Introduction to Probabilty - Data Science - Statistics

Professor Somesh Kumar
Department of Mathematics
Indian Institute of Technology Kharagpur
Kharagpur – 721302

smsh@maths.iitkgp.ac.in

- **The term data is quite old. There are instructions for collecting data for agriculture, number of workers, crime, business etc. in Kautilyas's Arthshastra. There are references to data collection by governments in ancient Rome, Greece**

- **Data Scientist: - A person who is better at statistics than any software engineer and better at software engineering than any statistician. — Josh Wills**

- **Data analytics refers to techniques and methodologies to analyze, interpret data and aid in decision making for the practitioner in his/her domain.**
- **The first phase in any problem is identification of objectives and relevant data, its source, e.g., it could be primary sampling, secondary data, continuously evolving data etc.**

- **Primary data: Most government organizations collect primary data on its citizens : expenditure on health, education, food, luxuries; monthly income, employment status etc. In India NSSO and CSO usually do this. Governments and other organizations also routinely collect data on production in mining, manufacturing, output in services and**

agricultural sectors, sales – consumers' preferences, product prices, locations of warehouses, factories, workers in industries – their numbers, working hours, pay packages, workplace facilities; lifetime data – births and deaths, mortality due to various factors, distribution of populations in geographic zones; data in sports – performances, outcomes, players in

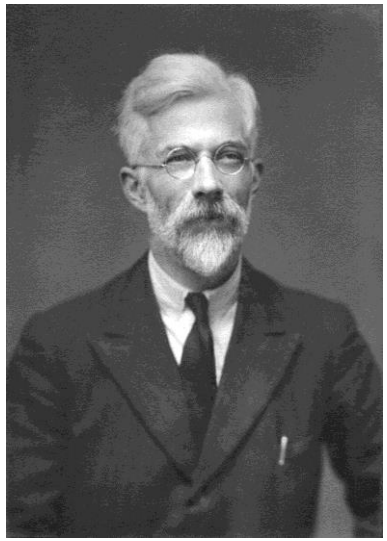
different categories according to sports, expertise, earnings, longevity in sports; health data – proportion and prevalence of different diseases in populations, expenditure on treatments, recovery rates, survival times; and so on.

- **Secondary Data: Most of other sectors make use of primary data collected by other organizations for their limited objectives. If**

the primary data is from a reliable source, then the secondary data can be used to get formal inferences. The use of secondary data is helpful as it saves times and resources for other organizations so that they can effectively spend their resources on devising strategies for their implementations.

- **It is important to understand the difference between complete enumeration and sampling from populations. A simple example is – exit polls and election results.**
- **The early methodologies for sampling and then development of statistical methods to analyze these have been developed by**

Karl Pearson



R. A. Fisher

Jerzy Neyman

E.S. Pearson

P C Mahalanobis

P.V. Sukhatme

One of the important books:

Sampling Techniques by – W.G. Cochran, Wiley

- **Major Steps in a Sample Survey**
- **Identifying objectives**
- **Specifying the population to be sampled, for example, workers in dangerous tasks in mines, strength of walls in underwater tunnels, life of equipments engaged in hazardous tasks etc.**
- **Data to be collected – relevant data to be taken. Many times useless information is**

also collected. This usually contaminates data. For example in a survey on health status if there are questions on family income, there may be avoidance in receiving feedback.

- **Degree of precision required – many times overemphasis on 100% accuracy results in delay, cost escalation, whereas it may not**

contribute too much to fulfilling the objectives of the study.

- **Methods of measurement – in a health survey, it may be questionnaire based or based on actual check up by medical professionals. In safety check up in industrial unit, it could be based on interviews with the managers, previous**

records or actual safety audit by professionals.

- **The frame – identification of unit of population to be sampled, for example – electronic item, manufacturing unit for that electronic item, individuals, families, colonies, towns etc.**
- **Some Important Sampling Methods**

1. **Simple Random Sampling (with or without replacement)**
2. **Stratified random sampling**
3. **Systematic sampling**
4. **Single stage, two stage and multi-stage sampling**
5. **Double Sampling**
6. **Subsampling**
7. **Cluster sampling**

8. Bootstrapping

- **The Pretest – it always improves the quality of the survey**
- **Organization of Field Work – identifying workers, training them, early checking of first data sets, improvement in methods**
- **Cleaning and Preparation of Data – the data may have missing values, incorrect entries, inconsistencies. These should be**

very properly inspected to make the data conducive for further modelling and analysis.

- **Statistical Modelling - Data is manipulated to extract information out of it. The mathematical foundation of Data Science is Statistics and Probability. Without having a clear knowledge of statistics and probability, there is every possibility of**

misinterpreting data and reaching at incorrect conclusions. This leads to harmful consequences for companies, industries, economy etc. Therefore, probability and statistics play most important role in Data Science and Data Analytics.

- In the remaining part of this lecture I will explain some important topics of Probability and Statistics.

