Data Visualisation 1

Code ▾

Kapil Paranjape

08/03/2018

Coarse Statistics

We study the process of analysing data based on visualisation.

To begin with we start with the same data set that we had earlier.

```
mydata <- read.csv("dat.csv")
dim(mydata)
```

```
[1] 180   2
```

```
colnames(mydata)
```

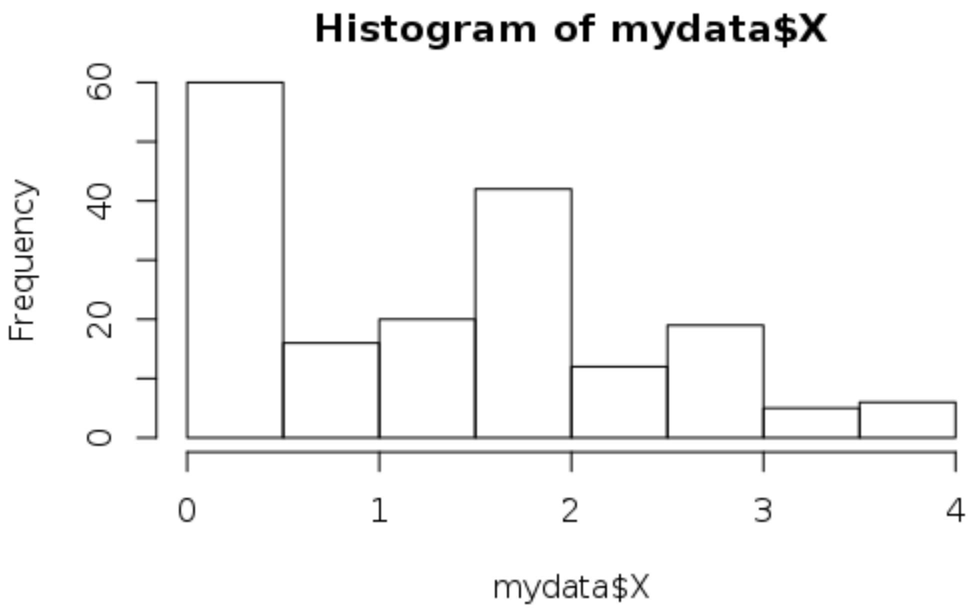
```
[1] "X" "Y"
```

As seen earlier this has 180 entries in 2 columns named `X` and `Y`. We can find the summary information of these columns.

```
summary(mydata$X)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   0.000   1.500   1.472   2.000   4.000 
```

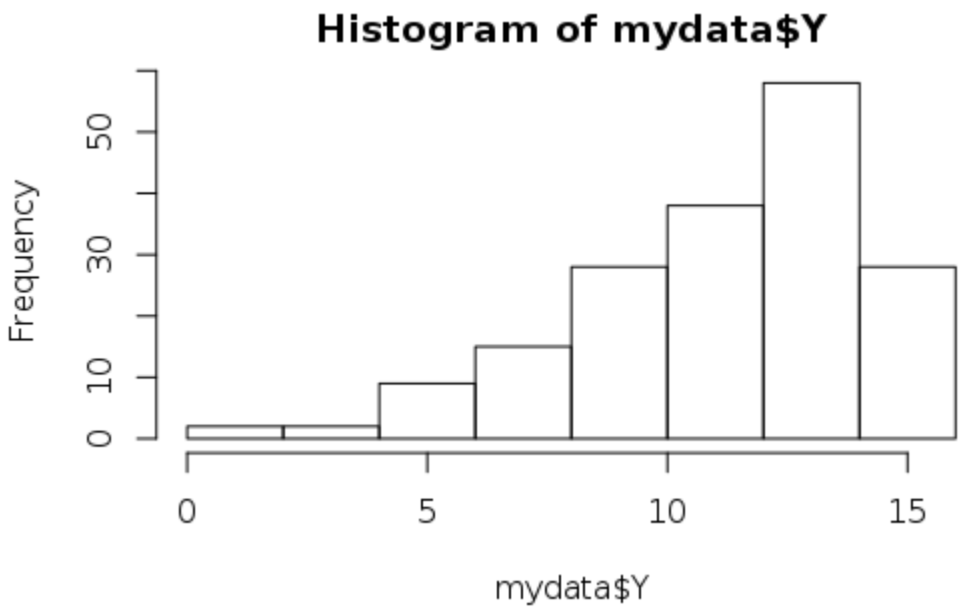
```
hist(mydata$X)
```




```
summary(mydata$Y)
```

