Tutorial for AP Lab Assignment 3

Assignment : Applied Probability Lab Assignment 3

Subject Code : CS 215

Assignment Submission : **30/03/2023**

Department of Computer Science and Engineering
National Institute of Technology Silchar, Assam, India

Q.1)

Histogram :

- **The purpose of a histogram is to graphically summarize the distribution of a univariate data set.**

The histogram graphically shows the following :

- center (i.e. the location) of the data
- spread (i.e. the scale) of the data
- skewness of the data
- presence of outliers; and
- presence of multiple modes in the data.

  - These features provide strong indications of the proper distributional model for the data.

- The most common form of the histogram is obtained by **splitting the range of the data into equal-sized bins** (called classes).

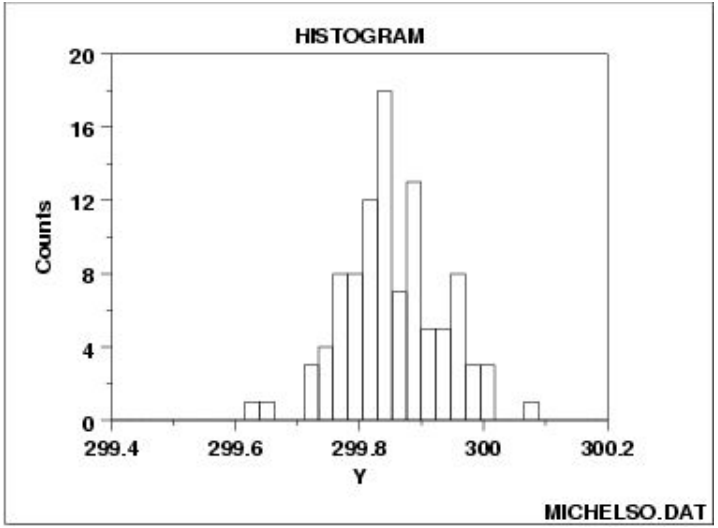


Fig 1 : The above plot is a histogram of the **Michelson speed of light data set**.

Types of tailed histograms :

- **Short-Tailed Histogram :**
  - For a short-tailed distribution, the tails approach zero very fast.
  - Such distributions commonly have a truncated ("sawed-off") look.
  - The classical short-tailed distribution is the uniform (rectangular) distribution in which the probability is constant over a given range and then drops to zero everywhere else, we would speak of this as having no tails, or extremely short tails.

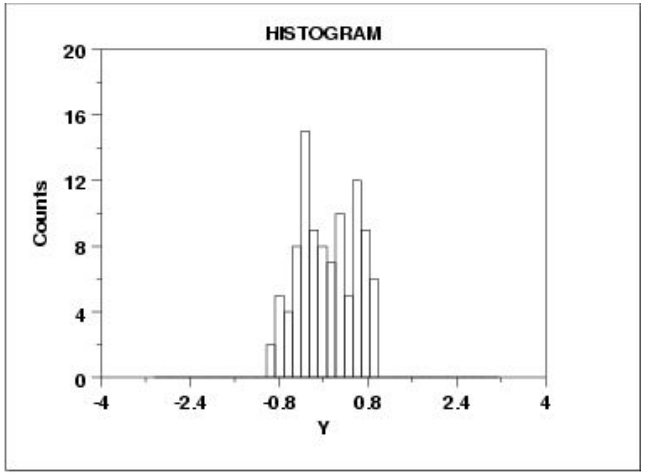


Fig 2: Example of Short-Tailed Histogram

Types of tailed histograms :

- **Sample R code of Short Tailed Histogram :**

  - R creates histogram using hist() function.The basic syntax for creating a histogram using R is : hist(v,main,xlab,xlim,ylim,breaks,col,border)

Example :

```
n = 100
#short tailed: Uniform (on [0,2]) with mu = 1 and sigma = 0.3333;
short <- runif(n,min=0,max=2)

RandomData <- short
title <- "Right tailed Distribution"

std<-sd(RandomData)
m <- mean(RandomData)
hist(RandomData, xlab="Data", freq = FALSE, main=title)

curve(dnorm(x, mean=m, sd=std), col="blue", lwd=2, add=TRUE)
lines(density(RandomData,adjust=3),col = "red", lwd=2)
```

Types of tailed histograms :

- **Long-Tailed Histogram :**
  - For a long-tailed distribution, the tails decline to zero very slowly, and hence one is apt to see probability a long way from the body of the distribution. **The classical long-tailed distribution is the Cauchy distribution.**