- **The terms such as chance, randomness, uncertainty, probably etc. have been in use since time immemorial. People realized long back that things do not happen as planned.**
- **Examples of uncertainty in events:**

(a) Presence/absence of mineral deposits in a specified area

(b) Estimate of total deposits of minerals

(c) Safety required for extracting mineral deposits from specifies mines

(d) Quality of mineral deposits from specified mines

(e) Amount of rainfall during a monsoon season in specified area/state/country over a day/week/month/season

(f) Weather tomorrow

(g) Height that a child born today will achieve when he/she is adult

(h) The total amount of food grains production this year

(i) The longevity of a person

(j) The height of storm surge in a cyclone

(k) The time to get cured after taking medicine for a certain disease

(l) Systolic/dylostic blood pressure of a person

(m) Number of students scoring more than 75% marks in a test

(n) total life of an equipment, say a tubelight

- This is plain truth: everyone ought to keep a sharp eye on main chance. –Plautus (Roman playwright, 200 B.C.)

- **It is a common notion that randomness is an indispensable ingredient of creative arts..... Randomness is an intrinsic feature of human thought, not something which has to be artificially inseminated, whether through dice, decaying nuclei, random number tables, or what-have-you. It is an insult to human creativity to imply that it relies on arbitrary sources. – Hofstadler**

- **The theory of probability originated in the middle of seventeenth century. The studies by Fermat (1601-1665), Pascal (1623-1662), Huygens (1629-1695), James Bernoulli (1654-1705) led to laying down foundations of probability theory.**
- **The physician and mathematician Gerolamo Cardano (1501-1575) was probably the first to develop systematic**

theory of probability. He was a compulsive gambler and his interest in probability was to find probabilities of various outcomes in dice problems.

Gerolamo Cardano

- **His work was published in 1663 long after his death in a 15 page book consisting of 32**

small chapters “Liber de ludo aleae” -> meaning “The Book on Games of Chance”.

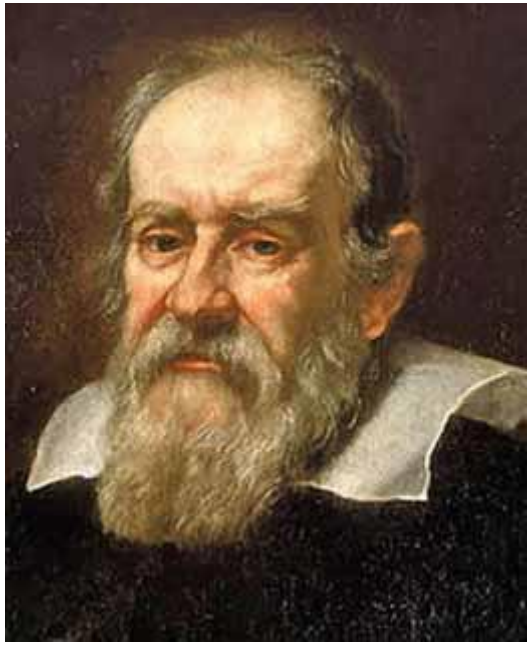
- **One may read Ore’s “Cardano: The Gambling Scholar”.**
- **He solved the following problem:**
- **How many throws of a fair die do we need in order to have an even chance of at least one six?**

- **Cardano gave an erroneous reasoning to get the answer as 3.**
- **Correct Solution: Let *A* be the event " a six shows in one throw of a die". Then *P(A)*=1/6. Then assuming independence of throws;**

  ***P*(a six shows at least once in *n* throws) = $1 - (5/6)^n$.**

- **Now solving $1 - (5/6)^n > 0.5$ gives $n > 3.8$. So the number of throws is 4.**
- **Galileo Galilei (1564-1642) was asked the following problem.**
- **Suppose three dice are thrown and the three numbers obtained are then added. The total scores of 9, 10, 11, and 12 can be obtained in six different combinations.**

Why then is a total score of 10 or 11 more likely than a total score of 9 or 12?



Galileo Galilei

- **Solution: We find the combinations for each of 9, 10, 11, and 12.**

9	**1, 2, 6**	**6 permutations**	**25 cases**
	1, 3, 5	**6 permutations**	
	1, 4, 4	**3 permutations**	
	2, 2, 5	**3 permutations**	
	2, 3, 4	**6 permutations**	
	3, 3, 3	**1 permutation**	
10	**1, 3, 6**	**6 permutations**	**27 cases**
	1, 4, 5	**6 permutations**	

	2, 2, 6	**3 permutations**	
	2, 3, 5	**6 permutations**	
	2, 4, 4	**3 permutations**	
	3, 3, 4	**3 permutations**	
11	**1, 4, 6**	**6 permutations**	**27 cases**
	1, 5, 5	**3 permutations**	
	2, 3, 6	**6 permutations**	
	2, 4, 5	**6 permutations**	
	3, 3, 5	**3 permutations**	
	3, 4, 4	**3 permutations**	

12	**1, 5, 6**	**6 permutations**	**25 cases**
	2, 4, 6	**6 permutations**	
	2, 5, 5	**3 permutations**	
	3, 3, 6	**3 permutations**	
	3, 4, 5	**6 permutations**	
	4, 4, 4	**1 permutation**	

