DATA DRIVEN DECISION

BUDDHANANDA BANERJEE

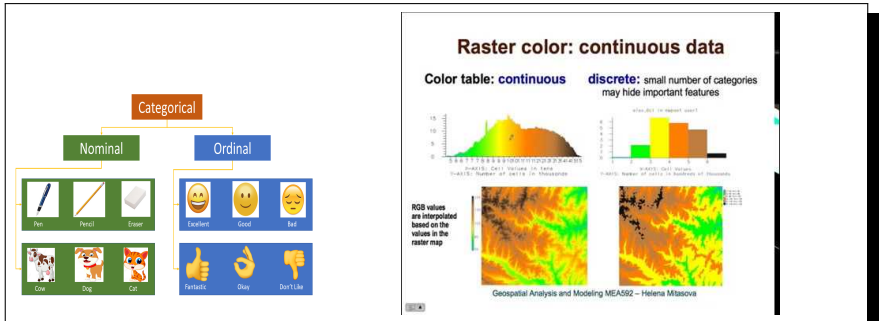
1. Data

Qn: Why do we collect data or samples ?
Ans: To know about the entire population.
Qn: What do we want to know about a population ?
Ans: One or more variable(s) / attribute(s) of interest.
Qn: What is an unique characterization of any feature(s)?
Ans: Its (joint) probability distribution. **[WHY???]**

▷ **Types of Data**
- **Categorical**
  - ∗ **Nominal** : Names only, without ordering
  - ∗ **Ordinal** : Names and ordering
- **Numerical**
  - ∗ **Countable** : Values from integers or its function
  - ∗ **Uncountable :** Values from an interval

Just as we always encounter numbers accompanied by units rather than standalone figures, similarly, data in the real world represents the tangible expression of the abstract concept of random variables.

Date: Last updated May 18, 2024.



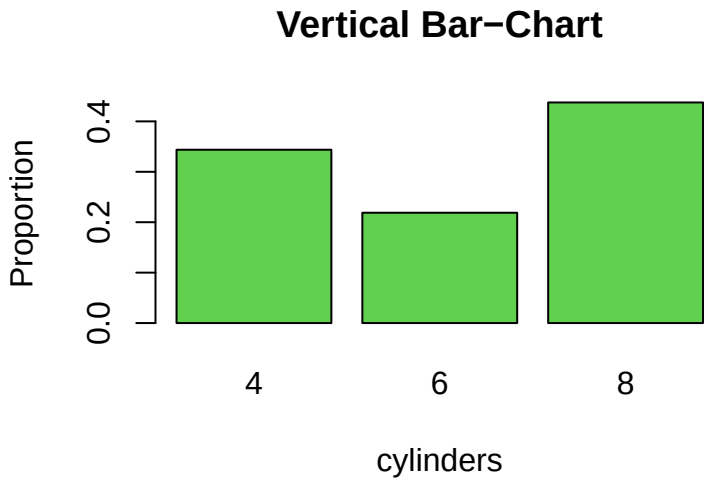
Example of data

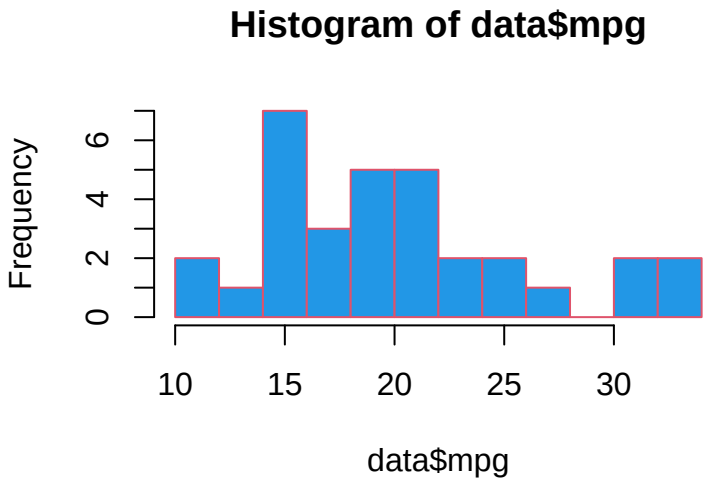
These data are considered to be the realized values of a mathematical object (function) which is known as **random variable.** A **random variable** is usually denoted as X and its realized value is denoted as $x \in \mathbb{R}$.

