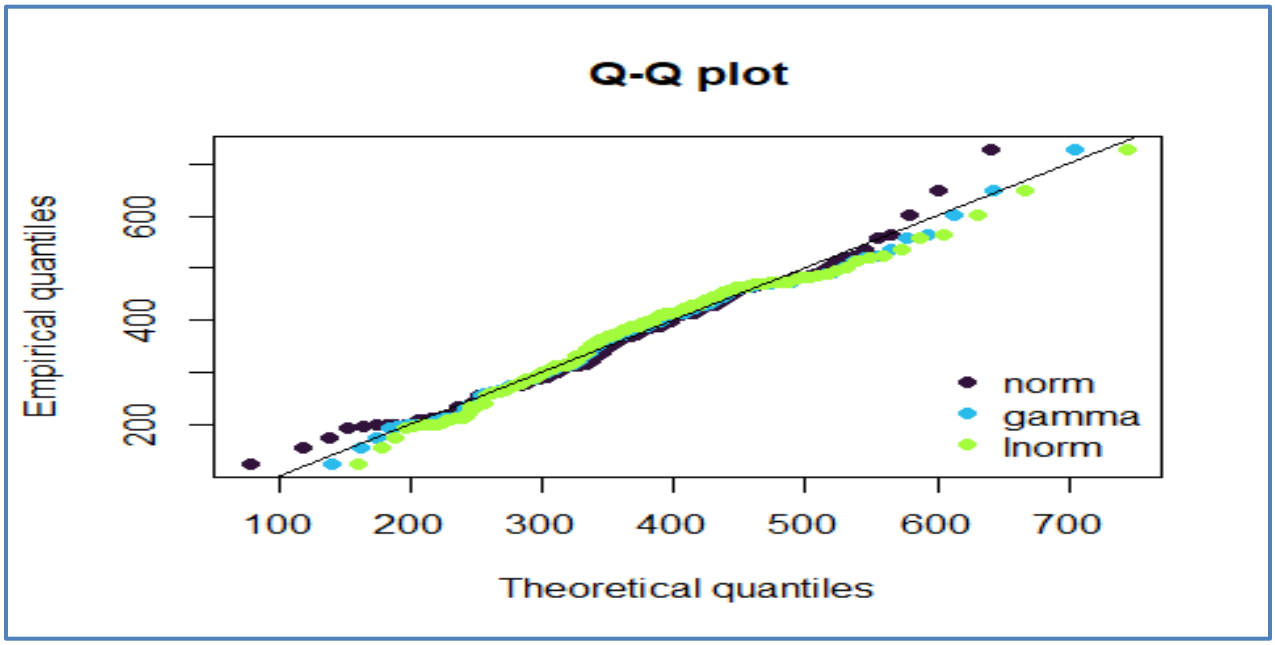


Figure 10 : Q-Q Plot for theoretical Distributions

```
ppcomp(all_fits, fitpch = 20, fitcol = viridis(3, option = "C"))
```

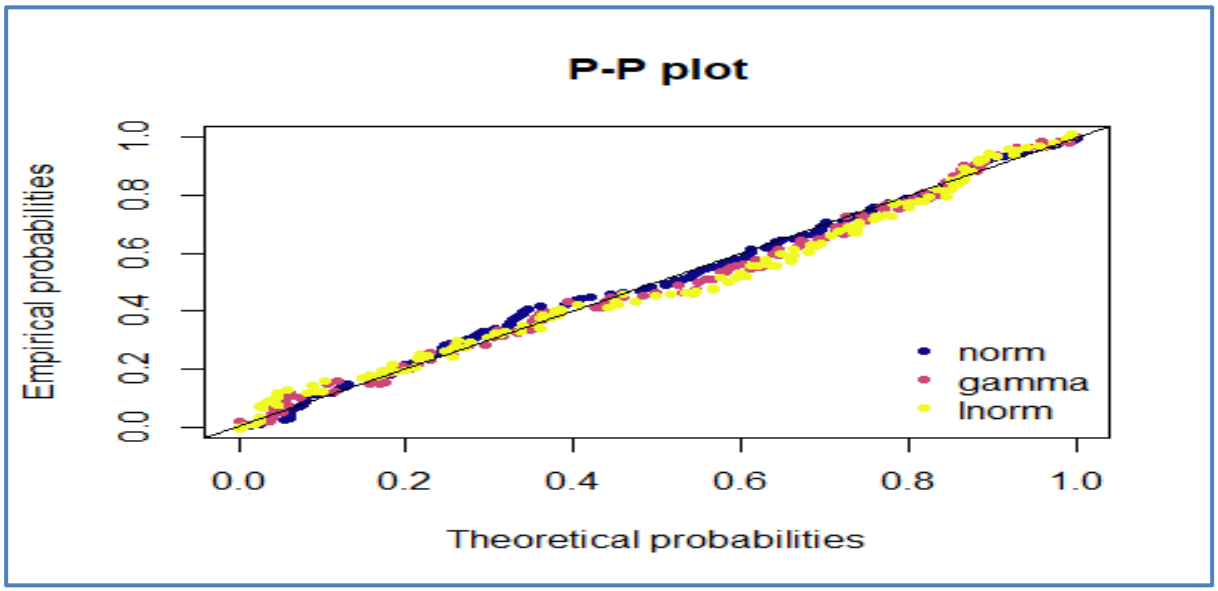


Figure 11 : P-P Plot for theoretical distributions

- Comment the obtained results by comparing the different fits based on the graphs you produced.

With the ***denscomp*** function, the density distribution was plotted for each fitted distribution, normal fitted distribution was more coherent at the peak than the other two fitted distribution, but then the **gamma** and **log normal** distribution were coherent with the data at the **tail**. Similarly the cumulative distribution frequency was plotted for empirical distribution (by default uses Hazen`s rule for the plotting positions) and for each fitted distribution and this plot further affirms the statement that gamma and log normal distribution is in more conjunction with the data at the tails. Q-Q plot compares the proximity of empirical quantiles (found from the data) with the theoretical quantiles, from the Q-Q plot we can deduce the quantiles of the data (empirical quantiles) are in more proximity with the **Gamma fitted distribution**. P-P plot is plotted to see the fit of data with the distribution or how closely the empirical probabilities align with the fitted model, but here in this case it seems that all three models (normal, gamma and log normal) predicts the same probabilities as predicted by empirical models.

- Get the distributions parameters of each fit obtained with the moment matching estimation method using summary

```
summary(norm_fit)
## Fitting of the distribution ' norm ' by matching moments
## Parameters :
##          estimate
## mean 358.9568
## sd       104.0541
## Loglikelihood:          -897.4497       AIC:    1798.899        BIC:    1804.894

summary(gamma_fit)

## Fitting of the distribution ' gamma ' by matching moments
## Parameters :
##                  estimate
## shape 11.90052559
## rate        0.03315308
## Loglikelihood:          -896.2208       AIC:    1796.442        BIC:    1802.436

summary(lognormal)

## Fitting of the distribution ' lnorm ' by matching moments
## Parameters :
##                  estimate
## meanlog 5.8428594
```

- The log likelihood, the Aikake Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are measures of goodness-of-fit. The higher the log likelihood, the better the fit. The lower the AIC and BIC, the better the fit. based on these criteria, which is the best fit for the Loire annual discharge at Blois?

  Based on the above summary, it is concluded that **gamma distribution** is the best fit for the Loire Annual Discharge as it has lower AIC and BIC (**1796.442** & **1802.436**) respectively, as compared to the measures of other distribution.

- Use the function fitdistto fit successively a normal, log normal and gamma distribution to the annual stream flow (from which missing values were removed), this time based on the maximum likelihood method.

```
norm_fit1 <- fitdist(annual_Q_without_NA, distr = "norm", method
= c("mle"))

gamma_fit1 <- fitdist(annual_Q_without_NA, distr = "gamma",  method = c("mle"))

lognormal1 <- fitdist(annual_Q_without_NA, distr = "lnorm",  method = c("mle"))

all_fits1 <- list(norm_fit1,gamma_fit1,lognormal1)
```

- Get the distributions parameters of these new fits using summary

```
summary(norm_fit1)

## Fitting of the distribution ' norm ' by maximum likelihood
## Parameters :
##              estimate Std. Error
## mean 358.9568              8.553190
## sd          104.0541          6.048021
## Loglikelihood:                -897.4497          AIC:        1798.899          BIC:        1804.894
## Correlation matrix:
##              mean sd
## mean           1     0
## sd             0     1

summary(gamma_fit1)

## Fitting of the distribution ' gamma ' by maximum likelihood
## Parameters :
##                      estimate         Std. Error
## shape 11.41056780 1.288768928
## rate           0.03178586 0.003666481
## Loglikelihood:                -896.1521          AIC:        1796.304          BIC:        1802.299
## Correlation matrix:
##                       shape              rate
## shape 1.0000000 0.9775398
## rate          0.9775398 1.0000000

summary(lognormal1)

## Fitting of the distribution ' lnorm ' by maximum likelihood
## Parameters :
##                   estimate Std. Error
## meanlog 5.838737 0.02511498
## sdlog           0.305537 0.01775812
## Loglikelihood:                -898.6546          AIC:        1801.309          BIC:        1807.304
## Correlation matrix:
##                   meanlog sdlog
## meanlog                  1          0
## sdlog                    0          1
```

- Compare the parameters and goodness-of-fit measures obtained based on the two parameter estimation methods. Comment.