- **So we can see that probability of 10 or 11 is more than 9 or 12.**

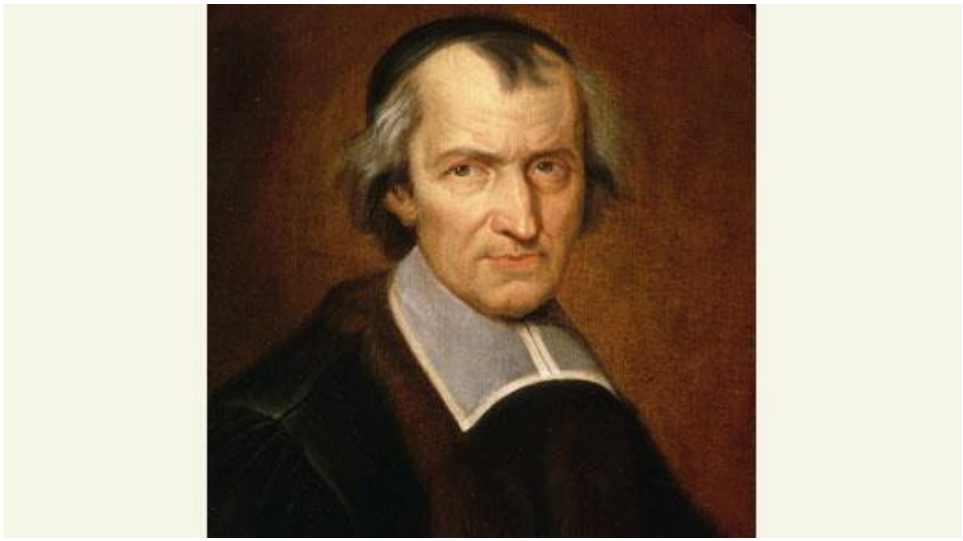
- **Although Cardano was the first to study probability, it was correspondence between two French mathematicians Pascal and Fermat which led to the development of a proper mathematical theory of probability. The main problem that they discussed was posed by a gambler friend Chevalier de Mere, the so called “The problem of points”**

(1654). This was related to division of prize money between two players.

Blaise Pascal

Pierre de Fermat

- **The term “Probability” as a numerical measure was first used in the book “La Logique, ou l’Art de Penser (1662, English title: Port Royal Logic)) by Antoine Arnauld (1612-1694) and Pierre Nicole (1625-1695).**



Antoine Arnauld

Pierre Nicole

- **Dutch astronomer and mathematician Christiaan Huygens (1629-1695) was aware of problems discussed by Fermat and Pascal, in particular “The gambler’s ruin problem” and published a solution in his book (English title: On the calculations in games of chance (1657)).**

Christiaan Huygens

- **The Gambler' Ruin Problem: Two players *A* and *B* have initial amounts of money Rs. *M* and Rs. *(N – M)* respectively. They play**

a game in which the probability that A wins a round is p and the probability that B wins a round is $q = 1 - p$. Each time a player wins he/she gets Rs. 1 from the other player. What is the probability that A will eventually get all of B's money?

Solution: Let $P_{M,N}$ = probability that A will eventually win all Rs. N (that is, ruins B), starting with Rs. M. Then

$P_{M,N} = p\,P_{M+1,N} + q\,P_{M-1,N}$

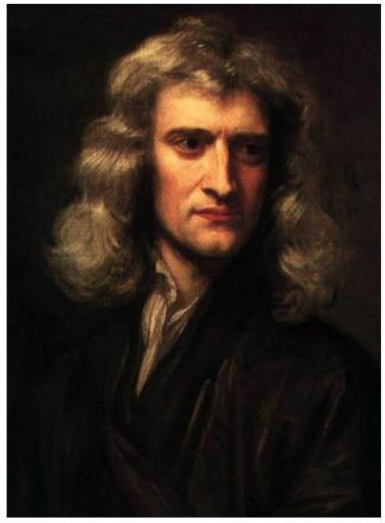
with $P_{N,N} = 1$, $P_{0,N} = 0$.

The solution to this difference equation is

$$P_{M,N} = \begin{cases} \dfrac{(q/p)^M - 1}{(q/p)^N - 1}, & \text{if } p \neq q \\ \dfrac{M}{N}, & \text{if } p = q \end{cases}$$

Huygens solved one special case of this problem, when the game involved tossing of

three dice. Player *A* wins if number 11 shows and Player *B* wins if number 14 shows. Both players start with 12 pieces of money. Huygens got the chances of *A* winning to that of *B* as 244,140,625 to 282,429,536,481.



Sir Isaac Newton

(1643 – 1727)

- **The following problem was posed by member of Parliament Samuel Pepys (1633-1703) to Sir Isaac Newton (1623-1727).**

- **Problem: A asserts that he will throw at least one six with six dice. B asserts that he will throw at least two sixes by throwing 12 dice. C asserts that he will throw three sixes by throwing 18 dice. Which of the three stands the best chance of carrying out his promise?**

- **Solution: Let X denote the number of sixes when n balanced dice are thrown independently, $n = 1, 2, 3,....$**
  Then $X \sim$Bin $(n, 1/6)$.
- **For A, $X \sim$Bin (6, 1/6) and**
  $P(X \geq 1) = 1 - P(X = 0) = 1 - (5/6)^6 = 0.665$
- **For B, $X \sim$Bin (12, 1/6) and**
  $P(X \geq 2) = 1 - P(X = 0) - P(X = 1)$
  $= 1 - (5/6)^{12} - 12\,(1/6)(5/6)^{11} = 0.619.$

- **For C, X~Bin (18, 1/6) and**
  $P(X \geq 3)$
  $= 1 - P(X = 0) - P(X = 1) - P(X = 2)$
  $=1-(5/6)^{18}- 18\,(1/6)(5/6)^{17} - 153(1/6)^2(5/6)^{16}$
  $= 0.597.$
- **Thus A is more likely to carry out his promise than B and B is more likely to carry out his promise than C.**

- **Though Newton arrived at the correct answer his solution was not completely correct. Whenever, somebody approached Newton with a probability problem, he used to recommend him to Abraham de Moivre (1667-1754).**
- **Matching Problem: Suppose n couples go to a dance floor and husbands are randomly**

paired. What is the probability that at least one husband and wife are paired correctly?

