Probability and Statistics

Debolina Ghatak

Indian Institute of Technology (Jammu)

July 26, 2020

Outline

Probability

– A branch of mathematics that studies random phenomenon.

Uses

To name a few areas where probability theory is used:

- Statistics
- Physics
- Finance
- Computer Science
- Cryptology
- Computer Science
- Machine Learning/Artificial Intelligence

Problem of points: Beginning of probability

One of the classical problems that led to the discovery of modern probability theory is the **Problem of points** or the **Problem of division of stakes**.

- Game between two players who have equal chance of winning for a prize money S.
- Game consists of r rounds.
- Due to unavoidable circumstances all the rounds cannot be completed and game stops before completion of r rounds.

Question: How can the money be divided among the players "Fairly"?

Proposed Solutions

- *Luca Pacioli* gave the first solution to the problem in 1494 stating to divide the prize money according to the proportion of rounds won by each player.

  Drawback: If one player wins two rounds out of, say, 21 we wins the total prize. This seemed to be really unfair because anything could have happened in the remaining rounds.

- *Tartaglia* proposed a new solution where the above problem was resolved. He proposed to take the ratio of the size of the lead and the total number of rounds and award that amount to the winning side. The rest of the money to be divided equally.

  Drawback: However, this solution, too, was not free of faults. A game resulting in 6-8 and 8-10 (21 rounds) have same size of lead and second length of game resulting in similar distribution of money but in the later case there is clearly a much higher chance of the leader winning the game than in the former case due to difference in remaining rounds.

Fermat-Pascal Solution

The problem remained unsolved until 1654 when the two famous mathematicians Pascal and Fermat came up with their solution. Their solution is till now the best known solution to the given problem and can be describes as follows:

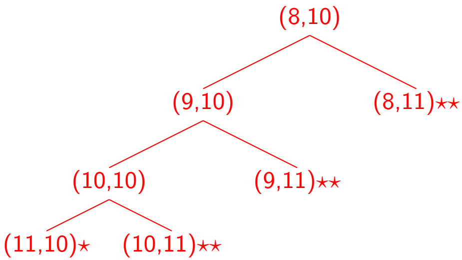
- If the first player needs s_1 more rounds to win and the second player needs s_2 more rounds to win then the game would surely end in the next $s_1 + s_2 - 1$ rounds.
- Considering an equal chance of winning for each player in all the upcoming matches one can see in how many possibilities each player is winning.
- The stake is divided according to the ratio of possibilities among remaining $2^{s_1+s_2-1}$ possibilities.

Elaborative Example

Consider a tennis match between two players.

- 21 rounds are to be played.
- If a player wins 11 rounds he wins.
- Suppose the game stops at 8-10.

- Weight given to Player 2: $\frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{7}{8}$
- Weight given to Player 1: $\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$
- Thus Player 1 gets **S/8** and Player 2 gets **7S/8** prize money.



Mathematical Definition

"**It is remarkable that this science, which originated in the consideration of games of chance, should become the most important object of human knowledge**" — Marquis de Laplace.

Book Ref. *Analytical Theory of Probability*.

- Ω usually denotes the *sample space* or a set of all possible outcomes corressponding to a random experiment.
- Mathematically, the probability function, defined by *Kolmogorov*, is a function that maps from an event space (a sigma field defined on Ω) to $[0, 1]$ and satisfies *the three axioms by Kolmogorov*.
- Examples of some known probability functions are:
  - Classicial probability function defined for finite sample spaces.
  - Geometric probability functions defined for cases where the sample space is a subset of the *n* dimentional Euclidean Space.