We observe that when the estimation method is changed from **Moment Method Estimation (mme)** to Maximum **Likelihood Estimation (mle)** the log likelihood, AIC and BIC values changes but not significantly and still after changing of estimation method the AIC and BIC for Gamma distribution is the lowest **(1796.304 & 1802.299)** respectively. Hence till this point of the analysis It is of the opinion that Annual data set is best fitted by

Gamma distribution.

1.2. ANALYSIS ON MONTHLY STREAM FLOW

Depending on your Chamilo group number, you will work on the following monthly data:

- Plot the series of monthly discharge for month **February**.

  : Note our Group is working for **February Flows.**

```
plot(data$Year, data$Feb, xlab = 'Year', ylab = 'February Flows  (m3/sec)', main = "February Stream Flows VS Years", col.lab =  "Black", pch = 18, cex= 1,col = viridis(n=100, option = "H"), bg = "Yellow", font.lab = 2, bty= "n", cex.axis = 0.9, ylim =  c(0,1600), xlim = c(1863,2019))
```

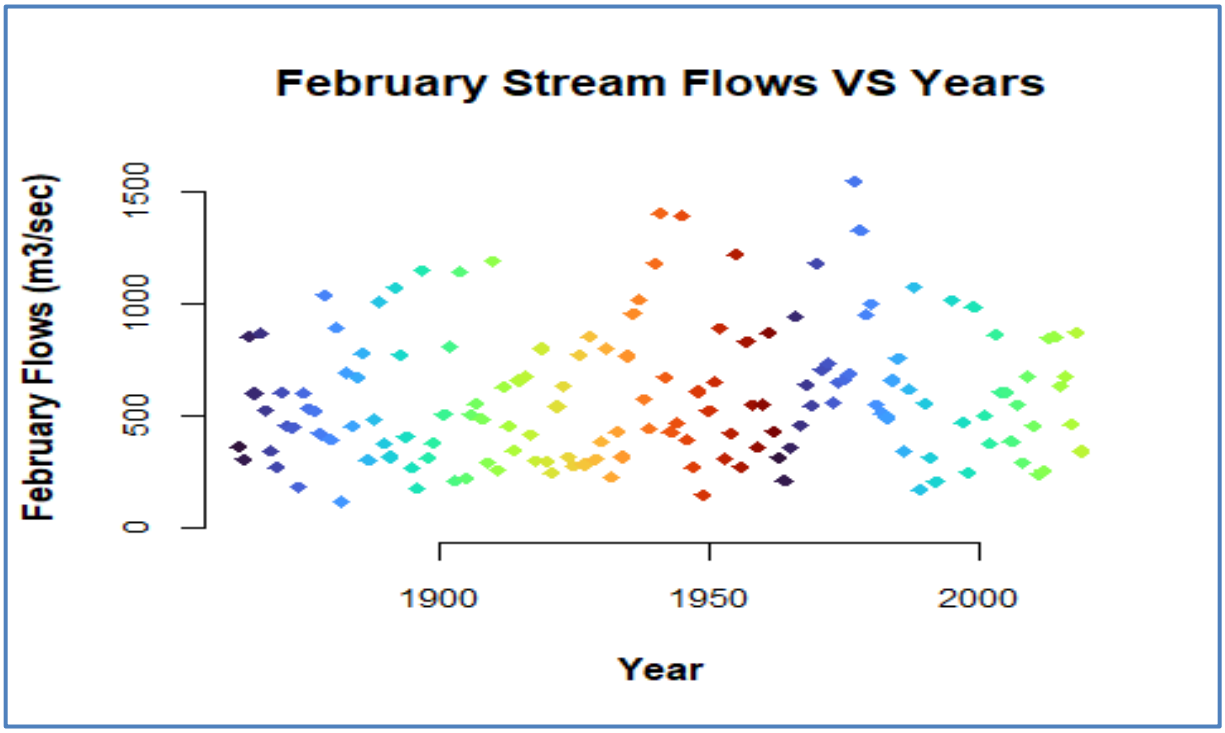


Figure 12 : February Stream Flow VS Years

- Analyse the distribution of monthly discharge in month **February.**

```
Feb_Q_without_NA <- data$Feb[which(!is.na(data$Feb))]
print(Feb_Q_without_NA)
```

```
##   [1]  361.07  303.45  850.00  599.64  866.79  521.03  343.57
267.50  602.50
##  [10]  453.79  447.50  180.54  601.43  529.66  521.43  423.21
1037.14  392.76
##  [19]  892.14  115.71  690.36  450.34  672.14  776.79  300.00
482.76 1008.21
##  [28]  375.00  315.89 1070.34  771.43  406.43  268.57  177.59
1146.79  308.75
##  [37]  376.79  506.07  807.14  210.54 1144.48  216.96  506.07
552.50  486.21
##  [46]  288.93 1189.29  255.00  626.55  451.07  342.86  657.50
673.45  411.79
##  [55]  298.21  798.57  292.41  241.96  541.07  630.36  314.83
275.36  771.07
##  [64]  282.86  850.00  304.29  383.93  799.64  225.00  425.71
315.71  765.00
##  [73]  955.17 1017.14  570.00  442.86 1178.97 1401.79  671.07
425.71  466.90
##  [82] 1392.86  392.50  267.68  606.90  145.71  521.61  650.36
888.62  308.21
##  [91]  420.36 1218.57  270.17  831.07  545.00  361.43  548.97
867.86  429.79
## [100]  313.61  209.79  355.68  943.21  455.32  637.28  542.79
1180.00  706.43
## [109]  735.76  557.43  645.46  658.57  686.17 1545.00 1327.50
952.14  997.24
## [118]  545.54  508.50  490.39  657.59  755.32  340.04  616.71
1072.41  167.14
## [127]  555.36  312.04  203.83 1011.67  469.83  248.11  986.15
502.86  375.72
## [136]  859.22  603.35  605.87  384.74  547.72  288.01  673.45
455.90  237.97
## [145]  256.47  845.85  851.70  628.52  674.91  459.12  872.13
341.63
```

```
plotdist(Feb_Q_without_NA, histo = TRUE, demp = TRUE, col =
viridis(n=5, option = "E"), pch = 20, type = "b" , lty = "solid",  font.lab = 2, bty= "n", cex.axis = 0.9,)
```

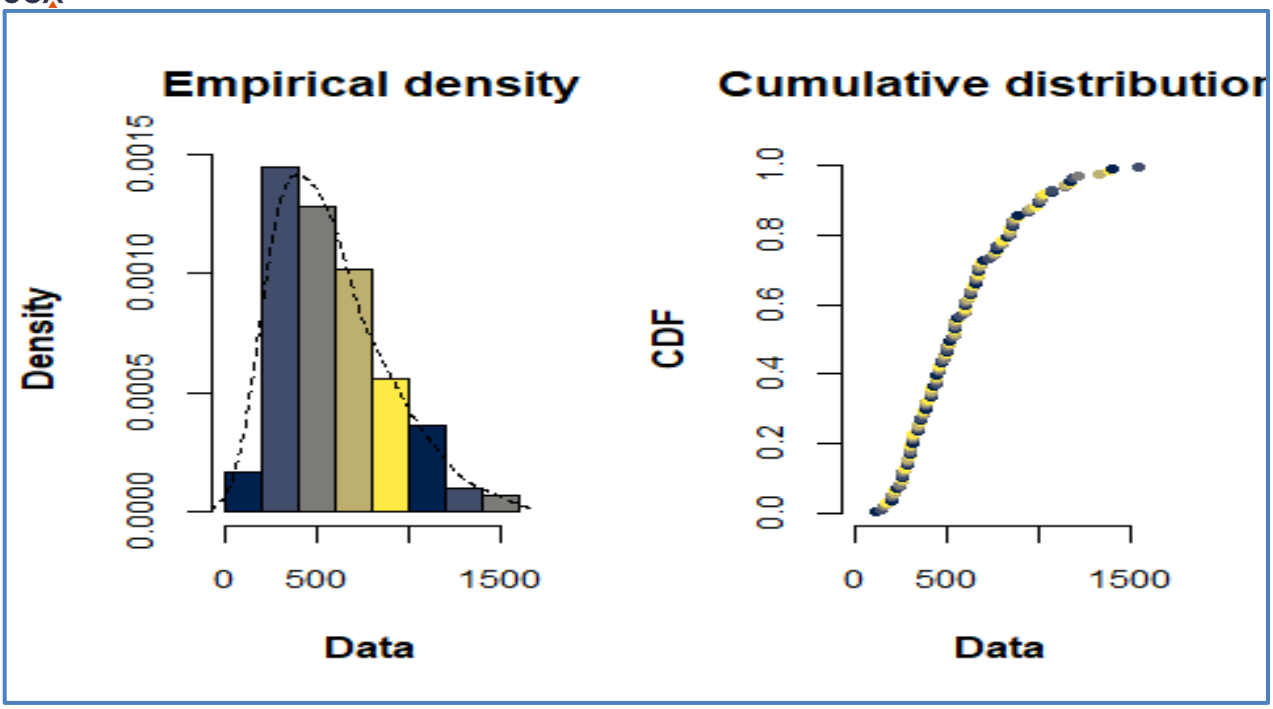


Figure 13 : Emperical Distribution and CDF for February Flows

```
summary(data$Feb)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   115.7   343.4   525.6   581.6   766.5  1545.0       2
```