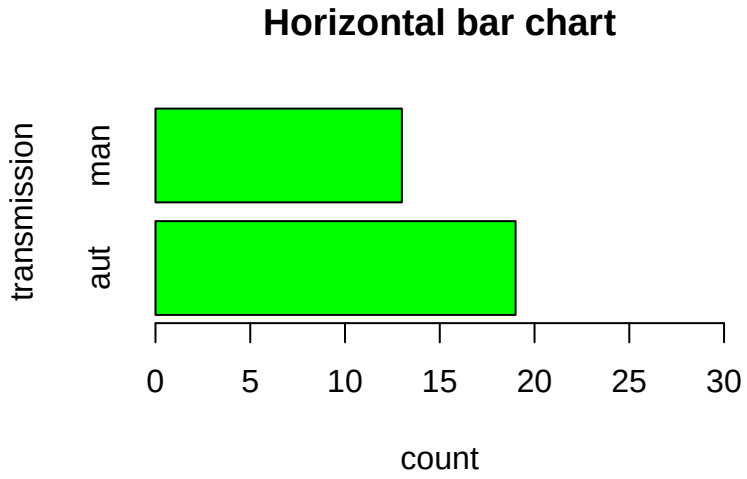
2. Diagrammatic summary of data

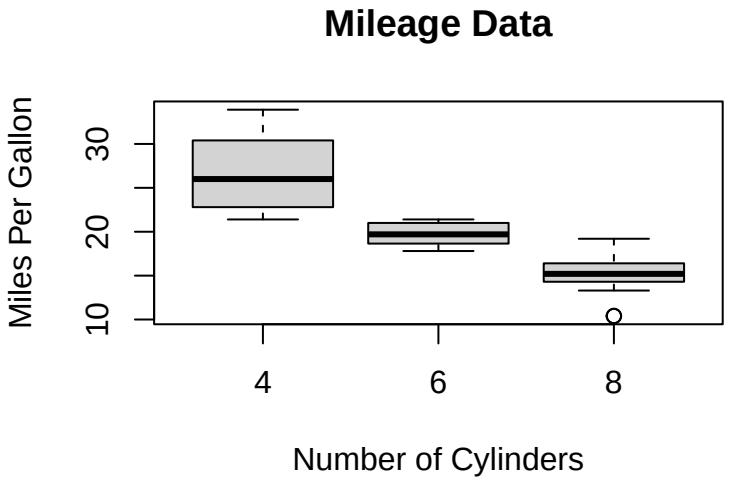
Description: The Motor Trend Car Road Tests (mtcars) data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models). Henderson and Velleman (1981) comment in a footnote to Table 1: 'Hocking [original transcriber]'s noncrucial coding of the Mazda's rotary engine as a straight six-cylinder engine and the Porsche's flat engine as a V engine, as well as the inclusion of the diesel Mercedes 240D, have been retained to enable direct comparisons to be made with previous analyses.'
Source: Henderson and Velleman (1981), Building multiple regression models interactively. Biometrics, 37, 391–411.