Conditional Probability and Independence

- *Conditional Probability* of A given B is given by,

$$P[A \mid B] \equiv P[\text{ Occurance of } A \mid B \text{ has already occurred}]$$

- Two events A and B are said to be independent if the occurance of one does not affect the occurance of the other. In such case,

$$P[A \mid B] = P[A] \text{ , and } P[B \mid A] = P[B]$$

Probability Distributions

Random Variables are numerical quantities associated with a random experiment. Although the values of a random variable are not fixed, the probability of a random variable taking values in a certain range has a fixed value.
Every random variable is identified by:

- It's *Cumulative Distribution Function (CDF)* $F : \mathbb{R} \to [0, 1]$,

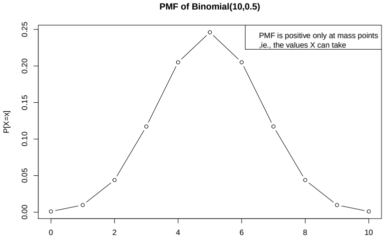
$$F(x) = P[X \leq x] \text{ , for all } x \in \mathbb{R}.$$

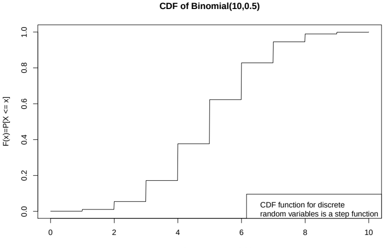
- It's *Probability Mass Function (PMF)* or *Probability Density Function (PDF)* for discrete and continuous valued random variables respectively.

Example discrete: Binomial random variable

Number of successes in fixed number of trials

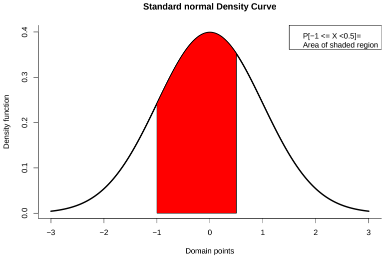
- Let X denote number of heads in 10 tosses of an unbiased coin.
- X can take values between 0 to 10 with certain probabilities.
- X is Binomial (10,0.5).

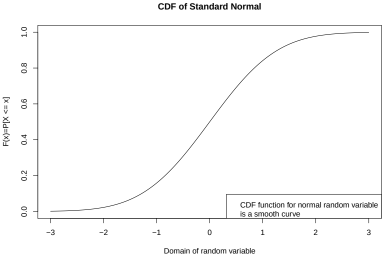




Example Continuous: Standard Normal random variable

A standard normal random variable can take any real value but has fixed probability of taking value in any given region.





Density function is given by

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \text{ , for all } x \in \mathbb{R}$$

Random Sets

A set of random variables defined on same probability space are called random sets. One can define,

- Random Vectors
- Random Sequence
- Random Matrix
- Stochastic Process

Real-life Scenario with Data

Suppose two players Mr. Red and Mr. Green are competing in a tennis game. We want to calculate *the probability that Mr. Red wins against Mr. Green*.

- Q How to know? No way of guessing.
- Q But what can we get? Results of Matches.
- E Say we conduct 20 games and scorecard of Mr. Red is given by

$$WWLWWLLWLWWLWLLWLWWL$$

Using this information or "data" we can try to guess the true value of the probability.

Population and Sampling

- A *Population* is an entire set of possible data values. Is usually unknown and involve unknown parameters.
- The *Sample* are the observed values.
- In the previous example, *WWLWWLLWLWWLWLLWLWWL* is the observed sample which comes from a population of all possible results in 20 matches.

Parametric Estimation Problem

Suppose our data comes from some known distribution F_θ with some unknown parameter $\theta \in \Theta$ which needs to be guessed.
Like in the previous example, the data, is given by the random variables,

$$X_i = \begin{cases} 1 & \text{if Mr. Red wins against Mr. Green} \\ 0 & \text{if Mr. Green wins against Mr. Red} \end{cases} \quad \text{for } 1 \leq i \leq 20$$

where each $X_i \sim Binomial(1, p)$ and p is unknown that needs to be guessed.
In a similar situation guessing the value of the unknown parameter of interest from the available data is known as the problem of estimation or point estimation.
Some famous estimators known in literature are:

- The *Minimum Mean-Squared Estimator* given by,

$$\min_{T(X_1, X_2, \cdots, X_n)} (T(X_1, X_2, \cdots, X_n) - \theta)^2$$

- The *Maximum Likelihood Estimator* given by the statistic, which when chosen as θ, maximizes the likelihood of the available data.