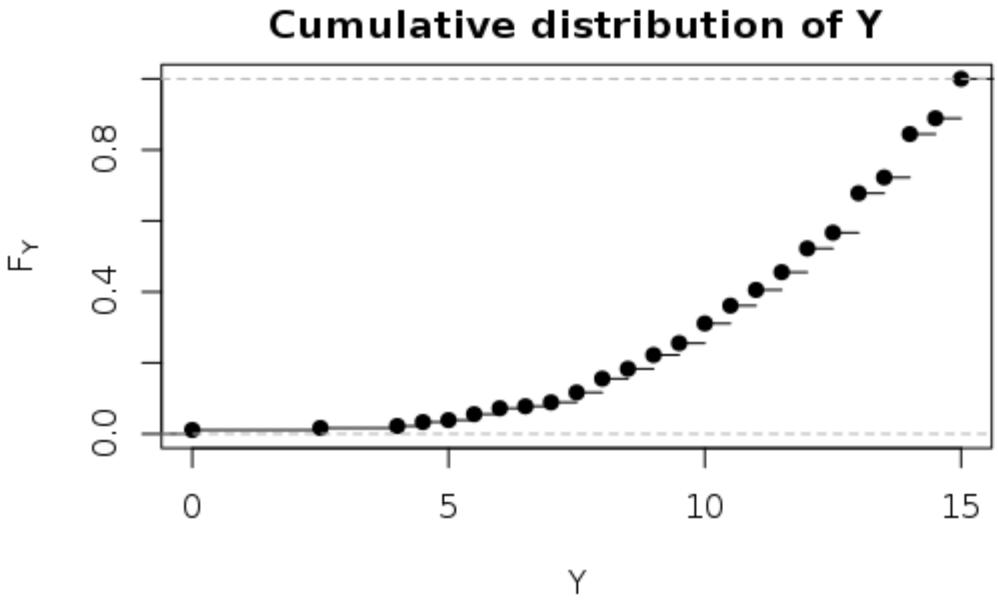
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    9.50   12.00   11.41   14.00   15.00
```

```
hist(mydata$Y)
```



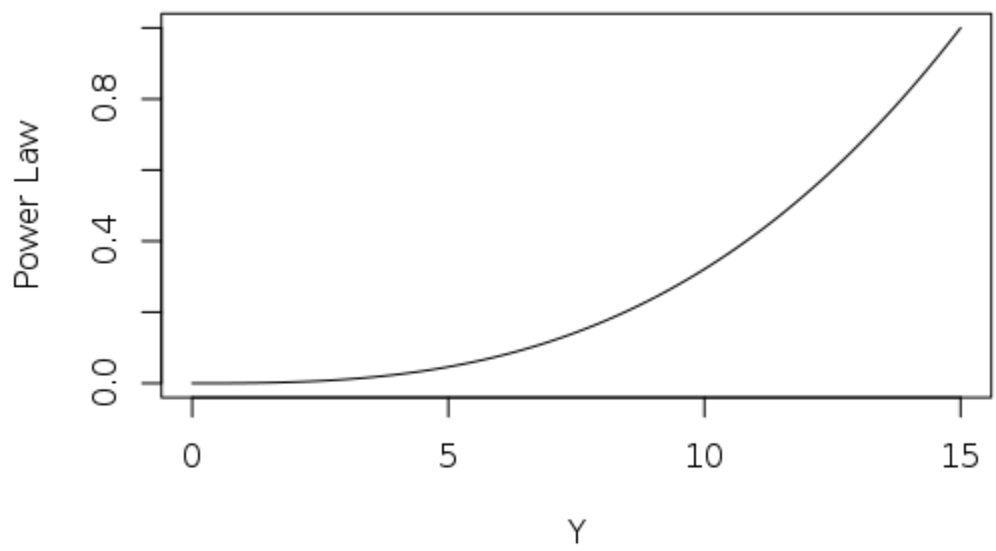
Next we look at the (empirical) cumulative distribution of the distribution of `Y`.