- **This problem is the best known contribution of French mathematician Pierre Remond de Montmort (1678-1719) and had origins in the works and comments by John Bernoulli (1667-1748), Jacob Bernoulli (1654-1705) (the two were**

brothers and archrivals) and Nicholas Bernoulli (nephew) (1687-1759)

- **Let *A* be the event that at least one pair is correctly matched. Then**

  $$A = B_1 \bigcup B_2 \bigcup \cdots \bigcup B_n,$$

  where B_i is the event *that* i^{th} pair is correctly matched.
- **By the general addition rule**

$$P(A) = P\left(\bigcup_{i=1}^{n} B_i\right) = \sum_{i=1}^{n} P(B_i) - \sum_{i<} \sum_{j} P(B_i \cap B_j$$

$$+ \sum_{i<} \sum_{j<} \sum_{k} P(B_i \cap B_j \cap B_k) - \cdots + (-1)^{n+1} P\left(\bigcap_{i=1}^{n} B_i\right)$$

$$= \binom{n}{1} \frac{(n-1)!}{n!} - \binom{n}{2} \frac{(n-2)!}{n!} + \binom{n}{3} \frac{(n-3)!}{n!}$$

$$- \cdots + (-1)^{n+1} \frac{1}{n!}$$

$$= \frac{1}{1!} - \frac{1}{2!} + \frac{1}{3!} - \cdots + \frac{(-1)^{n+1}}{n!}$$

- **As $n \to \infty$, this probability converges to $1 - e^{-1} \approx 0.6321$.**
- **This is perhaps the first exponential limit in probability.**
- **Golden Theorem of Jacob Bernoulli (1713)**
- **Theorem: The probability of observing an equal number of heads and tails when a fair**

coin is tossed $2n$ times is approximately $1/\sqrt{\pi n}$ for large n.

- **Proof: Let X denote the number of heads in $2n$ independent tosses of a fair coin.**
- **Then X~Bin($2n$, ½)**
- $P(X=n)=\frac{1}{2^{2n}}\binom{2n}{n}=\frac{1}{2^{2n}}\frac{2n!}{(n!)^2}$
- **Use Stirling's formula $n!=\sqrt{2\pi}\,n^{n+1/2}\,e^{-n}$ for large n, we get**

- $P(X=n)=\frac{1}{2^{2n}}\frac{\sqrt{2\pi}(2n)^{2n+1/2}e^{-2n}}{\left(\sqrt{2\pi}(n)^{n+1/2}e^{-n}\right)^2}$
- $=\frac{1}{\sqrt{2n}}$
- **For n=10 this is 0.1262, for n =100, it is 0.0399, for n =10000, it is 0.004 and for n =1000000, it is 0.0004.**

Jacob Bernoulli

Lepicallidus

Bizi Takip Edin!