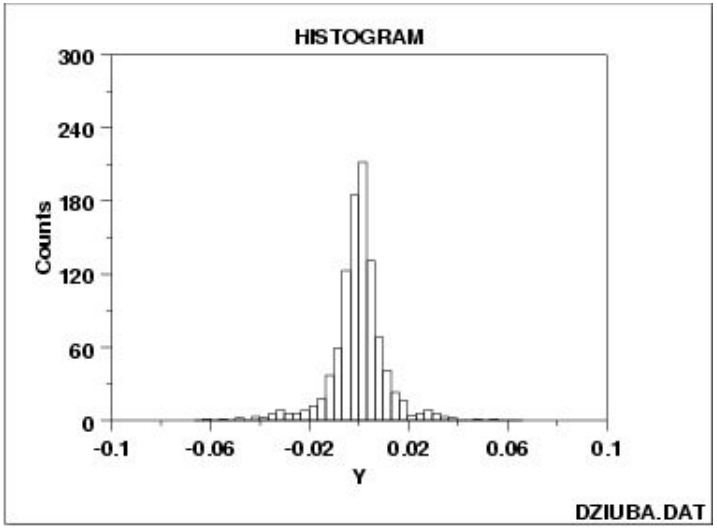


Fig 3: Example of Long-Tailed Histogram

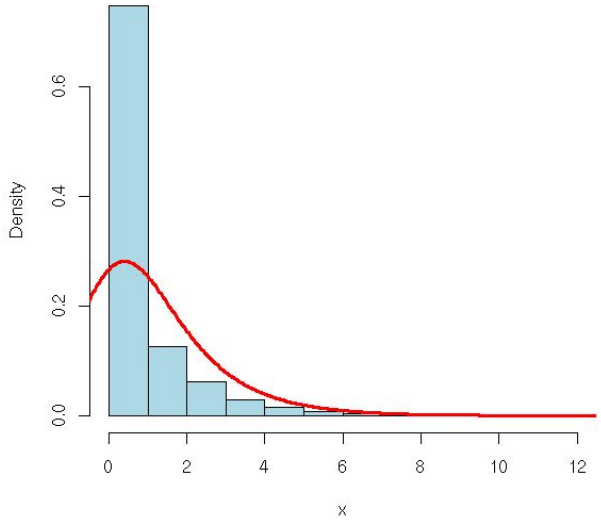
Types of tailed histograms :

- **Sample R code of Long Tailed Histogram :**

  - R creates histogram using hist() function.The basic syntax for creating a histogram using R is : hist(v,main,xlab,xlim,ylim,breaks,col,border)

Example 1 :

```
N <- 10000
x <- rgeom(N, .5)
hist(x,
    xlim=c(min(x),max(x)), probability=T,
nclass=max(x)-min(x)+1,
    col='lightblue',
    main='Geometric distribution, p=.5')
lines(density(x,bw=1), col='red', lwd=3)
```



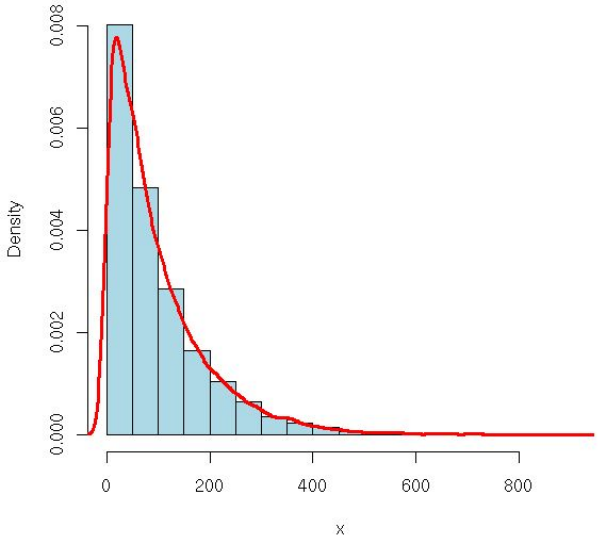
Types of tailed histograms :

- **Sample R code of Long Tailed Histogram :**

  - R creates histogram using hist() function.The basic syntax for creating a histogram using R is : hist(v,main,xlab,xlim,ylim,breaks,col,border)

Example 2 :

```
N <- 10000
x <- rgeom(N, .01)
hist(x,
     xlim=c(min(x),max(x)), probability=T, nclass=20,
     col='lightblue',
     main='Geometric distribution, p=.01')
lines(density(x), col='red', lwd=3)
```



Types of tailed histograms :

- **Skewed-Tailed Histogram :**
  - For skewed distributions, it is quite common to have one tail of the distribution considerably longer or drawn out relative to the other tail.
  - A **"skewed right"** distribution is one in which the tail is on the right side.
  - A **"skewed left"** distribution is one in which the tail is on the left side.