```
plot.ecdf(mydata$Y, xlab=expression(Y), ylab=expression(F[Y]), xlim=c(0,15),
main="Cumulative distribution of Y")
```



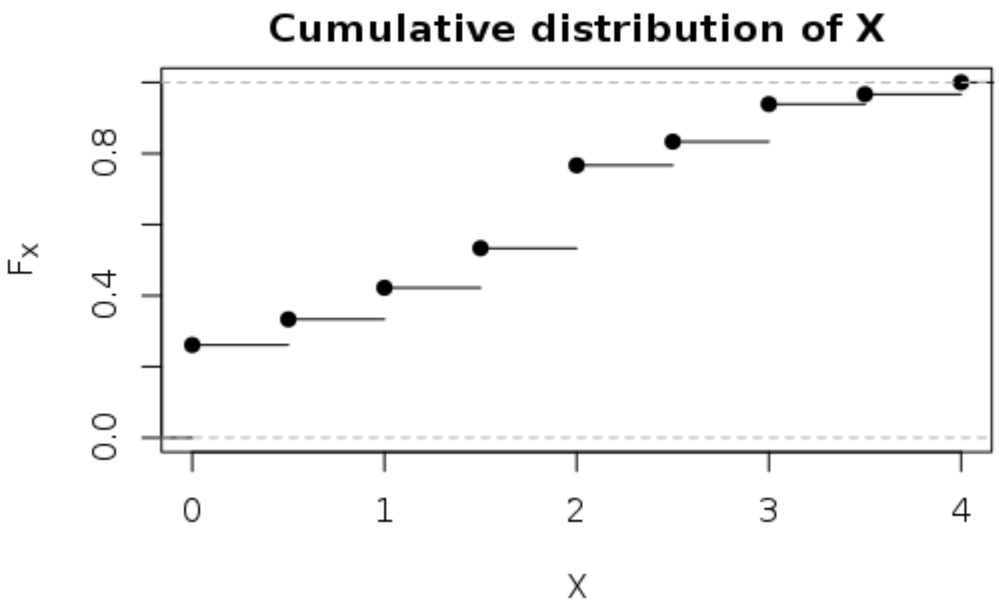
We can “eyeball” this curve and find that it superficially looks similar to the following curve.

```
curve((x/15)**2.8,0,15,ylab="Power Law",xlab="Y")
```



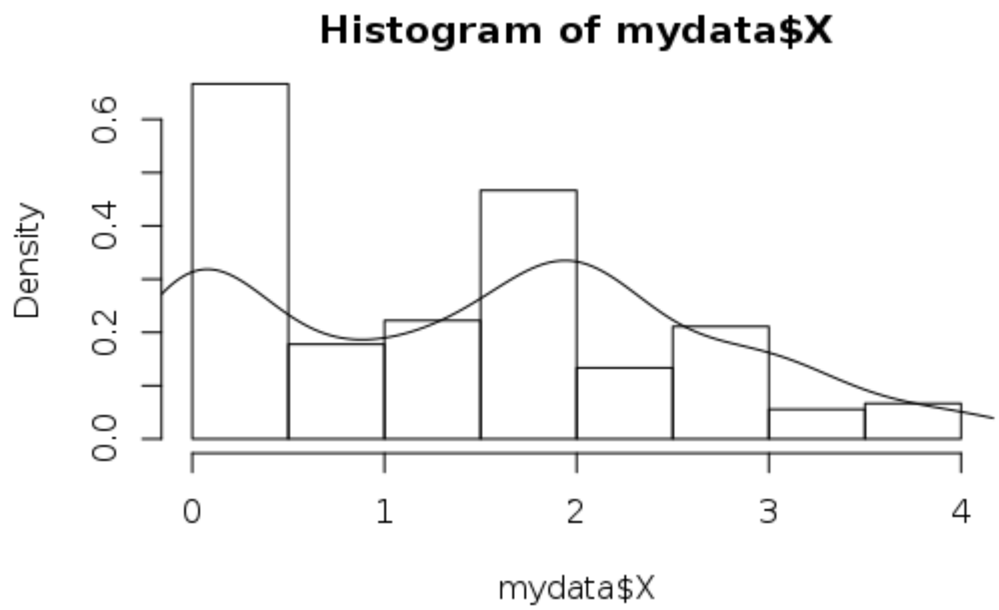
We can try to do the same for `X` .