From the above graph we can conclude that the February flow data corresponds to **positively skewed distribution** as majority of values cluster to the left tail and right tail is longer. The summary of the data further confirms this statement as for the case of positively skewed data **mean > median.**

- Fit a normal, log normal and gamma distribution to these data by the moment matching estimation (MME) method. Plot the corresponding results. Comment.

```
norm_fit2 <- fitdist(Feb_Q_without_NA, distr = "norm", method = "mme")

gamma_fit2 <- fitdist(Feb_Q_without_NA, distr = "gamma", method = c("mme"))

lognormal2 <- fitdist(Feb_Q_without_NA, distr = "lnorm", method = c("mme"))

all_fits2 <- list(norm_fit2, gamma_fit2, lognormal2)

denscomp(all_fits2, datacol = viridis(5, option = "H"), fitlwd = 2)
```

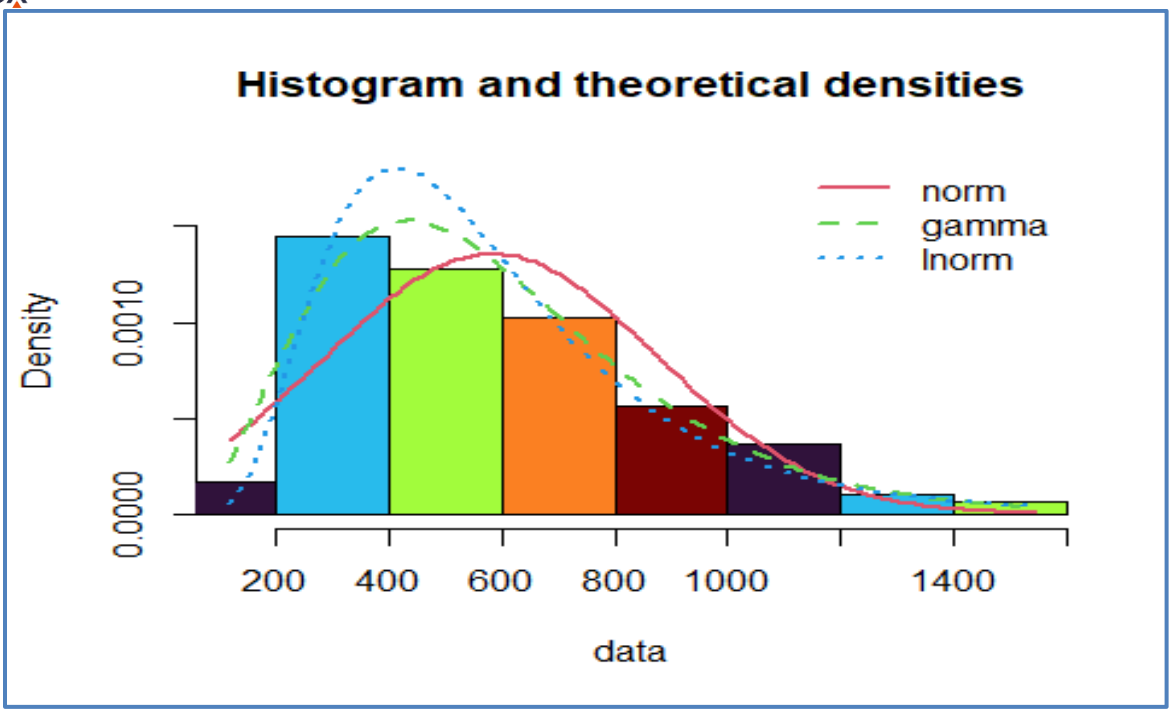


Figure 14 : Theoretical Distributions for February Flows

```
cdfcomp(all_fits2, datapch = 20, datacol = viridis(10, option =  "H"), fitlwd = 1)
```

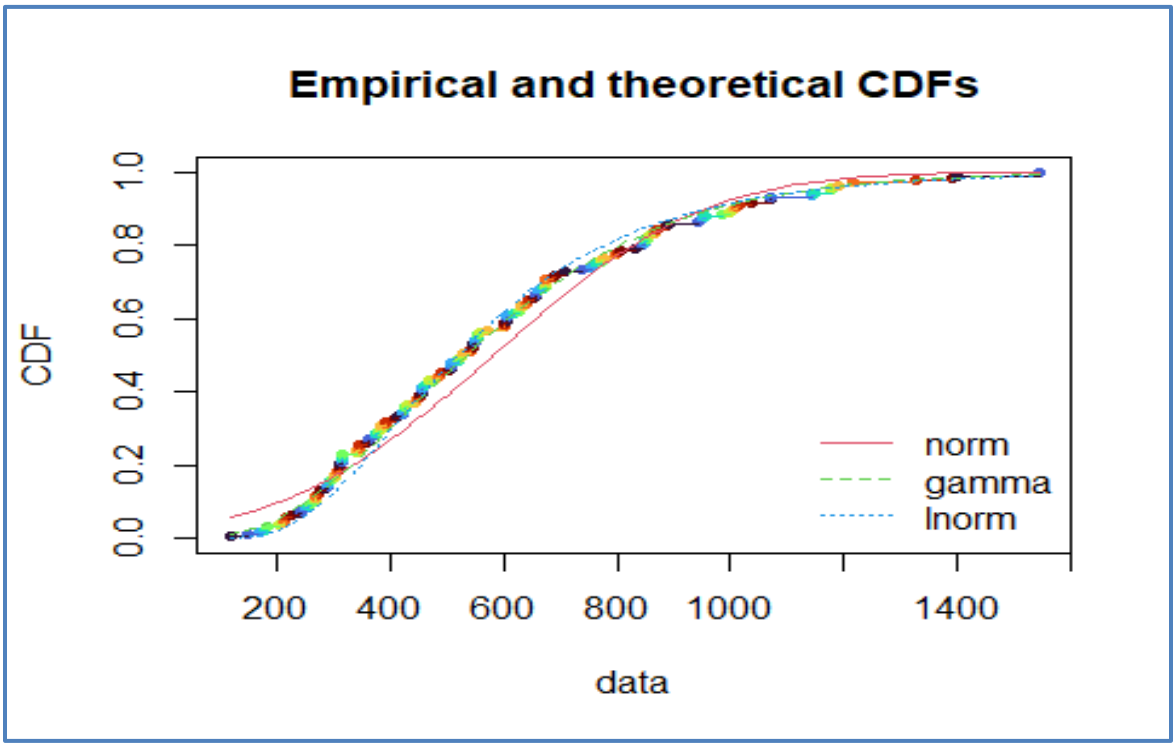


Figure 15: Empirical and Theoretical CDFs for February Flows

```
qqcomp(all_fits2, fitpch = 19, fitcol = viridis(5, option = "H"))
```