$$y' = p(x)y + q(x)y^n.$$

John Bernoulli

A. N. Kolmogorov

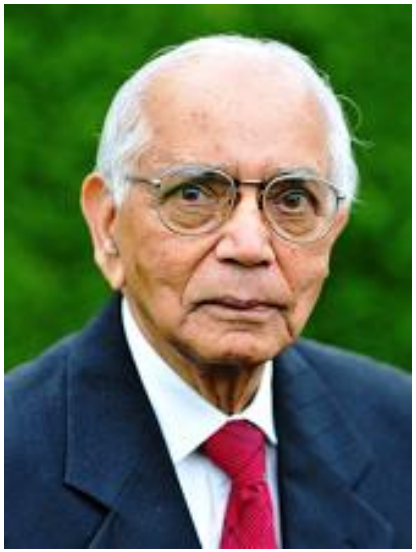
Pierre-Simon Laplace

Richard von Mises

Bruno de Finetti

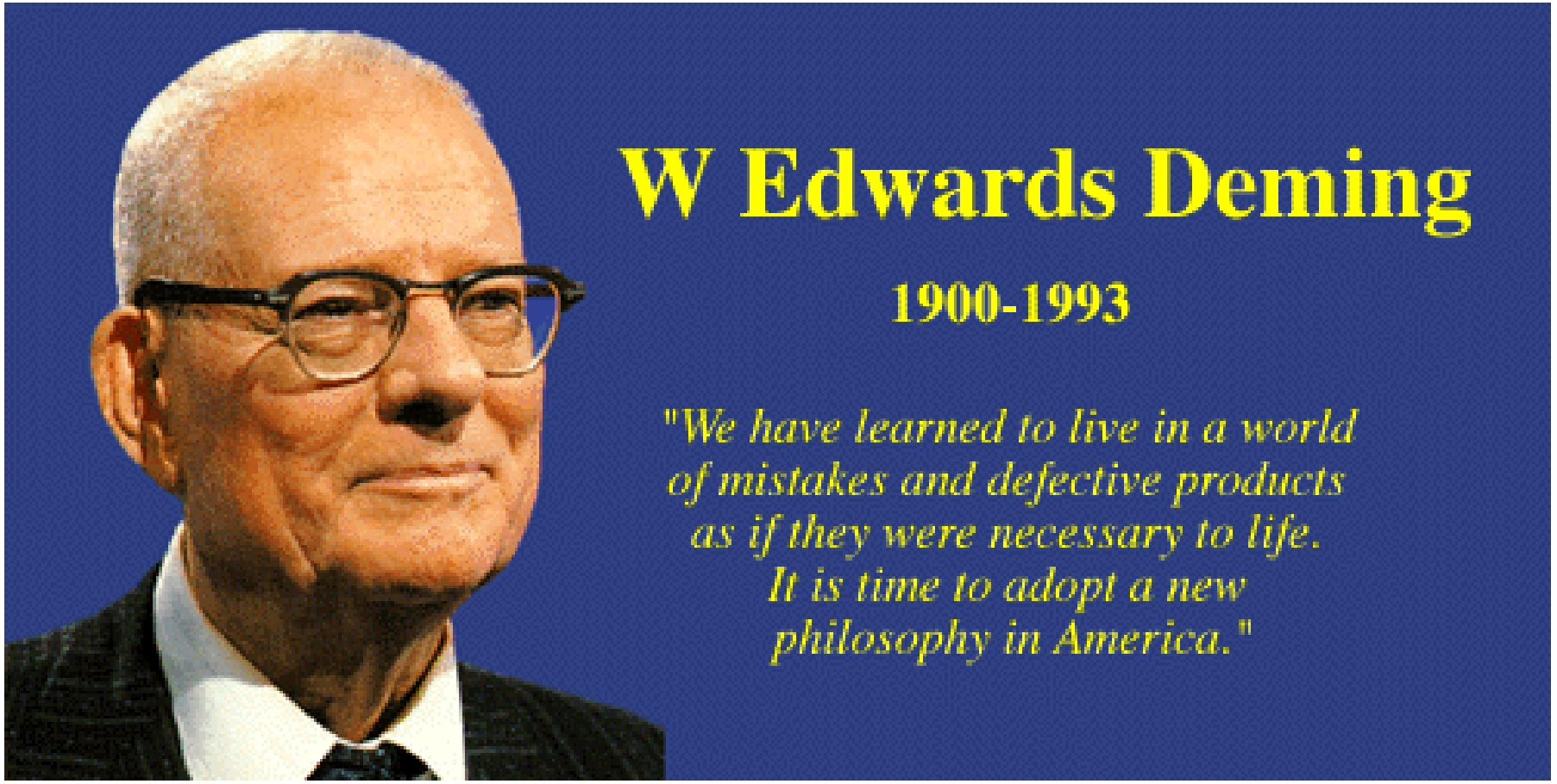
R.C. Bose

S.N. Roy

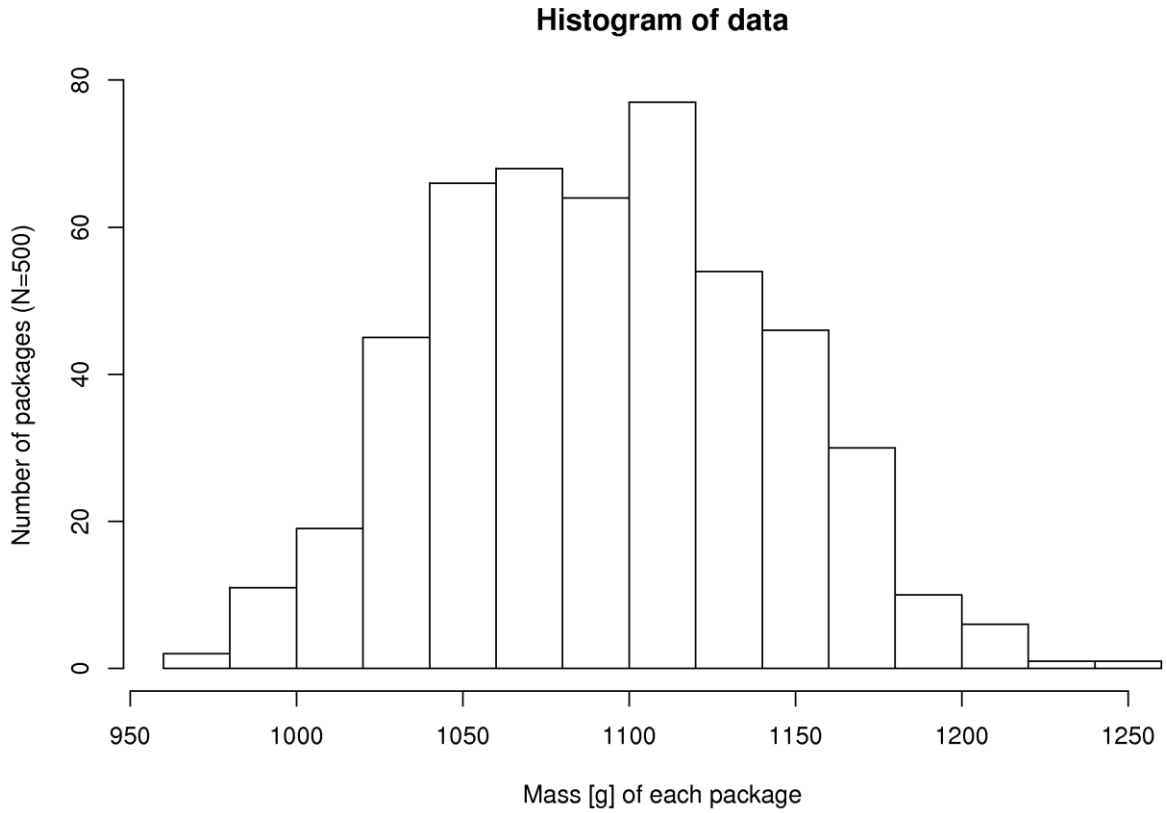


C. R. Rao

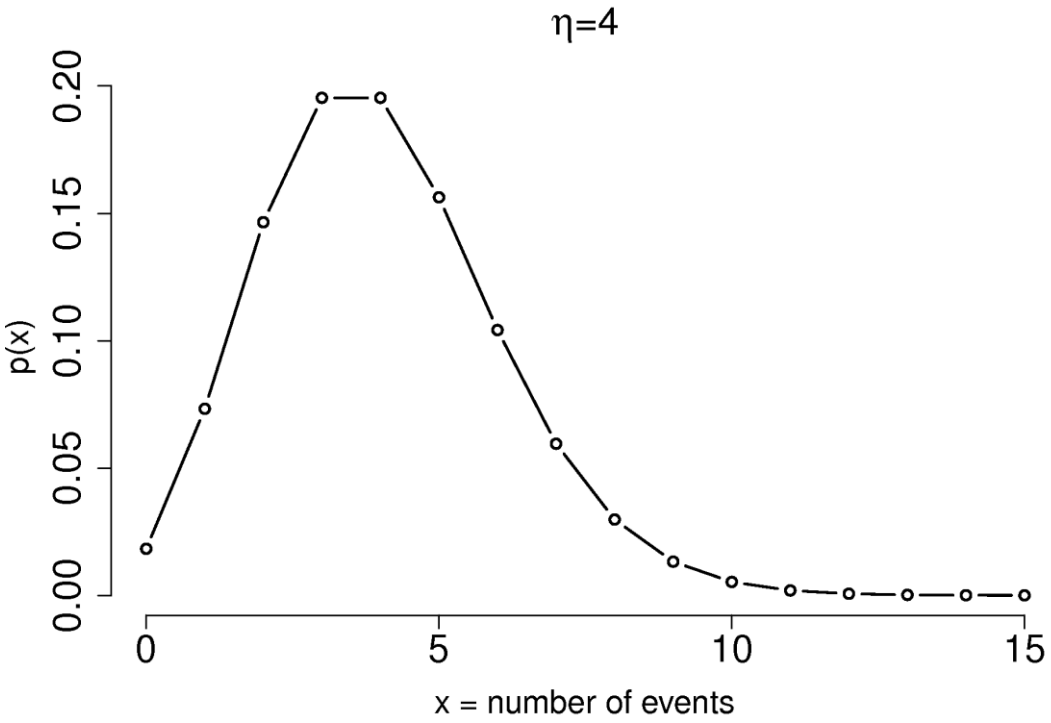
Abraham Wald


W Edwards Deming
1900-1993
"We have learned to live in a world
of mistakes and defective products
as if they were necessary to life.
