Exercise 1

1. Calculate for the following files `dat1.dat`, `dat2.dat`, `dat3.dat`:
    - Mean and Standard Deviation
    - Median
    - Histogram

      Plot histograms (density, not frecuency) and curves for the distributions
2. Which distributions are they? Why

> For installing modeest: R documentation `conda install -c conda-forge r-modeest`

Statistics: Uncertainties and Errors

Author: Laura V. Trujillo T. lvtrujillot@unal.edu.co

Introduction

The interest of the following *exercise* is to give an overview of the basic concepts in statistics such as **mean**, **median** and **standard deviation** of a bunch of data in order to analyze and understand its behavior quantitatively and qualitatively (histograms, distribution curves and so on)

1. **Mean** : it is used to derive the central tendency of the data. In other words, it is the average of the data set.
2. **Median**: it is the middle of the data set.
3. **Standard Deviation** it is a measure of how spread out the data set is around the mean.

```
d1 <- read.table("dat1.dat",header=FALSE)
d2 <- read.table("dat2.dat",header=FALSE)
d3 <- read.table("dat3.dat",header=FALSE)
```

Mean, Median and Standard Deviation

`dat1.dat`

```
cat("Data1 \n The standard deviation is: ", sd(d1$V1),
    "\n The Median is: ", summary(d1$V1)[3],
    "\n The Mean is: ", summary(d1$V1)[4],
   "\n The first and third quartiles: ", summary(d1$V1)[2],",", summary(d1$V1)[5])
```

```
Data1
```

```
  The standard deviation is:  1.414748
  The Median is:  1.975919
  The Mean is:  1.9863
  The first and third quartiles:  0.7467177 , 3.225604
```

`dat2.dat`

```
cat("Data2 \n The standard deviation is: ",sd(d2$V1),
    "\n The Median is: ", summary(d2$V1)[3],
    "\n The Mean is: ", summary(d2$V1)[4],
   "\n The first and third quartiles: ", summary(d2$V1)[2],",", summary(d2$V1)[5])
```

```
 Data2
  The standard deviation is:  1.414748
  The Median is:  2
  The Mean is:  1.9863
  The first and third quartiles:  1 , 3
```

`dat3.dat`

```
cat("Data3 \n The standard deviation is: ",sd(d3$V1),
    "\n The Median is: ", summary(d3$V1)[3],
    "\n The Mean is: ", summary(d3$V1)[4],
   "\n The first and third quartiles: ", summary(d3$V1)[2],",", summary(d3$V1)[5])
```

```
 Data3
  The standard deviation is:  1.414748
  The Median is:  2.001954
  The Mean is:  1.9863
  The first and third quartiles:  1.023417 , 2.936421
```

The following table summarizes the results obtained:

Data	Mean	Median	Standard Deviation	First Quartile	Third Quartile
`dat1`	1.9863	1.9759	1.4147	0.746	3.22
`dat2`	1.9863	2.000	1.4147	1.0	3.0
`dat3`	1.9863	2.001	1.4147	1.023	2.93

Plots

The main interest of this section is to see qualitatively the behavior of the data given so as to describe it with a known distribution curve (i.e. Poisson, Gaussian and so on)

```
#Distribution curve for dat1
x1 <- seq(-0.5, 4.4, length.out=1000 )
y1 <- dunif(x1, min=-0.45, max=4.39)

#Distribution curve for dat2
x2 <- seq(0, 9, by=1)
y2 <- dbinom(x2, size = 10, prob = 0.18)
# Distribution curve for dat3
x3 <- seq(-4, 8, length.out=1000)
y3 <- dnorm(x3, mean=mean(d3$V1), sd=sd(d3$V1))
```

```
h <- hist(d2$V1, breaks=100, plot=FALSE)
h$counts=h$counts/sum(h$counts)
```

```
par(mfrow=c(2,2))
hist(d1$V1, freq=FALSE, main="Data 1",ylim=c(0, 0.25), col="cyan",breaks = 10,
xlab="V1")
lines(x1, y1, col="red", lw=2)
plot(h, ylim=c(0, 0.30), col="cyan",main="Data 2",ylab="Density",  xlab="V1")
lines(x2,y2, col="red", type="p")
hist(d3$V1, freq=FALSE,main="Data 3", ylim=c(0, 0.35),col="cyan", xlab="V1")
lines(x3, y3, col="red", lw=2)
```