```
plot.ecdf(mydata$X, xlab=expression(X), ylab=expression(F[X]), xlim=c(0,4),
main="Cumulative distribution of X")
```



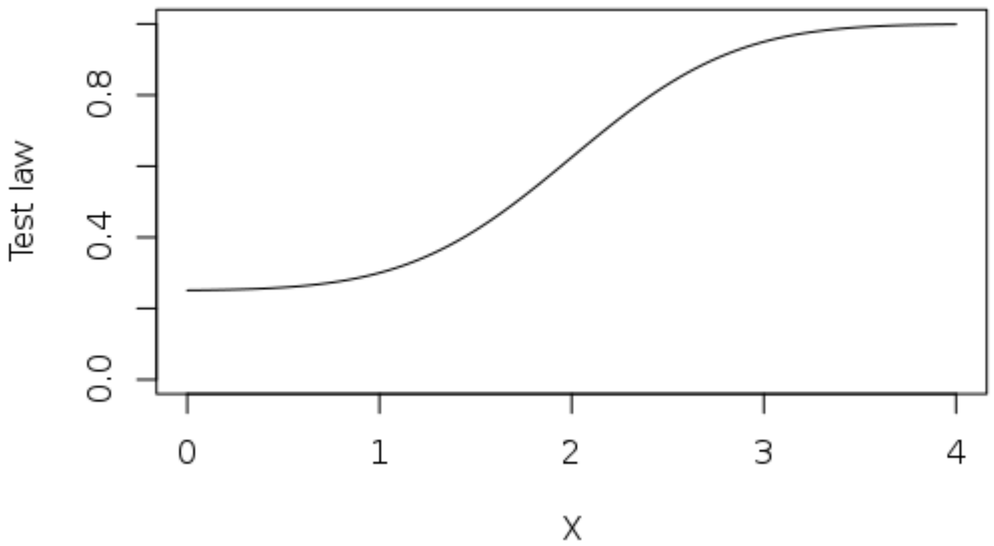
This is more difficult to see as a “shape”. So sometimes it is better to ask R what a density plot of this would look like *if* it were points from a continuous distribution.

```
hist(mydata$X,prob=T)
lines(density(mydata$X))
```



It looks like the superposition of two distributions. One centred at 2 with standard deviation about 2/3 and other a sharper “delta” distribution centred at 0. The relative weight is given by the frequency of 0 which is about 0.25.