It is time to adopt a new
philosophy in America."

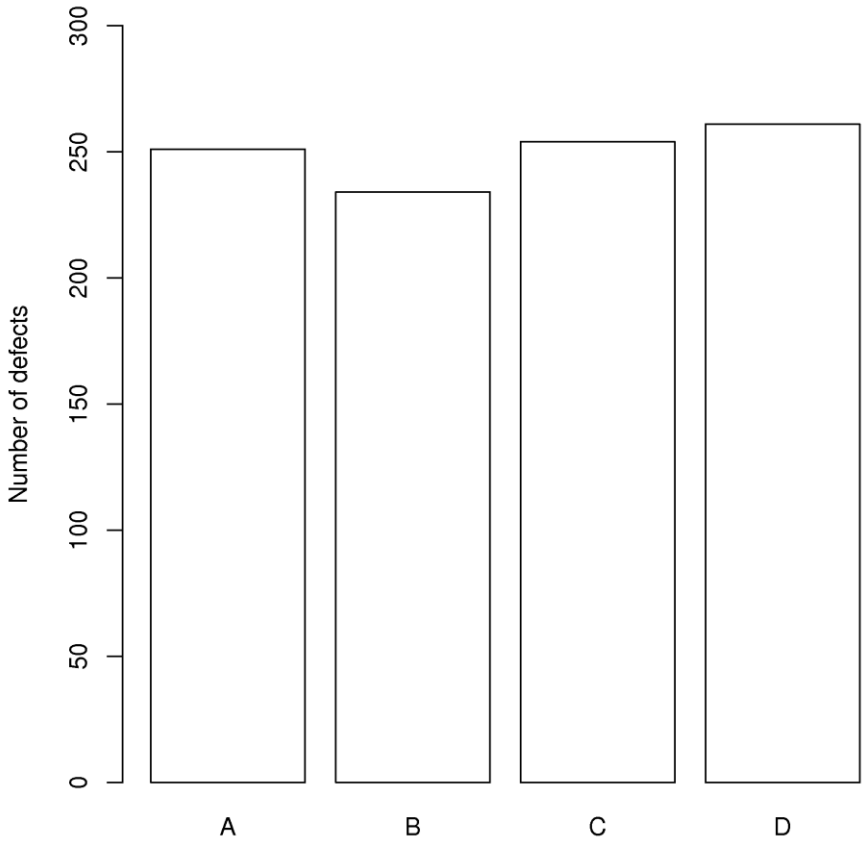
- **There are various methods which are used for drawing conclusions from the data based on objectives.**
- **Modelling of Distribution – In most applications first aspect is to look for the distributional model which will correctly describe the given data. This is done using histograms, frequency curves, cumulative frequency curves.**