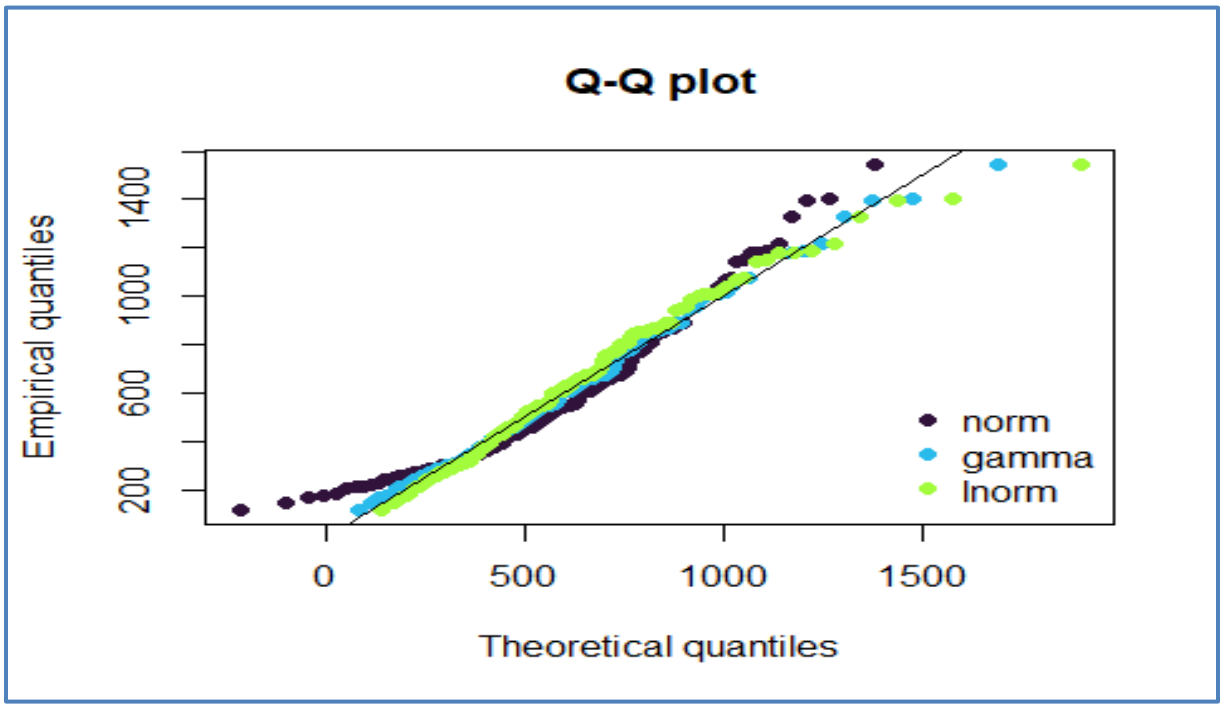


Figure 16 : Q-Q plot for theoretical distribution of February Flows.

```
ppcomp(all_fits2, fitpch = 20, fitcol = viridis(3, option = "C"))
```

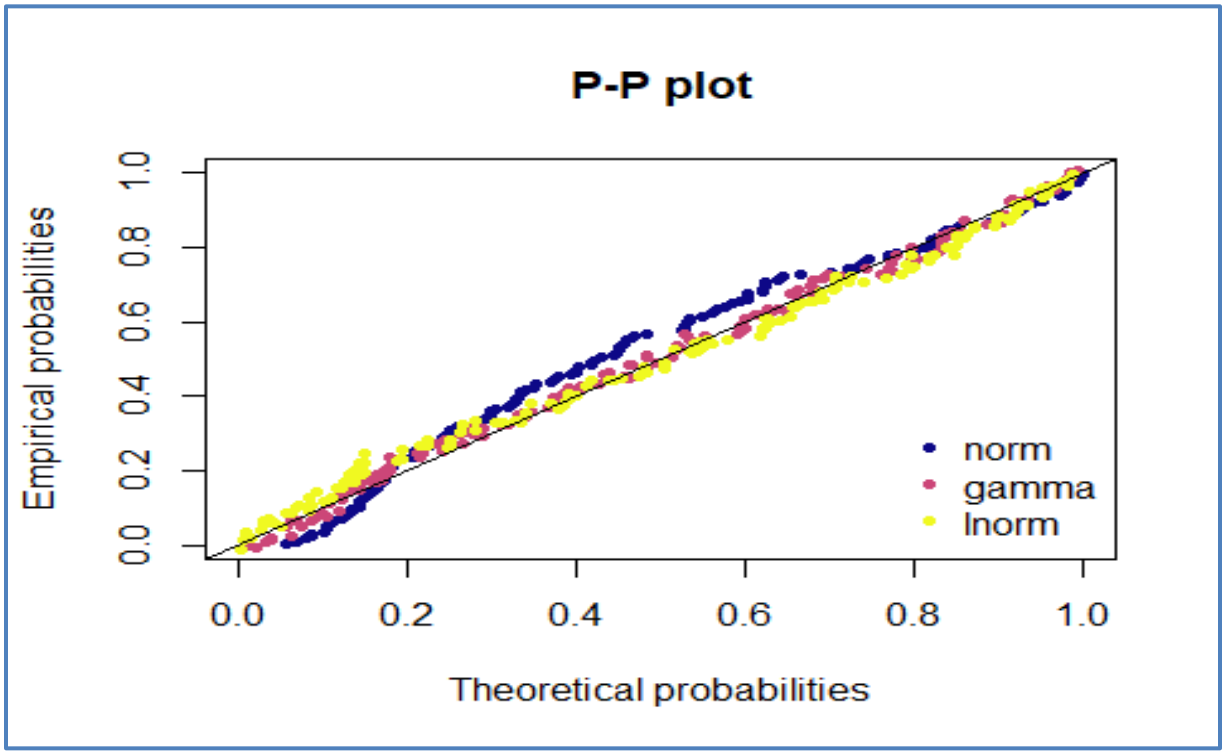


Figure 17 : P-P Plot for Theoretical Distribution of February Flows

From the plots we see that normal distribution is not a good fit for February flows but graphically it can be concluded that gamma and log normal almost predicts the same.

- Choose one of the three fitted distributions as representative of the monthly discharge at Blois. Justify your choice. Give the distribution parameters of such distribution.

```
summary(norm_fit2)

## Fitting of the distribution ' norm ' by matching moments
## Parameters :
##            estimate
## mean 581.6043
## sd         293.7648
## Loglikelihood:              -1079.461          AIC:       2162.922          BIC:       2168.97

summary(gamma_fit2)

## Fitting of the distribution ' gamma ' by matching moments
## Parameters :
##                   estimate
## shape 3.919727682
## rate        0.006739509
## Loglikelihood:              -1064.067          AIC:       2132.135          BIC:       2138.182

summary(lognormal2)

## Fitting of the distribution ' lnorm ' by matching moments
## Parameters :
##                   estimate
## meanlog 6.2521749
## sdlog           0.4766875
```

In order to choose between the Gamma and Log normal, measures of goodness-fit-test were computed and by the aid of it, **Gamma distribution** is selected as representative for the month of February flows. The choice was made on the basis of lower values of AIC and BIC which for the gamma distribution were **2132.135** & **2138.182** respectively, which in comparison to other two fitted distribution (Normal and Log normal) are lower .The parameters for representative distribution are **Shape = 3.919727682** and **Rate = 0.006739509.**

- In the first part of **February** this year, the mean Loire River discharge at Blois was around 165 m^3/s. supposing that the average at the end of the month is 165 m^3/s, what is its probability? What must be the average discharge in the next 15 days to reach the expected value in **February** at Blois? Comment.

```
pgamma(165, shape = 3.919727682, rate = 0.006739509)

## [1] 0.02990726
```

The probability of getting **165 m3/sec** in the month of February this year is approximately **3 %** and the solution to find the average discharge is as follows:

Considering a discharge of 165 m3/s in the first part of February, to get the average discharge (x) in the next 15 days to reach the expected value in February at Blois (210.6 m3/s), we evaluate thus;

$$\mu(Feb)=[16\,\mu(Feb\,2020)+15\,x)/28$$

where μ(Feb 2020)=165 m3/s μ(Feb) = 581.60 m3/s and x is the expected discharge of the next 15 days found to be **942 m3/s**.

- Give an estimate of the probability of the current monthly value of 170 m^3/s.

```
pgamma(170, shape = 3.919727682, rate = 0.006739509)

## [1] 0.03276672
```

The probability of getting 170m3/sec is **3.27 %.**

2. PART II – STATISTICAL TESTS

Objectives of Part II

- To become familiar with the use of statistical tests with R/RStudio
- To apply conformity and homogeneity tests on the mean and the variance: Student's, Wilcox (or Mann-Whitney) and Fisher-Snedecor tests
- To apply goodness-of-fit tests: Pearson's χ^2 and Kolmogorov-Smirnov's tests

Short reminder on statistical tests

The different types of statistical tests depending on the objectives:

- **Conformity tests** allow the comparison of a characteristic of a sample (such as mean, standard deviation, …) with a reference value (or norm)
- **Homogeneity tests** allow comparing two data samples
- **Goodness-of-fit tests** check if a given sample can be considered as coming from a specific parent population (distribution)
- **Stationarity tests** include trend tests for highlighting a slow drift in the process and change point tests for identifying abrupt changes from a given date
- **Autocorrelation tests** verify whether there is a statistical temporal dependence between datasets
- **Parametric tests**: we know or assume that a particular parametric distribution (such as the normal law) is an appropriate representation for the data and/or the test statistic. A parametric test concerning a physical process of interest can reduce to a test pertaining to a distribution parameter (such as the normal mean or the normal standard deviation)
- **Non-parametric (or distribution-free) tests**: are conducted without assumptions that particular parametric forms are appropriate in a given situation

The **5 steps** of any hypothesis test:

1. Identify a **test statistics** that is appropriate to the data and the question at hand
2. Define a **null hypothesis H_0**
3. Define an **alternative hypothesis H_1**
4. Obtain the **null distribution**, which is the sampling distribution for the test statistics if the null hypothesis H_0 is true
5. Compare the observed test statistics to the null distribution. If the test statistic falls in a sufficiently improbable region of the null distribution, **H_0 is rejected**. If the test statistic falls within the range of ordinary values described by the null distribution, the test statistic is seen as consistent with H_0, which is then **not** rejected.