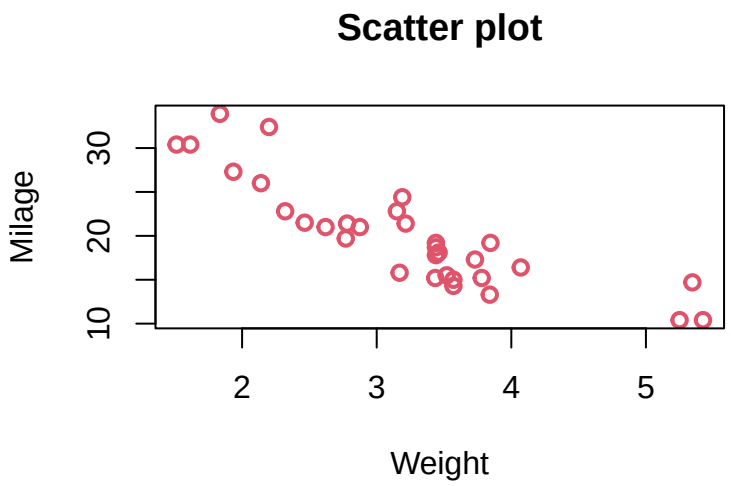
Output of R Code 1

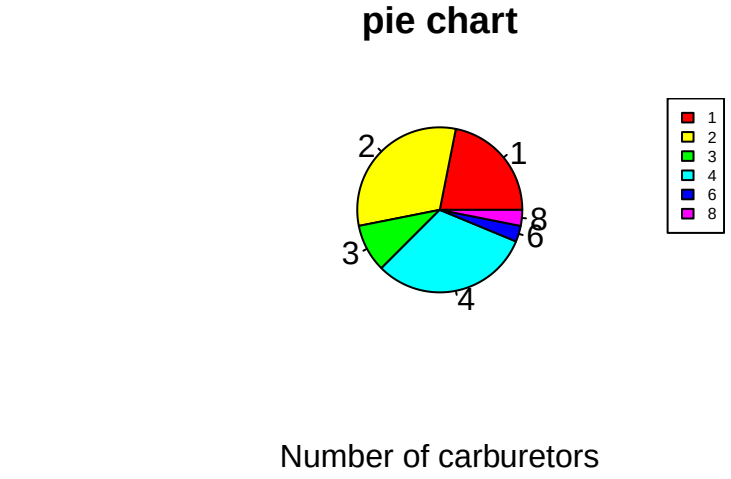












3. Numerical Summary of data

Central Tendency: Central tendency refers to the typical or central value around which data points tend to cluster. It gives an idea of the "average" or "typical" value in a dataset. The most common measures of central tendency are: mean, median and mode.

Mean:: The arithmetic average of all the values in a dataset. It is calculated by adding up all the values and then dividing by the number of values.

Median:: The middle value in a dataset when it is arranged in ascending or descending order. If there is an even number of values, the median is the average of the two middle values.

Mode:: The value that appears most frequently in a dataset.

Dispersion: Dispersion measures how spread out the values in a dataset are from the central value (i.e., the measure of central tendency). It provides information about the variability or spread of the data points. Common measures of dispersion include

Range:: The difference between the maximum and minimum values in a dataset.

Variance:: A measure of how much the values in a dataset vary from the mean. It is calculated by taking the average of the squared differences between each value and the mean.

SD:: Standard DeviationThe square root of the variance.

IQR:: Interquartile Range is he range between the first quartile (25th percentile) and the third quartile (75th percentile) of the dataset. It represents the spread of the middle 50% of the data.

Measure	Sample analog
Central Tendency	**Average**
Mean	$m_1' = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Median	Middle most point of the sorted data
Mode	Data with highest frequency
Dispersion	**Spread**
Variance	$m_2 = s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$
sd	$\sqrt{m_2}$
Mean absolute deviation	$g(m) = \frac{1}{n}\sum_{i=1}^{n} \lvert x_i - m \rvert$, m = median
Rage	max − min
Inter quartile range	$Q_3 - Q_1$
Skewness	**Tilt**
Skewness (Moment)	$g_1 = \frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right]^{3/2}}$
Skewness (Quartile)	$(Q_1 + Q_3 - 2Q_2)/(Q_3 - Q_1)$
Picked-ness	**Concentration**
Kurtosis	$g_2 = \frac{m_4}{m_2^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right]^2}$

Qn: If we know all these measures of population/sample do we know all about the population ?????

Ans: NO. !!!!!!!!

Output of R Code 2

```
## Summary of Miles/(US) gallon
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   15.43   19.20   20.09   22.80   33.90
## Var= 36.3241
## Skewness(m)= 0.6404399
## Kurtosis= 2.799467
## MAD= 4.634375
## Range= 23.5
## Inter Quartile Range= 7.375
## Skewness(Q)= -0.02372881
## [1] "Mode_cyl: 8"
## Proportion table  of Number of cylinders
##
##       4       6       8
## 0.34375 0.21875 0.43750
```