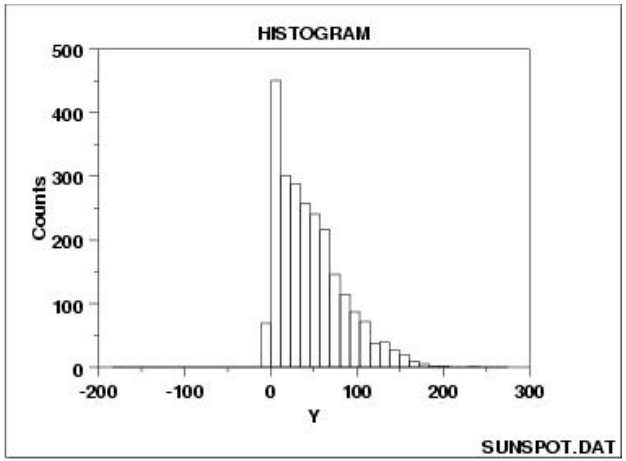


Fig 4: Example of Skewed Right Histogram

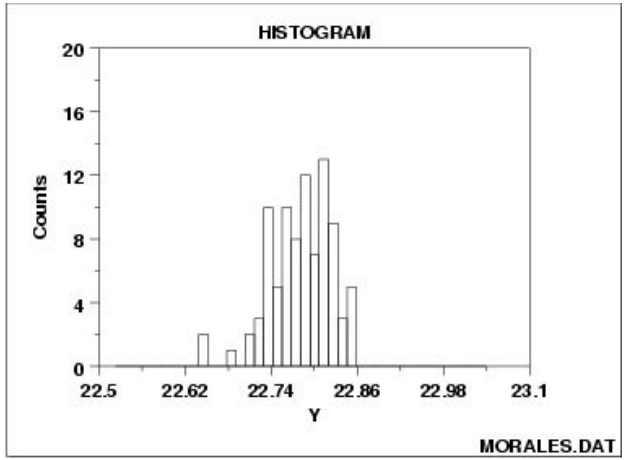


Fig 5: Example of Skewed Left Histogram

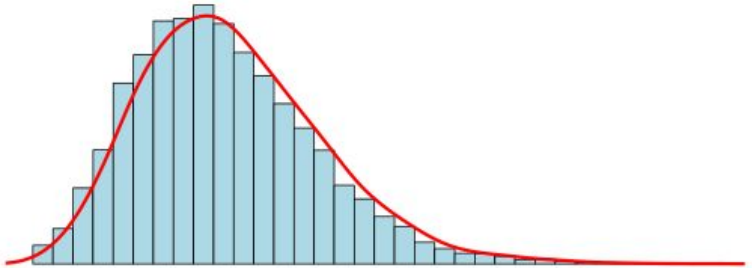
Types of tailed histograms :

- **Sample R code of Skewed Tailed Histogram :**

  - R creates histogram using hist() function.The basic syntax for creating a histogram using R is : hist(v,main,xlab,xlim,ylim,breaks,col,border)

Example 1 :

```
N <- 10000
x <- rnbinom(N, 10, .5)
hist(x,
xlim=c(min(x),max(x)), probability=T,
nclass=max(x)-min(x)+1,
  col='lightblue', xlab=' ', ylab=' ',
axes=F,main='Positive Skewed')
lines(density(x,bw=1), col='red', lwd=3)
```



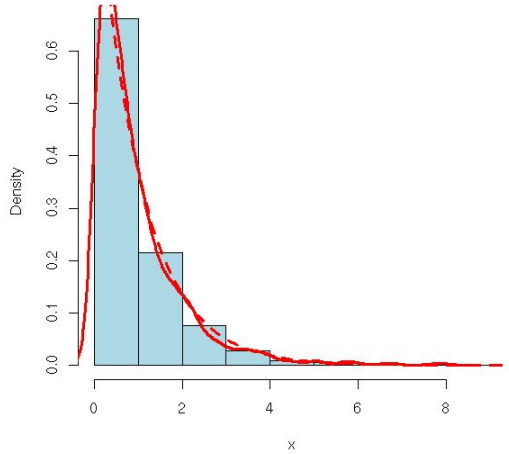
Types of tailed histograms :

- **Sample R code of Skewed Tailed Histogram :**

  - R creates histogram using hist() function.The basic syntax for creating a histogram using R is : hist(v,main,xlab,xlim,ylim,breaks,col,border)

Example 2 : Right Skewed

```
 n <- 1000
x <- rexp(n)
hist(x, probability=T,
     col='light blue', main='Exponential
Distribution')
lines(density(x), col='red', lwd=3)
curve(dexp(x), xlim=c(0,10), col='red',
lwd=3, lty=2,
      add=T)
```



Q.2)

Sample R program to plot multiple time series using ggplot2 data visualization package.