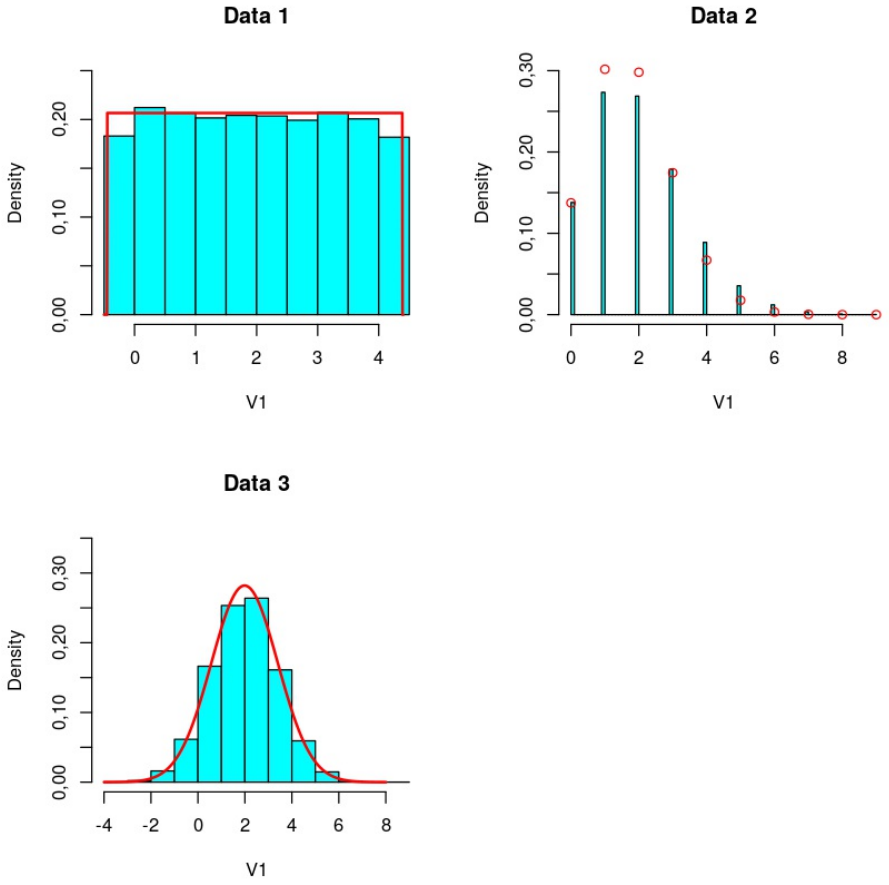

Figure 1. Plots for Data 1, Data 2, Data 3

For `Data1` is evident that its behavior is alike the **uniform distribution** since almost all the outcomes are equally likely.

For `Data2` is not so evident. Nevertheless, there are a couple of observations that it can be derived from the plot:

- It is discrete, meaning that it would be reasonable to choose the binomial or geometric distribution.
- It looks like it has two possible outcomes, for instance when you flip a coin you either get tails or heads. In that sense, the most appropriate distribution curve to choose is the **binomial**.

For `Data3` is observed that it is continuous, symmetric and the outcomes are concentrated in the middle of the range. Those characteristics matches the ones for the **normal or gaussian distribution**

Conclusions

In closing, despite the resemblance between the mean, median and standard deviation of the set of data given (`dat1.dat` , `dat2.dat` , `dat3.dat`) (see table attached) the shape of the distributions for each sample were quite differently and therefore it can be affirmed that it is necessary to look out the distribution of the data.

References

```
citation()
```

```
To cite R in publications use:

  R Core Team (2019). R: A language and environment for statistical
  computing. R Foundation for Statistical Computing, Vienna, Austria.
  URL https://www.R-project.org/.

A BibTeX entry for LaTeX users is

  @Manual{,
    title = {R: A Language and Environment for Statistical Computing},
    author = {{R Core Team}},
    organization = {R Foundation for Statistical Computing},
    address = {Vienna, Austria},
    year = {2019},
    url = {https://www.R-project.org/},
  }

We have invested a lot of time and effort in creating R, please cite it
when using it for data analysis. See also ‘citation("pkgname")’ for
citing R packages.
```