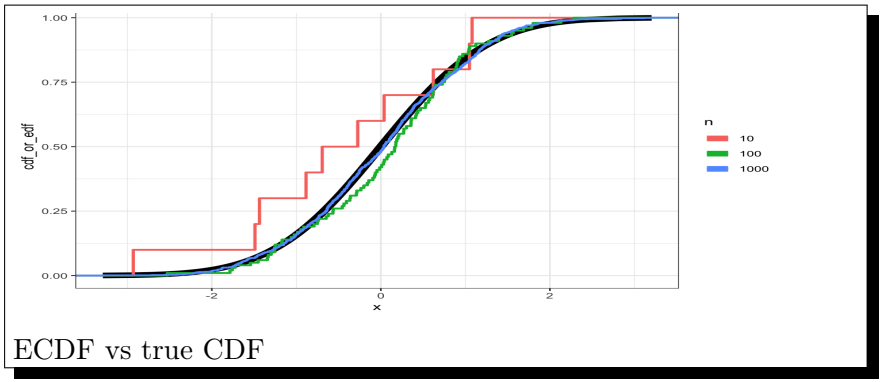
4. Probabilistic Characterization of data

In statistics, an empirical distribution function (commonly also called an empirical cumulative distribution function, eCDF) is the distribution function associated with the empirical measure of a sample. This cumulative distribution function is a step function that jumps up by $1/n$ at each of the n data points. Its value at any specified value of the measured variable is the fraction of observations of the measured variable that are less than or equal to the specified value.

Definition 1. Let $(X_1 \ldots, X_n)$ be random samples from a population distribution of which has the common cumulative distribution function $F(t)$. Then the empirical cumulative distribution function (ECDF) is defined as

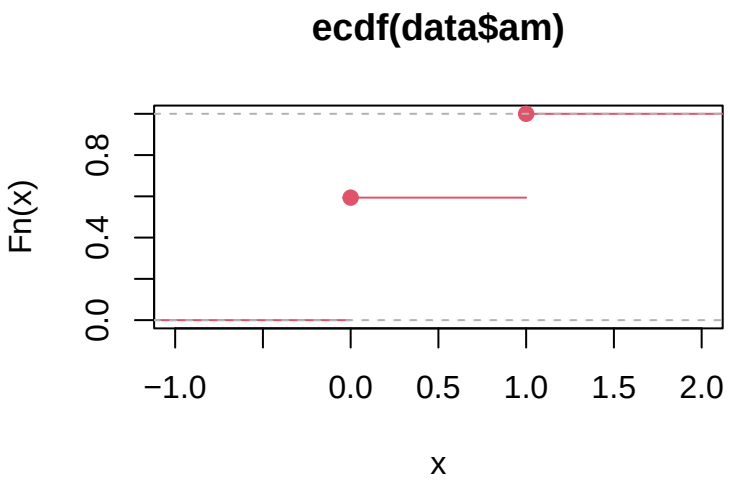
$$\widehat{F}_n(t) = \frac{\text{number of elements in the sample} \leq t}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_i \leq t},$$

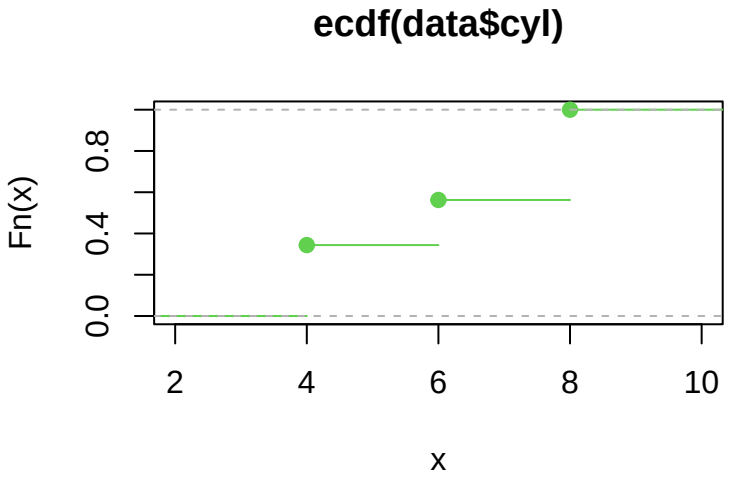
where $\mathbf{1}_A$ is the indicator of event A.