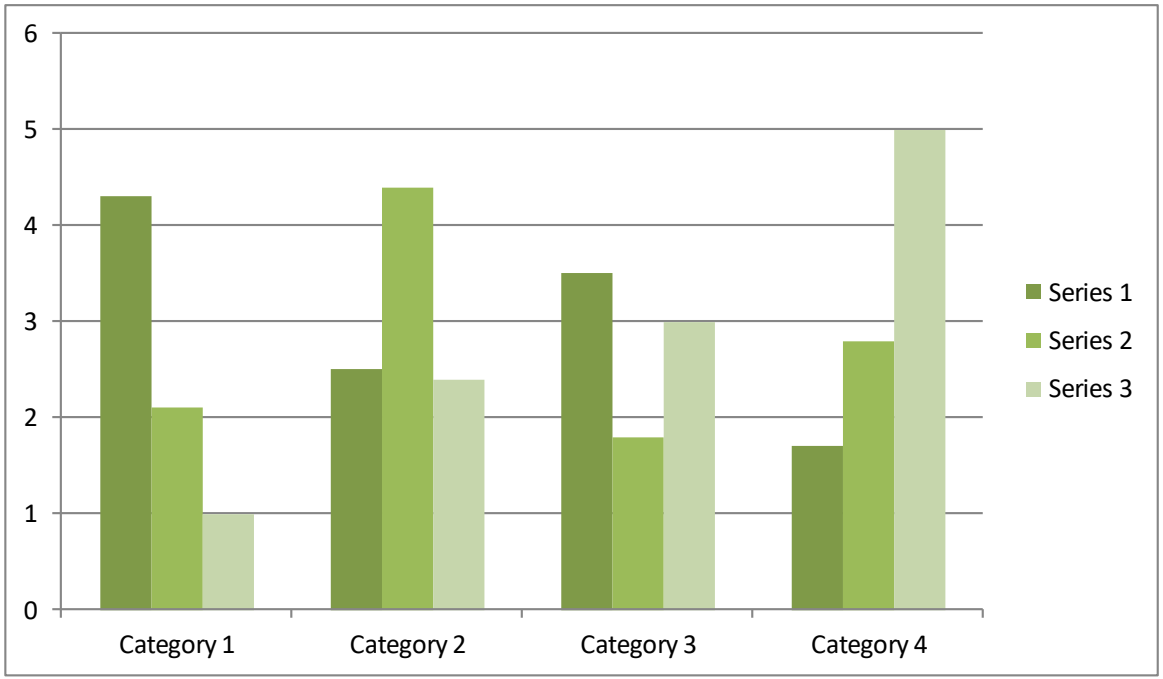
- **Each package contains somewhat more than 1 kg to completely fill the carton. So the distribution is slightly varying than the normal.**