The significance level α (also called the rejection level or the test level) corresponds to a threshold probability (i.e. an accepted probability) of incorrectly rejecting the null hypothesis when it is in fact true. It is also known as the type I error and is associated to one (in the case of a one-sided test) or two (in the case of a two-sided test) critical value(s)

delimitating the rejection interval(s).

Each test statistic is characterized by a given test variable whose observed value is associated to a probability called the p-value. The p-value corresponds to the observed rejection probability of the null hypothesis (i.e. the black areas in the figure at left).

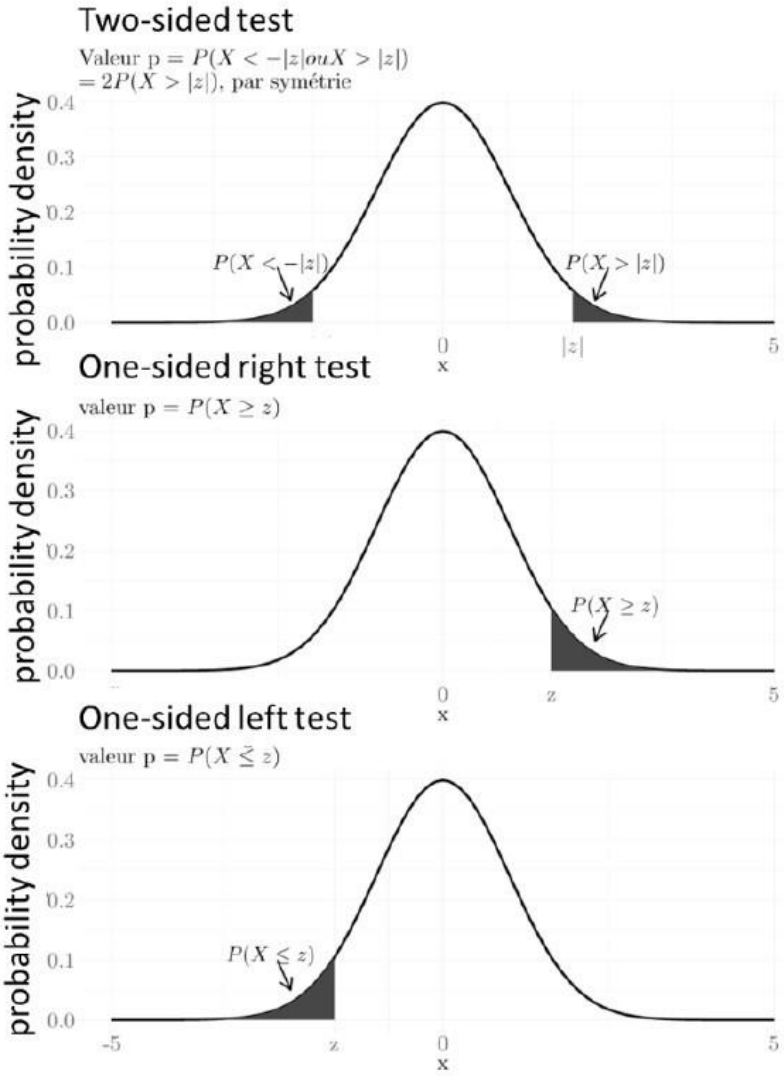


Figure modified from https://fr.wikipedia.org/wiki/Valeur_p#/media/Fichier:Valeur-p.jpg

In a **one-sided test**, the null hypothesis is rejected at a level of significance α if the p-value is greater than $1 - \alpha$ (the test value is greater than the critical value in a one-sided right test) or lower than α (the test value is lower than the critical value in a one-sided left test).

In a **two-sided test**, the null hypothesis is rejected at a level of significance α if the p-value is lower than $\alpha/2$ (the test value is lower than the left critical value) or greater than $1 - \alpha/2$ (the test value is greater than the right critical value).

The **confidence interval** $(1-\alpha)$ of the test corresponds to the uncertainty on the observed values at a given significance level α. The null hypothesis is rejected if the confidence

interval excludes the null hypothesis value. α = 5% is often chosen as significance level, corresponding to a 95% confidence interval.

2.1. CONFORMITY TEST ON ANNUAL DATA

A diatribe pits two hydrological experts against each other. The first states that the module (i.e. the interannual mean value) of the Loire River at Blois is 370 m3/s. The other one states that it is rather less than this value. We use a statistical test based on the available data set in order to determine who is right.

1. Calculate the module [m3/s] of the Loire River at Blois.

```
mean(annual_Q_without_NA)

## [1] 358.9568
```

A common conformity test used for comparing the mean of a normally distributed sample to a reference value is the Student's t-test. The t-test variable is defined as

$$t = \sqrt{n}\frac{m-\mu}{s}$$

Where n is the sample size, m is the mean of the sample values, μ is the reference (theoretical) mean value and s is the standard deviation of the sample values. The number of degrees of freedom is $n-1$.

In R, the Student test is applied using the function t.test(). Look at the help to understand the syntax of this function and how it works: type?t.testin the prompt.

2. Give H_0 (the null hypothesis) and H_1 (the alternative hypothesis) in the case of the t-test.

   H_0, μ = 370 m3/sec.

   H_1, μ <= 370 m3/sec.

3. Do you need a single or a two-sided test for answering the question?

   We need a **single sided** test for answering this question.

4. Perform the following Student's t-test and comment the results.

```
# level of significance = 5%
t.test(data$Annual, mu=370, alternative= "less", conf.level=0.95)

##
##     One Sample t-test
##
## data:         data$Annual
## t = -1.2867, df = 147, p-value = 0.1001
## alternative hypothesis: true mean is less than 370
## 95 percent confidence interval:
```

```
##            -Inf 373.1629
## sample estimates:
## mean of x
##     358.9568
```

The student test gives **-1.2867,** as test statistic and test has 147 degree of freedom, as the value of test statistic is small so we can roughly infer that sample mean is equal to 370m3/sec The p-value corresponding to test statistic is **0.10** or **10 %.**

5. Is H_0 rejected at a level of significance α of 5%?

Since the **P-value = 10%** and is greater than the level of significance α of 5% , so we **fail to reject the null hypothesis** and we can say that there is no significant difference in the sample and theoretical mean i.e μ = 370 m3/sec.

6. Same questions for a two-sided test.

In the case of two sided test . The null and alternative hypothesis could be written as following:

H_0, μ = 370 m3/sec

H_1, μ != 370 m3/sec.

```
# level of significance = 5%
t.test(data$Annual, mu=370, alternative= "two.sided", conf.level=0.95)

##
##     One Sample t-test
##
## data:         data$Annual
## t = -1.2867, df = 147, p-value = 0.2002
## alternative hypothesis: true mean is not equal to 370
## 95 percent confidence interval:
##     341.9963 375.9173
## sample estimates:
## mean of x
##
```

Since the **P-value = 20%** and is greater than the level of significance $\alpha/2$ = 2.5% and less than $1-\alpha/2$ = 97.5%, so we **fail to reject the null hypothesis** and we can say that there is no significant difference in the sample and theoretical mean. i.e μ = 370 m3/sec

2.2. HOMOGENEITY TESTS ON ANNUAL DATA

The inhabitants of Blois are divided as to whether or not the regime of the Loire River has changed over time. Some say that the module has changed over the last 30 years, while others say that it is the inter-annual variability that has changed. As a hydrologist, you suggest tests to them to provide a scientific basis for the discussion. Indeed, we will use two statistical tests, one applied on the **module (i.e. the inter annual mean)** and the other

applied on the **variance**, both calculated over two different periods: 1950-1979 and 1990-2019.

First, we will test the module. An appropriate test for comparing the means of two samples is the Student's t-test if the two samples are normally distributed and have the same variance. In this case, the t-test variable is defined as:

$$t = \frac{m_1 - m_2}{\sqrt{\frac{s^2(n_1+n_2)}{n_1 n_2}}}$$

Where n_1 and n_2 are the sample sizes of the two samples, m_1 and m_2 are the means of the two sample values ad s is calculated as:

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

Where s_1 and s_2 are the variances of the two sample values. The number of degrees of freedom is n_1+n_2-2.