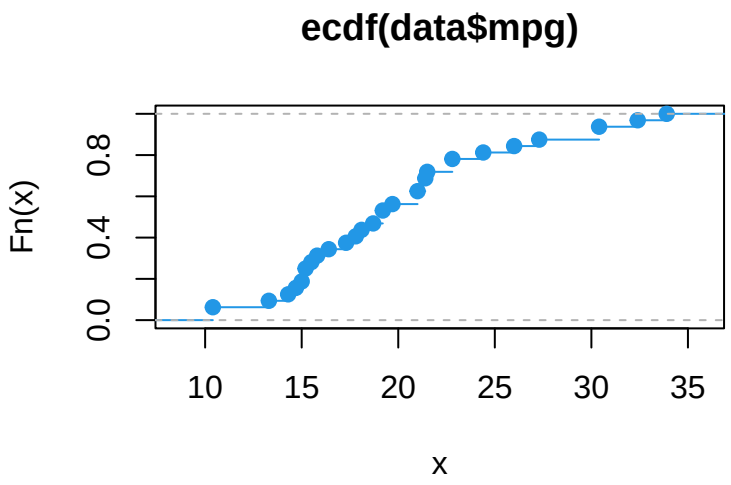
ECDF vs true CDF

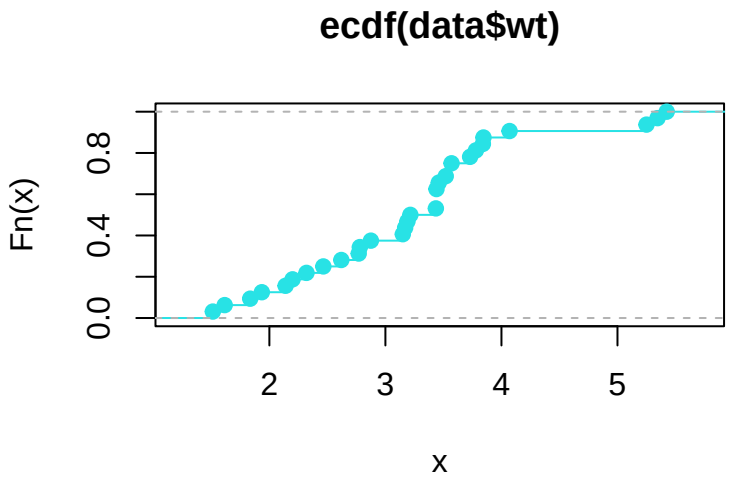
Output of R Code 3






```
## Proportion table  of Number of cylinders
##
##       4       6       8
## 0.34375 0.21875 0.43750
```





Qn: Is this information transferable in a summarized form ?

Ans: Not always.

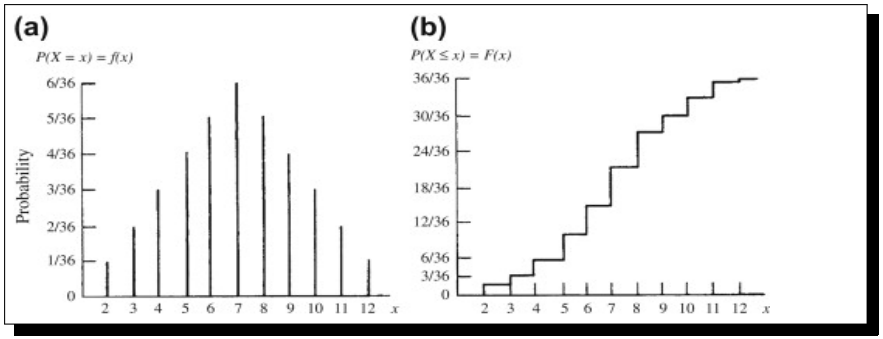
In probability theory and statistics, the cumulative distribution function (CDF) of a real-valued random variable X, or just distribution function of X, evaluated at x, is the probability that X will take a value less than or equal to x.

Definition 2. The **cumulative distribution function (c.d.f.)** $F : \mathbb{R} \to [0, 1]$, which uniquely characterizes a random variable, is define as

$F(x) = P(X \leq x) = P(X \in (-\infty, x])$ for all $x \in \mathbb{R}$.