Hypothesis Testing Problem

- In a general testing of hypothesis problem there is a certain claim which needs to be validated. This claim is also called the *null hypothesis* of the problem.
- Let $X_1, X_2, \cdots, X_n$ be a random sample from some distribution with distribution function $\{F_\theta : \theta \in \Theta\}$, where the functional form of F is known, but θ is unknown.
- The *null hypothesis* is a claim on θ falling into a certain region, say $\theta \in \Theta_1 \subset \Theta$. The *alternative hypothesis* is usually the complement of it, i.e., θ falling into the region $\Theta - \Theta_1$.
- Such tests are called *parametric tests*.

Example

Considering the tennis game as before, one may want to bet on Mr. Red because he believes Mr. Red has a probability of winning more than 0.75.
In such a scenario he would test,

$$H_0 : p > 0.75 \text{ against } H_1 : p \leq 0.75$$

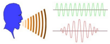
and bet if his test gets accepted.

Non-Parametric set up

Unlike the previously discussed situations, in real-life one usually comes accross data for which the distribution itself is unknown and not only some parameters. In such scenarios one usually wants to estimate the CDF or PDF/PMF of underlying data or test whether the data actually comes from a known distribution. Such kind of problems are dealt in non-parametric infernece.

Application in Digital Processing

- Input Source $\Rightarrow$ Binary data
- Noise generates in channel, eg., Thermal Noise, Man-Made noise, Atmospheric noise etc.
- Receiver decodes binary data.



BIT ERROR Probability

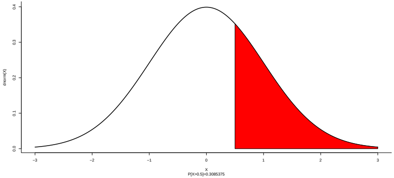
Accuracy is specified in terms of BIT-ERROR Rate (Probability of making bit error)

Additive Noise Model

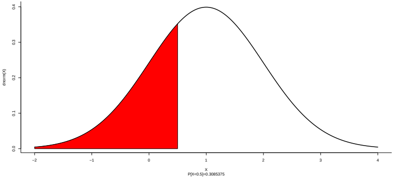
- $Y = X + N$
- X is binary, N is Gaussian noise, Y is received data which needs to be decoded.
- $P(\text{Making a Bit Error}) = P(Y > 0.5 \mid X = 0)P(X = 0) + P(Y < 0.5 \mid X = 1)P(X = 1)$.

P[Bit Error] for N(0,1) Noise

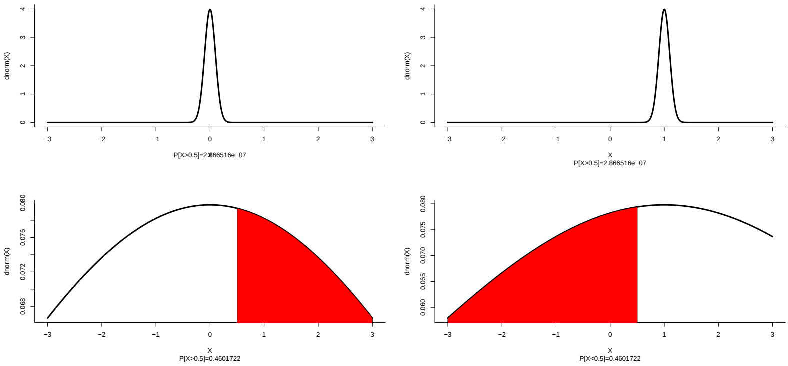
X=0



X=1



Effect of varying σ (Used scales are 0.1 and 5)



Remark: P(Making a Bit-Error) increases as σ increases.

Solution

Repeat each bit r times and generate $Y_1, Y_2, \cdots, Y_r$,

$$Y_i = X + N_i \text{ for } 1 \leq i \leq r$$

Thus, $\bar{Y} = X + \bar{N}$ and,

$$\bar{N} = \frac{N_1 + N_2 + \cdots + N_r}{r} \sim N(0, \sigma^2/r)$$

Thus noise variance is reduced resulting in reduction of probability of making bit error.

Thank You