1. Calculate the module [m^3/s] of the Loire River discharge during the two periods 1950-1979 and 1990-2019.

```
mean(data$Annual[88:117] , na.rm = T) # for period for, 1950-1979

## [1] 359.1347

mean(data$Annual[128:154], na.rm = T) # for period for, 1990-2019

## [1] 320.3168
```

2. Give H_0 and H_1 in the case of this t-test.

H_0, $\mu 1 = \mu 2$

H_1, $\mu 1 != \mu 2$

3. Test the homogeneity of the Loire River modules during 1950-1979 and 1990-2019:

```
t.test(data$Annual[88:117], data$Annual[128:154], conf.level=0.95)

##
##     Welch Two Sample t-test
##
## data:          data$Annual[88:117] and data$Annual[128:154]
## t = 1.4762, df = 48.971, p-value = 0.1463
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##     -14.02552         91.66122
## sample estimates:
## mean of x mean of y
##  359.1347  320.3168
```

4. Is it rejected or not at a level of significance α = 5%?

Since the **P-value = 14%** and is greater than the level of significance $\alpha/2$ of 2.5% and less than $1-\alpha/2$ of 97.5%, so we **fail to reject the null hypothesis** and we can say that there is no significant change in the module of two samples i.e $\mu 1 = \mu 2$.

Secondly, we will test the variance, which is an indicator of interannual variability. An appropriate test for comparing the two samples variances is the Fisher-Snedecor test. The F variable is defined as:

$$F = \frac{\frac{n_1 s_1^2}{n_1 - 1}}{\frac{n_2 s_2^2}{n_2 - 1}}$$

where n_1 and n_2 are the sample sizes of the two samples, m_1 and m_2 are the means of the two sample values, s_1 and s_2 are the variances of the two sample values. The number of degrees of freedom is $n_1 + n_2 - 1$.

4. Calculate the variance [m^3/s] of the Loire River discharge during the two periods 1950-1979 and 1990-2019.

```
var(data$Annual[88:117])                # for period for, 1950-1979

## [1] 10547.25

var(data$Annual[128:154], na.rm = T) # for period for, 1990-2019

## [1] 7477.205
```

5. Give H_0 and H_1 in the case of the Fisher-Snedecor test.

H_0, $s_1^2 = s_2^2$

H_1, $s_1^2 \neq s_2^2$

6. Test the homogeneity of the variances of the Loire River discharge during 1950-1979 and 1990-2019:

```
var.test(data$Annual[88:117], data$Annual[128:154], conf.level=0.95)

##
##     F test to compare two variances
##
## data:         data$Annual[88:117] and data$Annual[128:154]
## F = 1.4106, num df = 29, denom df = 21, p-value = 0.4184
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##     0.6089238 3.0982930
## sample estimates:
```

```
## ratio of variances
##                 1.410588
```

What is your conclusion considering as before a significance level α = 5%?

Since the **P-value = 41%** and is greater than the level of significance $\alpha/2$ of 2.5% and less than 1-$\alpha/2$ of 97.5% , so we **fail to reject the null hypothesis** and we can say that there is no significant change in the inter annual variability of two samples i.e true ratio of variances is equal to 1 OR $s1^2 = s2^2$

2.3. GOODNESS-OF-FIT TESTS ON MONTHLY DATA

In Part I you fitted a Normal, a LogNormal and a Gamma law to the data series of monthly mean river discharge [m^3/s] observed on the Loire River at Blois.

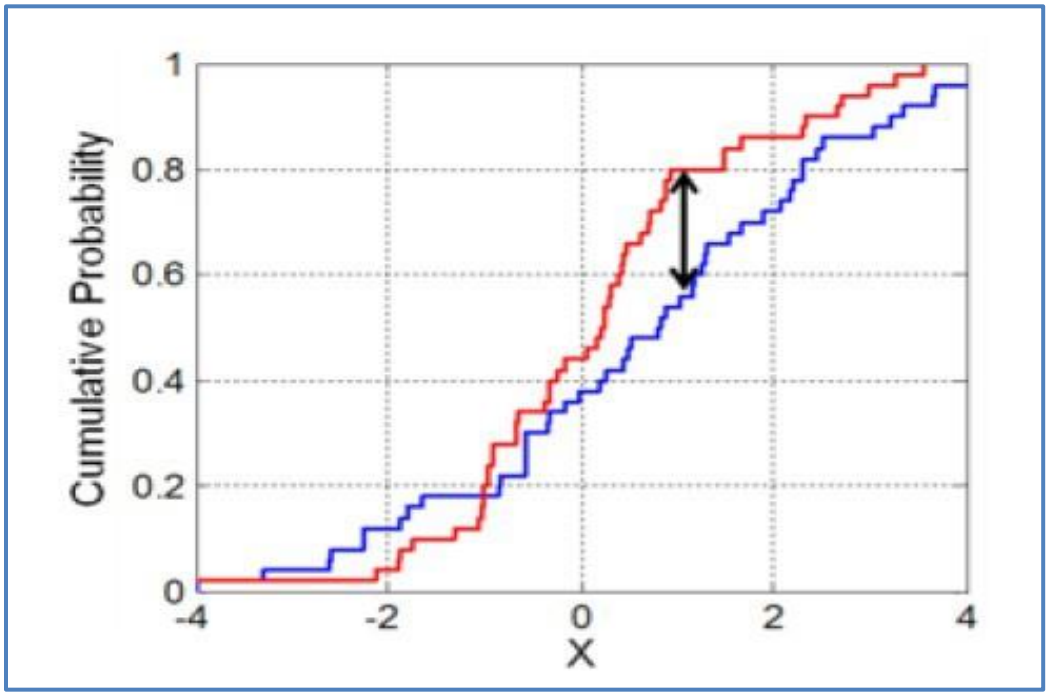
1. Recall the fitted parameter values for the three statistical laws by the maximum likelihood estimation (MLE) method.

```
summary(norm_fit1)

## Fitting of the distribution ' norm ' by maximum likelihood
## Parameters :
##            estimate Std. Error
## mean 358.9568            8.553190
## sd        104.0541        6.048021
## Loglikelihood:            -897.4497        AIC:      1798.899        BIC:      1804.894
## Correlation matrix:
##            mean sd
## mean          1    0
## sd            0    1

summary(gamma_fit1)

## Fitting of the distribution ' gamma ' by maximum likelihood
## Parameters :
##                   estimate        Std. Error
## shape 11.41056780 1.288768928
## rate          0.03178586 0.003666481
## Loglikelihood:            -896.1521        AIC:      1796.304        BIC:      1802.299
## Correlation matrix:
##                   shape             rate
## shape 1.0000000 0.9775398
## rate        0.9775398 1.0000000

summary(lognormal1)

## Fitting of the distribution ' lnorm ' by maximum likelihood
## Parameters :
##                  estimate Std. Error
## meanlog 5.838737 0.02511498
```

```
## sdlog       0.305537 0.01775812
## Loglikelihood:  -898.6546   AIC:  1801.309   BIC:  1807.304
```

```
## Correlation matrix:
##                meanlog sdlog
## meanlog              1     0
## sdlog                0     1
```

An appropriate test for the goodness-of-fit assessment is the Kolmogorov-Smirnov test, whose test variable D is the maximum distance between the empirical probability distribution and the fitted theoretical one:



The general expression of the Kolmogorov-Smirnov test in R is the following:

ks.test(empirical_dataset, “prob_distribution”, par1, par2)

e.g. for the gamma law in the case of October:

OctQ_withoutNA <- DataOct¿Oct))] ks.test(OctQ_withoutNA, “pgamma”, 1.8312, 0.0087)

```
ks.test(annual_Q_without_NA, "pnorm", 358.9568, 104.0541)

##
##      Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:          annual_Q_without_NA
## D = 0.067708, p-value = 0.5061
## alternative hypothesis: two-sided

ks.test(annual_Q_without_NA, "pgamma", 11.41056780, 0.03178586)
```

```
##
##      Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:          annual_Q_without_NA
## D = 0.061372, p-value = 0.6329
## alternative hypothesis: two-sided ks.test(annual_Q_without_NA, "plnorm",

5.838737 , 0.305537)

##
##      Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:          annual_Q_without_NA
## D = 0.079475, p-value = 0.3073
```

2. What is the result of a Kolmogorov-Smirnov test on the three statistical laws fittings?

From the KS test we conclude the same as we concluded from Section **Distribution fitting on annual stream flow** through **Graphical representation (Q-Q plots)** and through **measures of goodness-of-fit** that the sample distribution fits with the gamma distribution. Here we affirm the same conclusion by comparing the D-values for the 3 distribution, as the D-value is the least for gamma distribution i.e. **D = 0.061372** we can say that the sample distribution come from the gamma distribution.