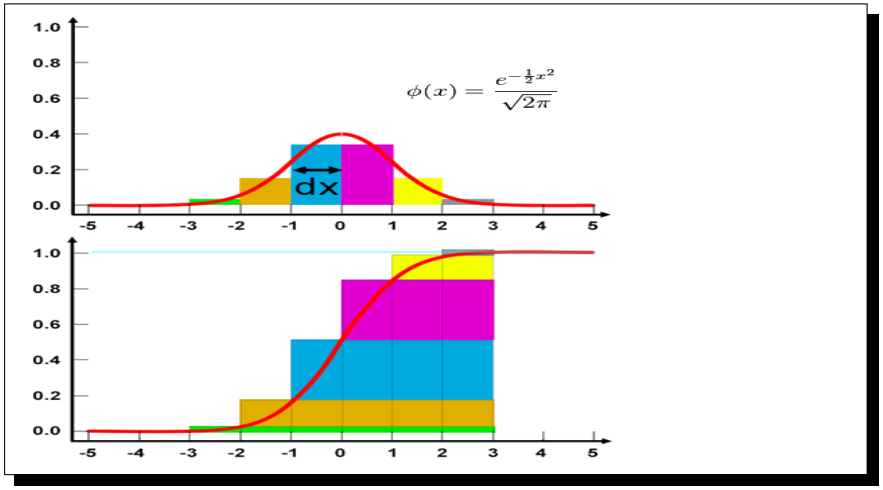
The CDF is also can equivalently represented through another non-negative function $f(x)$ called as probability density function (p.d.f.). The p.d.f. is a mathematical function that describes the likelihood of a random variable taking on a particular value within a given range. It represents the relative likelihood of observing different outcomes of the random variable. The integral of the PDF over a certain interval gives the probability that the random variable falls within that interval.

In broad terms, the cumulative distribution function (CDF) and probability density function (PDF) serve to uniquely define the distribution of a random variable. However, they are dependent on certain unknown parameter(s) that need to be estimated based on a given dataset. If the underlying assumption regarding the CDF and PDF holds true, the estimation of these parameter(s) becomes more precise, thereby rendering the information effectively reusable.



Definition 3. The **probability density function (p.d.f.)** of a random variable X with c.d.f $F_X(\cdot)$ is a nonnegative function $f(x)$ such that,

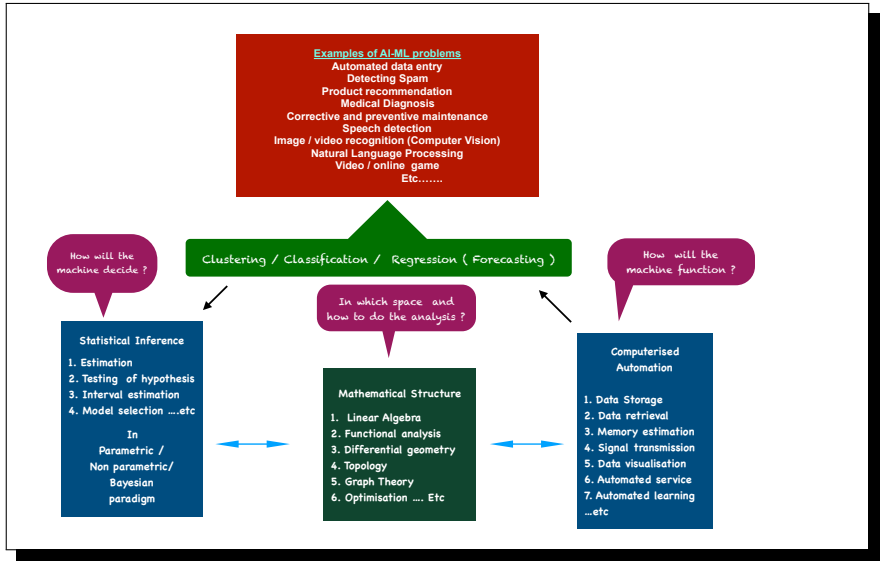
$$F(t) = \begin{cases} \sum_{x \le t} f(x) \cdot 1, & \text{for discrete } X, \\ \int_{x \le t} f(x)dx, & \text{for continuous } X. \end{cases}$$



5. Ask some questions to yourself

- ▷ When do you consider the rating of a product in amazon in Amazon may be valid and useful ? Why so ?
- ▷ Why do we represent the CAT / NET rank of a candidate in percentile?
- ▷ How does the chat prompt work at your smart phone?
- ▷ How does the Electronic Health Records (EHRs) and medical imaging data are used to develop data-driven solutions for disease diagnosis?
- ▷ Why do often you get calls/ sms/ mail for pre approved lone or credit card even though you have not applied for it ?