- **This data seems to model a Poisson distribution.**

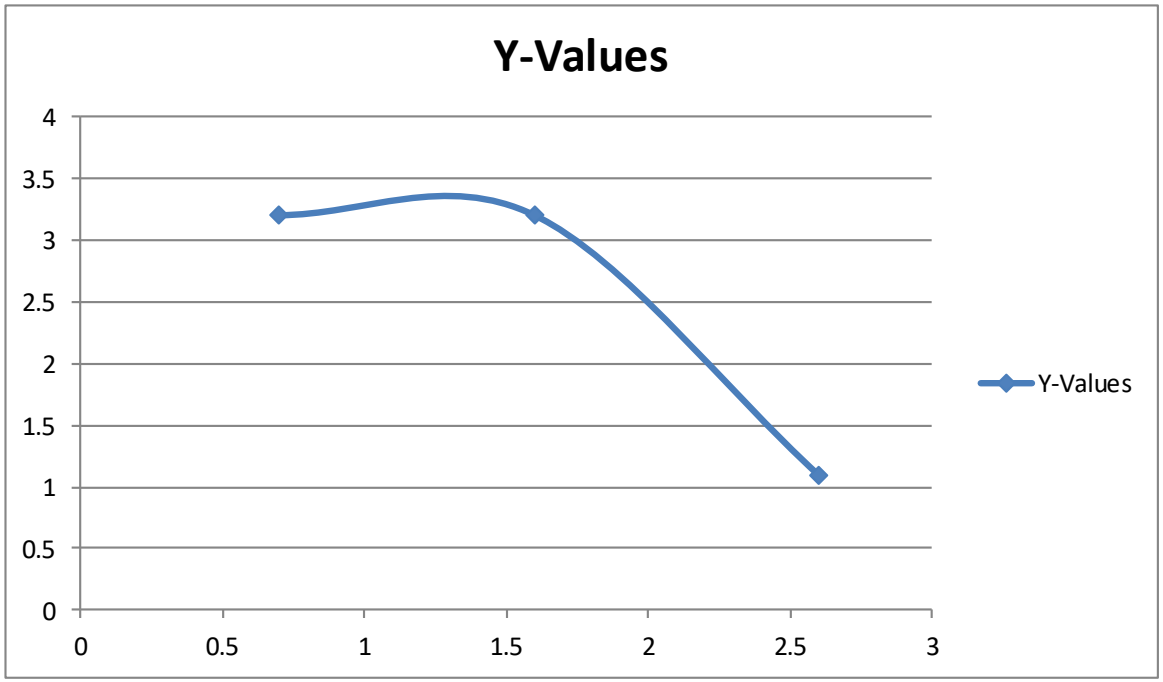


- **The above diagram suggests a uniform distribution.**

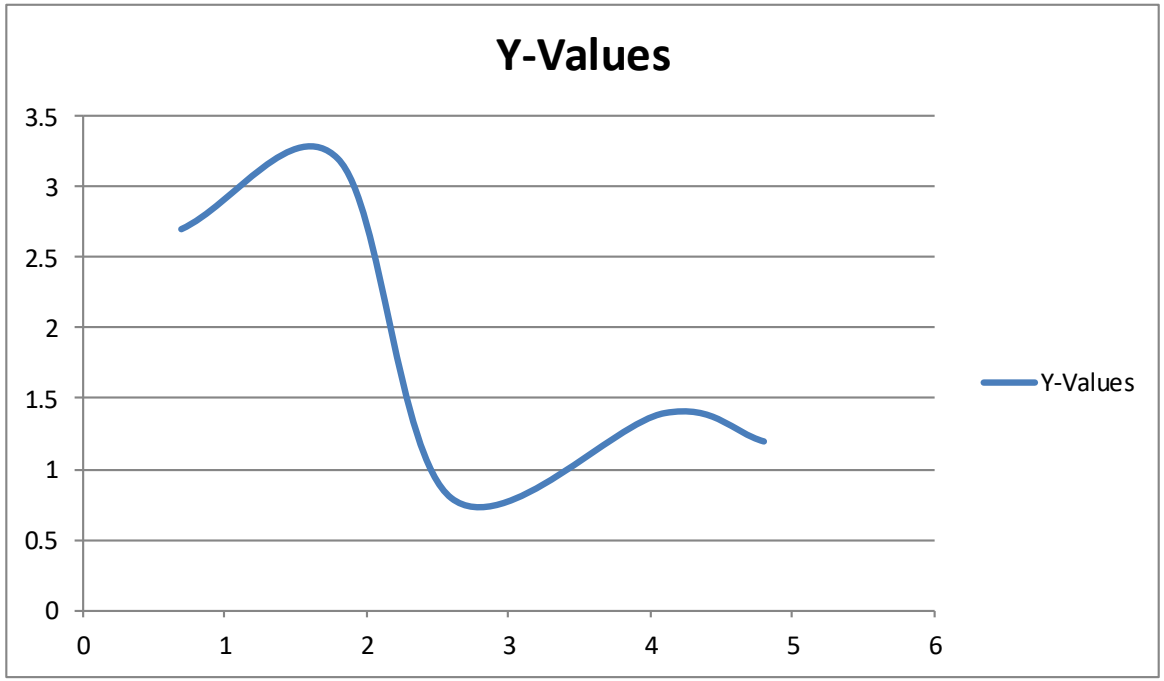


- **Here we have four different histograms. The first one represents a positively skewed distribution. The second one is symmetric distribution. The third is bi-modal and the fourth is negatively skewed.**

- **Regression Analysis – We often have several variables in the study. We need to find relationship between them For example, expenditure on safety and number of accidents in mines, rate of interest and rate of inflation, expenditure on girls' education and infant mortality rate/mother mortality rate etc.**



- **In the above scatter plot we see that the relation between X and Y values is represented by a second degree curve.**



- **In this plot we can see that the relationship is described by a mixture of curves. That is it changes with magnitude of values.**
- **We have various models for describing relationship between variables.**
- **If we have two variables X and Y, then we can use**

1. **Simple linear regression model,**
2. **Polynomial regression model**
3. **Logarithmic model**
4. **Exponential model**

- **If we have more than two variables, then relationship is described by multiple regression models which may be linear or non-linear. We also use logistic regression, indicator variables ridge regression models etc.**
-

- **Residual Analysis – We need to analyze the residuals to check the validity of fitted model.**
- **Multicollinearity**
- **Heteroscedasticity**
- **Stepwise regression**
- **Autocorrelated variables**
- **When the data is observed over time we have to use Time Series Models**