2.4. WORK ON OTHER MONTHLY DATA AND WRITE A REPORT

Continue working with the month corresponding to your group number.

- Perform a goodness-of-fit test for the retained fitted distribution of Part I.

```
summary(norm_fit2)

## Fitting of the distribution ' norm ' by matching moments
## Parameters :
##             estimate
## mean 581.6043
## sd          293.7648
## Loglikelihood:              -1079.461          AIC:       2162.922          BIC:       2168.97

summary(gamma_fit2)

## Fitting of the distribution ' gamma ' by matching moments
## Parameters :
##                     estimate
## shape 3.919727682
## rate          0.006739509
## Loglikelihood:              -1064.067          AIC:       2132.135          BIC:       2138.182

summary(lognormal2)

## Fitting of the distribution ' lnorm ' by matching moments
## Parameters :
##                     estimate
##   meanlog 6.2521749
##   sdlog         0.4766875
##   Loglikelihood:              -1065.74          AIC:       2135.48          BIC:       2141.528

ks.test(Feb_Q_without_NA, "pnorm", 581.6043 , 293.7648)

##
##     Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:          Feb_Q_without_NA
## D = 0.098582, p-value = 0.1042
## alternative hypothesis: two-sided

ks.test(Feb_Q_without_NA, "pgamma", 3.919727682, 0.006739509)

##
##     Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:          Feb_Q_without_NA
## D = 0.051622, p-value = 0.8127
## alternative hypothesis: two-sided
```

```
ks.test(Feb_Q_without_NA, "plnorm", 6.2521749, 0.4766875)
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  Feb_Q_without_NA
## D = 0.081593, p-value = 0.2637
## alternative hypothesis: two-sided
```

From the goodness Fit test we conclude that the Gamma Distribution fits the February Flow data which is the same conclusion as derived in the Part 1 of the report. The conclusion has been made on the fact that the value of **D = 0.051622** is the lowest for gamma distribution.

- Perform a homogeneity test on your study monthly discharge series, considering the following periods: 1910-1939, 1950-1079, 1990-2019 and a reference global period 1863-2019.

```
# Sample Mean
mean(data$Feb[48:77] , na.rm = T) # For 1910-1939

## [1] 536.9553

mean(data$Feb[88:117] , na.rm = T) # For 1950-1079

## [1] 677.1387

mean(data$Feb[128:154] , na.rm = T) # For 1990-2019

## [1] 548.1588
```

Null hypothesis & Alternative Hypothesis:

For T test.

H0, $\mu 1 = \mu 2$

H1, $\mu 1 != \mu 2$

For Fisher-Snedecor Test

H0, $s_1^2 = s_2^2$

H1, $s_1^2 != s_2^2$

```
# Student Test on Mean.
t.test(data$Feb[48:77],data$Feb, conf.level=0.95) # For 1910-1939

##
##     Welch Two Sample t-test
##
## data:         data$Feb[48:77] and data$Feb
## t = -0.8341, df = 44.717, p-value = 0.4087
## alternative hypothesis: true difference in means is not equal to 0
```

```
# Student Test on Mean.
##     -152.48238           63.18437
## sample estimates:
## mean of x mean of y
##     536.9553        581.6043

t.test(data$Feb[88:117],data$Feb, conf.level=0.95) # For 1950-1079

##
##     Welch Two Sample t-test
##
## data:         data$Feb[88:117] and data$Feb
## t = 1.485, df = 38.845, p-value = 0.1456
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##     -34.60825 225.67690
## sample estimates:
## mean of x mean of y
##     677.1387        581.6043

t.test(data$Feb[128:154],data$Feb, conf.level=0.95) # For 1990-2019

##
##     Welch Two Sample t-test
##
## data:         data$Feb[128:154] and data$Feb
## t = -0.62822, df = 38.813, p-value = 0.5335
## alternative hypothesis: true difference in means is not equal to 0
```

```
##     -141.14731           74.25631
## sample estimates:
## mean of x mean of y
##     548.1588        581.6043
```

For all three periods the **P-Values** are greater than $\alpha/2$= 2.5% and less than 1-$\alpha/2$ = 97.5 at a significance level of α = 5%. So for all the three periods we conclude that there is no significant change in the mean.

```
# Fisher-Snedcor Test on Variability
var.test(data$Feb[48:77],data$Feb, conf.level=0.95)# For 1910-1939

##
##     F test to compare two variances
##
## data:         data$Feb[48:77] and data$Feb
## F = 0.7922, num df = 29, denom df = 151, p-value = 0.4695
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##     0.4730959 1.4817591
## sample estimates:
## ratio of variances
##                  0.7922013

var.test(data$Feb[88:117],data$Feb, conf.level=0.95)# For 1950-1079

##
##     F test to compare two variances
##
## data:         data$Feb[88:117] and data$Feb
## F = 1.2319, num df = 29, denom df = 151, p-value = 0.4198
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##     0.7357006 2.3042498
## sample estimates:
## ratio of variances
##                    1.231934

var.test(data$Feb[128:154],data$Feb, conf.level=0.95)# For 1990-2019

##
##     F test to compare two variances
##
## data:         data$Feb[128:154] and data$Feb
## F = 0.67727, num df = 25, denom df = 151, p-value = 0.2544
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##     0.3934584 1.3315074
## sample estimates:
## ratio of variances
##                  0.6772687
```

For all three periods the P-Values are greater than $\alpha/2$= 2.5% and less than 1-$\alpha/2$ = 97.5% at a significance level of α = 5%. So for all the three periods we conclude that there is no significant change in the variability.

- Complete the report from Part I with:
  - On the basis of the goodness-of-fit test performed above, how do you evaluate your fitting? What is the expected value of the Loire River discharge for your studied month? What is the 10-year return value? What is the probability of the 2019 value?

From the goodness fit test we conclude that the Gamma Distribution fits the February Flow data which is the same conclusion as derived in the ***section 1.2*** **ANALYSIS ON MONTHLY STREAM FLOW** *from the graphical and measures of goodness-of-fit.*

```
summary(gamma_fit2)

## Fitting of the distribution ' gamma ' by matching moments
## Parameters :
##                   estimate
## shape 3.919727682
## rate        0.006739509
## Loglikelihood:              -1064.067          AIC:       2132.135          BIC:       2138.182

# expected value from the fitting for the month of February using the  matching moments
# exp_value_gamma <- shape*1/rate  exp_value_Feb_gamma <-
(3.919727682)*(1/0.006739509)  exp_value_Feb_gamma

## [1] 581.6043
```

The expected value(μ) from the gamma distribution is **581.6043 m3/sec**, which approximates the value of **581.6m3/sec** of the sample mean. From this it could be evaluated that gamma distribution is the best fit for the sample distribution.

```
P <- 1-1/10
qgamma(P,3.919727682,0.006739509)

## [1] 975.393
```

The value of discharge for 10 year return period from the fitted distribution is **975.393 m3/sec**

```
pgamma(341.63, 3.919727682,0.006739509)

## [1] 0.214183
```

The probability of getting **341.63 m3/sec** discharge value of 2019 is **21.4%**