Example 1 :

The ggplot2 package typically takes long data as input. For that reason, we first have to use the reshape2 package to convert our data frame from wide to long format.

We first need to install and load the reshape2 package, if we want to use the functions that are included in the add-on package:

```
install.packages("reshape2")
library("reshape2")
```

Now, we can reshape our data from wide to long format using the melt function:

```
data_long <- melt(data, id.vars = "year")
head(data_long)
```

Now,we need to install and load the ggplot2 package to draw our time series plot using ggplot2:

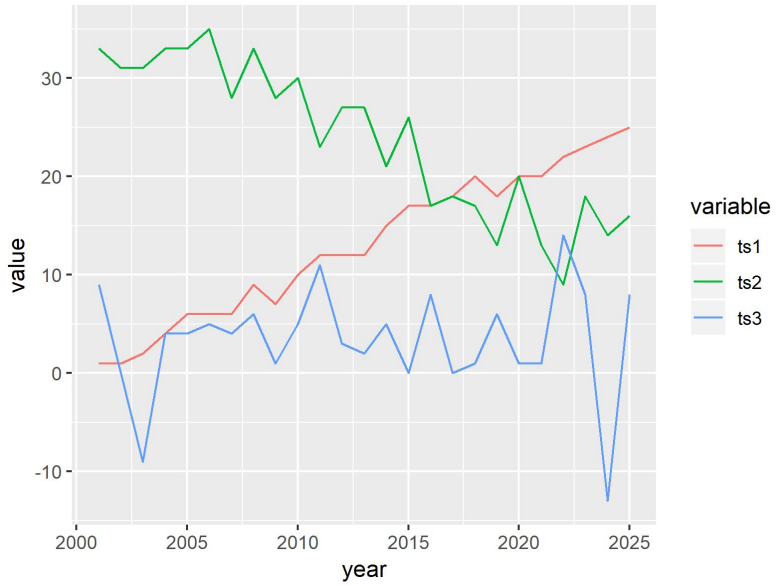
```
install.packages("ggplot2")
library("ggplot2")
```

Q.2)

Sample R program to plot multiple time series using ggplot2 data visualization package.

Finally, we can apply the ggplot and geom_line commands to draw multiple time series to a plot:

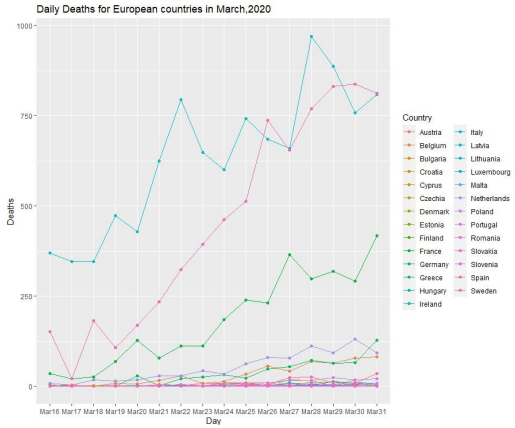
```
ggplot(data_long,
       aes(x = year,
           y = value,
           col = variable)) +
  geom_line()
```



Q.2)

Sample R program to plot multiple time series using ggplot2 data visualization package.

Example 2 : (Daily Death count using EU Covid death dataset)




```
library(ggplot2)
library(reshape2)
library(dplyr)

covid1 =(read.csv(file="EUCOVIDdeaths.csv",header=TRUE)[,-c(2)])

head(covid1)

covid_deaths <- melt(covid1,id.vars=c("Country"),value.name="value",variable.name="Day")

head(covid_deaths)

covid_plot <- ggplot(data=covid_deaths, aes(x=Day, y=value, group = Country,colour = Country))
+ geom_line() +labs(y= "Deaths", x = "Day")
covid_plot + ggtitle("Daily Deaths for European countries in March,2020")+geom_point()

covid_plot
```

Reference Link

Instructions :

Submission Date : 30/03/2023

For all the questions write clear definitions , explain with its characteristics with proper diagram , along with proper R code .(Code & Results Screenshot)