Data driven problem



R Code 1

```
#====================================
#Data Plotting Motor Trend Car Road Tests
#====================================
#write.csv(x = mtcars,file = "mtc.csv",row.names = F)
#data<- read.csv("mtc.csv")
data<-mtcars
# vertical bar chart
tab_cyl<-table(data$cyl)/nrow(data)
barplot(tab_cyl, xlab = "cylinders", ylab = "Proportion", main ="Vertical Bar-Chart" ,col=3)

# horizontal  bar chart
A<-table(data$am)
B<-c('aut', 'man')      # categorical data
barplot(A, names.arg = B, horiz = T,
        xlab ="count", ylab ="transmission", col ="green", xlim=c(0,30),
        main ="Horizontal bar chart")

#=======================
#histogram
#=======================
hg<-hist(x = data$mpg,probability = F,
         breaks = 10,right = T,col = 4,border = 2)
#=======================
# Box-plot
#=======================
boxplot(mpg ~ cyl, data , xlab = "Number of Cylinders",
```

```
        ylab = "Miles Per Gallon", main = "Mileage Data")
bp2<-boxplot(mpg ~ cyl, data , xlab = "Number of Cylinders",
        ylab = "Miles Per Gallon", main = "Mileage Data")
#print(bp2)
#=======================
# scatter plot
#=======================
plot(x = data$wt,y = data$mpg, type = "p", col=2,lwd=2,
     xlab = "Weight",
     ylab = "Milage",
     main = "Scatter plot"
)
#=======================
#pie-chart
#=======================
# Create data for the graph.
frequ<- c(7, 10,  3, 10,  1,  1 )
Number_of_carburetors<- c(1, 2,  3,  4,  6,  8)
# Plot the chart.
pie(frequ, labels = Number_of_carburetors,
    main = "pie chart", col = rainbow(length(frequ)) ,xlab="Number of carburetors")
legend("topright", c('1', '2',  '3',  '4',  '6',  '8'),
       cex = 0.5, fill = rainbow(length(frequ)))
```

R code 2

```
#====================================
#Data Summary Motor Trend Car Road Tests
```

```r
#===================================
data<-mtcars
# Summary of Miles/(US) gallon
cat("Summary of Miles/(US) gallon \n")
sm_mpg<-summary(data$mpg)
print(sm_mpg)
# Using moments package
cat("Var=",var(data$mpg),"\n")
library(moments)
skewness_m <- skewness(data$mpg)
cat("Skewness(m)=",skewness_m,"\n")
kurtosis_m<-kurtosis(data$mpg)
cat("Kurtosis=",kurtosis_m,"\n")
mad<-mean(abs(data$mpg-median(data$mpg)))
cat("MAD=",mad,"\n")
rg<-diff(range(data$mpg))
cat("Range=",rg,"\n")
Q1<-quantile(data$mpg,p=0.25)
Q2<-quantile(data$mpg,p=0.5)
Q3<-quantile(data$mpg,p=0.75)
cat("Inter Quartile Range=",Q3-Q1,"\n")
cat("Skewness(Q)=",(Q3+Q1-2*Q2)/(Q3-Q1),"\n")
mode_value <- as.numeric(names(sort(-table(data$cyl))[1]))
print(paste("Mode_cyl:", mode_value))

#Proportion table  of Number of cylinders
cat("Proportion table  of Number of cylinders \n")
```

```r
tab_cyl<-table(data$cyl)/nrow(data)
print(tab_cyl)

# plot(ecdf(data$am), col=2)
# plot(ecdf(data$gear), col=3)
# plot(ecdf(data$mpg),col=4)
# plot(ecdf(data$wt),col=5)
```

R Code 3

```r
#====================================
#Data Summary Motor Trend Car Road Tests
#====================================
data<-mtcars

#Proportion table  of Number of cylinders
cat("Proportion table  of Number of cylinders \n")
tab_cyl<-table(data$cyl)/nrow(data)
print(tab_cyl)

plot(ecdf(data$am), col=2)
plot(ecdf(data$cyl), col=3)
plot(ecdf(data$mpg),col=4)
plot(ecdf(data$wt),col=5)
```

Department of Mathematics, IIT Kgaragpur
URL: https://sites.google.com/site/buddhanandastat/
E-mail address: bbanerjee@maths.iitkgp.ac.in