An analysis of the stationarity of your studied monthly discharge: what is the best fit on the 3 sub-periods?

```
 # For 1910-1939
norm_fit3 <-            fitdist(data$Feb[48:77], distr = "norm", method = "mme")

gamma_fit3 <- fitdist(data$Feb[48:77], distr = "gamma", method =  c("mme"))

lognormal3 <- fitdist(data$Feb[48:77], distr = "lnorm", method =  c("mme"))

summary(norm_fit3)

## Fitting of the distribution ' norm ' by matching moments
## Parameters :
##              estimate
## mean 536.9553
## sd          257.9225
## Loglikelihood:              -209.1479          AIC:       422.2959          BIC:       425.0982

summary(gamma_fit3)

## Fitting of the distribution ' gamma ' by matching moments
## Parameters :
##                    estimate
## shape 4.334090797
## rate         0.008071604
## Loglikelihood:              -206.0809          AIC:       416.1617          BIC:       418.9641

summary(lognormal3)

## Fitting of the distribution ' lnorm ' by matching moments
## Parameters :
##                    estimate
## meanlog 6.1821116
## sdlog           0.4556387
## Loglikelihood:              -205.6092          AIC:       415.2184          BIC:       418.0208

ks.test(data$Feb[48:77], "pnorm", 536.9553,257.9225)

##
##      Exact one-sample Kolmogorov-Smirnov test
##
## data:          data$Feb[48:77]
## D = 0.16376, p-value = 0.3576
## alternative hypothesis: two-sided

ks.test(data$Feb[48:77], "pgamma", 4.3340907,0.008071604)

##
##      Exact one-sample Kolmogorov-Smirnov test
##
## data:          data$Feb[48:77]
```

```
## D = 0.13534, p-value = 0.5945
## alternative hypothesis: two-sided ks.test(data$Feb[48:77],

"plnorm",6.1821116 ,0.45563)

##
##      Exact one-sample Kolmogorov-Smirnov test
##
## data:          data$Feb[48:77]
## D = 0.15916, p-value = 0.3916
## alternative hypothesis: two-sided

#1950-1979
norm_fit4 <-            fitdist(data$Feb[88:117], distr = "norm", method =  "mme")

gamma_fit4 <- fitdist(data$Feb[88:117], distr = "gamma", method =  "mme")

lognormal4 <- fitdist(data$Feb[88:117], distr = "lnorm", method =  "mme")

summary(norm_fit4)

## Fitting of the distribution ' norm ' by matching moments
## Parameters :
##              estimate
## mean 677.1387
## sd          321.6364
## Loglikelihood:              -215.7708          AIC:       435.5416          BIC:       438.344

summary(gamma_fit4)

## Fitting of the distribution ' gamma ' by matching moments
## Parameters :
##                    estimate
## shape 4.432257084
## rate         0.006545568
## Loglikelihood:              -212.8351          AIC:       429.6701          BIC:       432.4725

summary(lognormal4)

## Fitting of the distribution ' lnorm ' by matching moments
## Parameters :
##                    estimate
## meanlog 6.4161532
## sdlog           0.4510496
## Loglikelihood:              -212.8558          AIC:       429.7115          BIC:       432.5139

ks.test(data$Feb[88:117], "pnorm", 677.1387,321.6364)
```

```
##
##      Exact one-sample Kolmogorov-Smirnov test
##
## data:          data$Feb[88:117]
## D = 0.13039, p-value = 0.6404
## alternative hypothesis: two-sided

ks.test(data$Feb[88:117], "pgamma",4.432257084,0.006545568)

##
##      Exact one-sample Kolmogorov-Smirnov test
##
## data:          data$Feb[88:117]
## D = 0.068713, p-value = 0.997
## alternative hypothesis: two-sided

ks.test(data$Feb[88:117], "plnorm",6.4161532,0.4510496)

##
##      Exact one-sample Kolmogorov-Smirnov test
##
## data:          data$Feb[88:117]
## D = 0.078265, p-value = 0.986
## alternative hypothesis: two-sided

# For 1990-2019
Feb_Q_1990_2019_without_NA <- data$Feb[which(!  is.na(data$Feb[128:154]))]
print(Feb_Q_1990_2019_without_NA)

##     [1]      361.07      303.45      850.00      599.64      866.79      521.03       343.57
602.50      453.79
## [10]       447.50      180.54      601.43      529.66      521.43      423.21     1037.14
392.76      892.14
## [19]       115.71      690.36      450.34      672.14      776.79      300.00       482.76
1008.21

norm_fit5 <-            fitdist(Feb_Q_1990_2019_without_NA, distr = "norm",  method = "mme")

gamma_fit5 <- fitdist(Feb_Q_1990_2019_without_NA, distr = "gamma",  method = c("mme"))

lognormal5 <- fitdist(Feb_Q_1990_2019_without_NA, distr = "lnorm",  method = c("mme"))

summary(norm_fit5)

## Fitting of the distribution ' norm ' by matching moments
## Parameters :
##             estimate
## mean 554.7677
```

```
## sd          236.3011
## Loglikelihood:          -178.9852          AIC:          361.9704          BIC:          364.4865

summary(gamma_fit5)

## Fitting of the distribution ' gamma ' by matching moments
## Parameters :
##                    estimate
## shape 5.511767488
## rate          0.009935271
## Loglikelihood:          -179.0463          AIC:          362.0927          BIC:          364.6089

summary(lognormal5)

## Fitting of the distribution ' lnorm ' by matching moments
## Parameters :
##                    estimate
## meanlog 6.2351867
## sdlog            0.4083204
## Loglikelihood:          -181.6019          AIC:          367.2037          BIC:          369.7199

ks.test(Feb_Q_1990_2019_without_NA, "pnorm",554.7677,236.3011)

##
##      Exact one-sample Kolmogorov-Smirnov test
##
## data:          Feb_Q_1990_2019_without_NA
## D = 0.11923, p-value = 0.8118
## alternative hypothesis: two-sided ks.test(Feb_Q_1990_2019_without_NA,

"pgamma",5.511767488,0.009935271)

##
##      Exact one-sample Kolmogorov-Smirnov test
##
## data:          Feb_Q_1990_2019_without_NA
## D = 0.080227, p-value = 0.9912
## alternative hypothesis: two-sided ks.test(Feb_Q_1990_2019_without_NA,

"plnorm",6.2351867,0.4083204)

##
##      Exact one-sample Kolmogorov-Smirnov test
##
## data:          Feb_Q_1990_2019_without_NA
```

From the above observation we can conclude that for the period **1910-1939**, p-values for all the three distribution are greater than $\alpha / 2$ = 2.5% and less than 1-$\alpha / 2$ = 97.5% at a significance level of α = 5%, so on the basis of lower D value of Gamma Distribution among all other distribution it is considered best for that period, but for the other two periods

1950-1979, 1990-2019 the p-value for the gamma and log normal distribution are greater than 1-α / 2 = 97.5% at a significance level of α = 5%, so we reject the hypothesis that these periods come from those distribution and for the two periods Normal distribution is the best fit.

Compare the parameters of the fitted laws and analyze the test results obtained above: is the series homogeneous.

The parameters for the 3 periods are:

1910-1939

Normal Distribution	*Gamma Distribution*	*Lognormal Distribution*
mean 536.9553	shape 4.334090797	mean log 6.1821116
sd 257.9225	rate 0.008071604	sdlog 0.4556387

1950-1979

Normal Distribution	*Gamma Distribution*	*Lognormal Distribution*
mean 677.1387	shape 4.432257084	meanlog 6.4161532
sd 321.6364	rate 0.006545568	sdlog 0.4510496

1990-2019

Normal Distribution	*Gamma Distribution*	*Lognormal Distribution*
mean 554.7677	shape 5.511767488	meanlog 6.2351867
sd 236.3011	rate 0.009935271	sdlog 0.4083204

From the above summarized data, it can be observed that the parameters for three periods for the monthly flow series vary from each other, but the homogeneity test on the three periods suggest that there is no significant change in the mean and variability and thus the series is homogeneous.

What is the expected value of the Loire River discharge for your studied month considering only the last 30 years for the fitting?

```
Feb_Q_1989_2019_without_NA <- data$Feb[which(!
is.na(data$Feb[127:154]))]
Feb_Q_1989_2019_without_NA

##    [1]    361.07    303.45    850.00    599.64    866.79    521.03    343.57
267.50    453.79
## [10]    447.50    180.54    601.43    529.66    521.43    423.21   1037.14
392.76    892.14
## [19]    115.71    690.36    450.34    672.14    776.79    300.00    482.76
1008.21    375.00
gamma_dist30 <- fitdist(Feb_Q_1989_2019_without_NA,distr =
"gamma",method = "mme")
summary(gamma_dist30)
```

```
## Fitting of the distribution ' gamma ' by matching moments
## Parameters :
##                   estimat
## shape 4.984353354
## rate        0.009304336
## Loglikelihood:              -185.4547          AIC:       374.9094          BIC:       377.5011
```

```
  # exp_value          = (shape*1/rate)
exp_value <- (4.984353354                    * 1/0.009304336)
exp_value
```

```
## [1] 535.7022
```

The expected value of discharge for data series of 30 years calculated by using the parameters of gammma distribution is **535.7022 m3/sec**.

What is the 10-year return value? What is the probability of the 2019 value? Comment.

```
qgamma(P, 4.984353354,0.009304336)
```

```
## [1] 856.9466
```

```
pgamma(212.5, shape = 4.984353354, rate = 0.009304336)
```

```
## [1] 0.05148769
```

The 10 year return value from the distribution of 30 years is **856.9466 m3/sec** The probability of 2019 value of **212.5m3/sec** for 30 year data series gamma distribution is **5%**.

References

Charles Lemarchand, R. R. (2014). Flagship Species Conservation and Introduced Species Invasion : Toxic Aspects Along Loire River (France). (M. L. Soloneski, Ed.) Pesticides - Toxic Aspects.

HYDROPORTAIL.(2015).hydro.eaufrance.Retrievedfrom https://www.hydro.eaufrance.fr/stationhydro/K447001001/synthese

Tutiempo Network, S. (n.d.). Retrieved from tutiempo: https://en.tutiempo.net/climate/ws-72450.html