```
curve(0.25+(1-0.25)*pnorm(x,2,2/3),0,4,ylab="Test law",ylim=c(0.0,1.0),xlab=
"X")
```



Visually, this looks reasonable. However, we can look for further statistical clues.

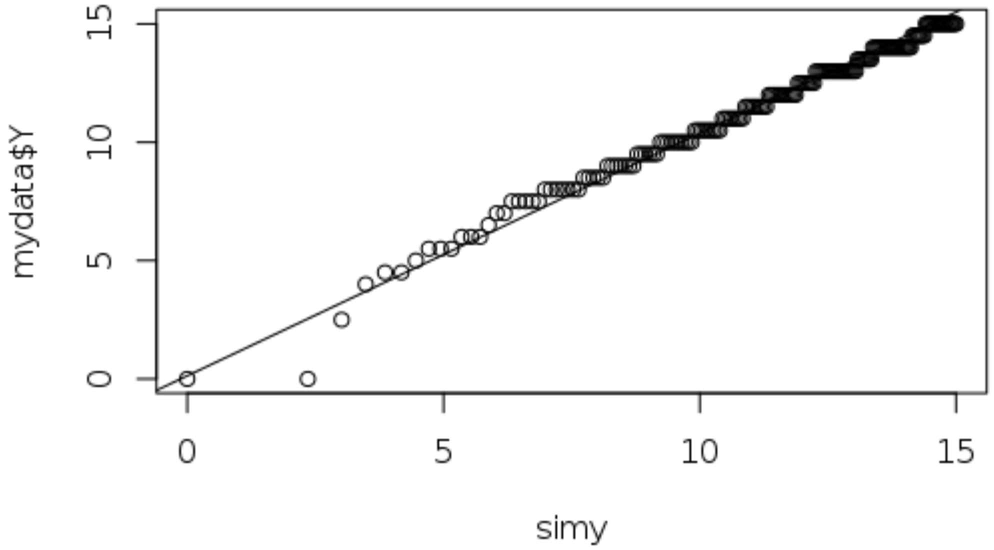
Deeper Analysis

There are two aspects that we need to work on. One is to find suitable “best” choice of parameters that will fit the distribution; for example, in the first case we can ask whether 2.8 is the optimal choice while in the second case we can wonder whether 0.25, 2 and 2/3 are the correct choice of constants. Next, having chosen these constants we need to have a way to test the hypothesis that the data fits the distribution with these parameters.

Before we do that, here is another way to visually compare the distribution, the quantile-quantile (q-q) plot.

To do this we need to create the quantile version of our test distribution. This is the inverse of the cumulative distribution function.

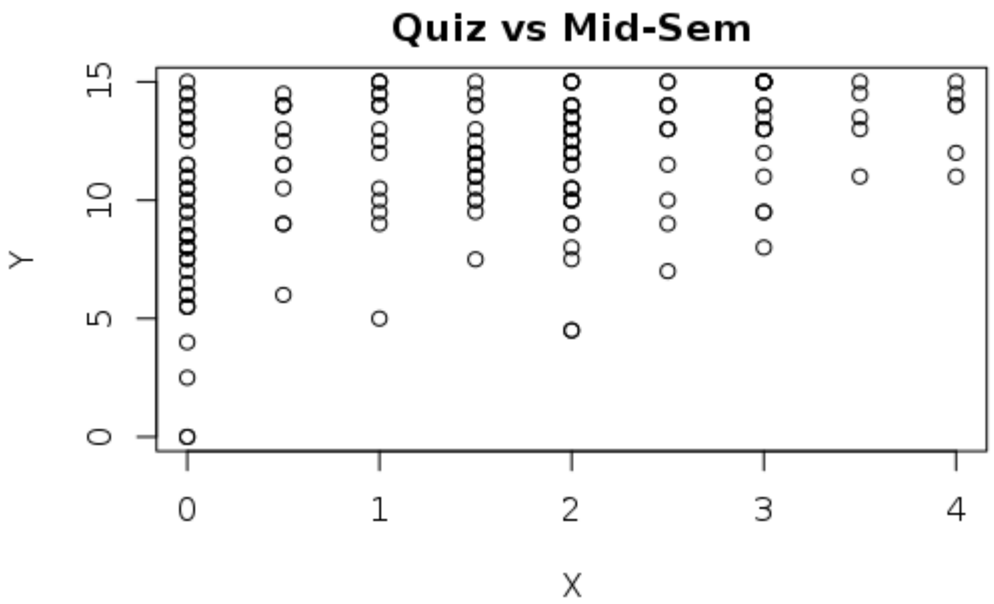
```
qy <- function(p) 15*(p**(1/2.8))
simy <- qy(seq(0,1,length.out=180))
qqplot(simy,mydata$Y)
qqline(mydata$Y,distribution = qy)
```



Upto a slight problem at the lower end it appears to fit quite neatly!

Finally, we may ask whether there is any relation between the `X` and `Y` variables.

```
plot(mydata,main="Quiz vs Mid-Sem")
```



We see that people who did well in the quiz generally did well in the examination! However, *some* people who scored 0 in the quiz (perhaps because they missed it!) also did well in the mid-sem.