Probability Distributions and Estimation of Parameters of some Continuous Distribution Functions

R. Srinivas
Associate professor of Mathematics, SGK Government Degree College, Vinukonda, Palnadu district, Andhra Pradesh, India

ABSTRACT: A probability distribution is a mathematical function that assigns the probabilities of different outcomes to the possible values of a random variable. It provides a way of modeling the likelihood of each outcome in a random experiment. There are two types of probability distributions namely—discrete probability distribution and continuous probability distribution. Estimation is one of the major areas of statistical inference. Statistical inference is the process by which conclusions from the sample data is used to draw conclusions about the population from which the sample was selected. The problem of estimation when some parameter is unknown has received considerable attention of statisticians in recent past. Problem of estimation can be found everywhere: in business, in science, as well as in everyday life. This paper explains Probability Distributions and Estimation of Parameters of Some Continuous Distribution Functions. In particular, the maximum likelihood, method of moment, and Bayes estimators has been derived. Through this paper expected to help the readers to understand the distribution and as well Estimation of Parameters, and to develop the knowledge base for its further applications.

KEYWORDS: Probability Distribution, Estimation of Parameters, Continuous Distribution, Bayes estimators, maximum likelihood.

I. Introduction

Definition 1: (Probability):

Let C be the sample space, that is, the set of every possible outcome of a random experiment [1]. The probability $P(C), C \subset c$ is a function of the outcome of the random experiment defined on subsets of the space C such that

(a) $P(C) \geq 0$,

(b) $P(C_1 \cup C_2 \cup \ldots) = P(C_1) + P(C_2) + \ldots$, where the sets C_i, $i = 1,2,3,\ldots$, are such that no two have a point in common (that is, where $C_i \cap C_j = \emptyset, i \neq j$), and

(c) $P(c) = 1$.

The probability function $P(C)$ tells us how the probability is distributed over various subsets C of the sample space c.

Definition 2: (Random experiment): An experiment in which all outcomes are known in advance, any performance of the experiment that results in an outcome is not known in advance and the experiment can be repeated under identical conditions, is called a random experiment [2]. The result of a statistical experiment is called an outcome.

Definition 3: (Random variable): Let (Ω, S) be a sample space. A finite single valued function which maps Ω into R is called a random variable if the inverse images under X of all Borel sets in R are events [3].

Definition 4: (Distribution function): Let X be an random variable defined on (Ω, S, P). Define a function F on R by $F(x) = P\{w : X(w) \leq x\}$ for all $x \in R$. F is non decreasing, $F(-\infty) = 0$, $F(\infty) = 1$. Then the function F is called the distribution function of the random variable X

Definition 5: (Continuous random variable): Let X be a random variable defined on (Ω, S, P) with distribution function F [4]. Then X is said to be continuous random variable if F is absolutely continuous that is if there exists a nonnegative function $f(x)$ such that, for every real number x, we have $F(x) = \int_{-\infty}^{x} f(t)\, dt$. The function f is called the probability density function of the random variable X. If X is a continuous random variable then we can define its probability density function as below [5].

Definition 6: (Probability density function): Every nonnegative real valued function f can serve as a probability density function of a continuous random variable X, if $f(x) \geq 0$, and satisfies $\int_{-\infty}^{\infty} f(x)\, dx = 1$

Definition 7: (Estimation of parameters): Suppose $F_\theta(x), \theta \in \Theta$ be a family of distribution functions and θ is taken to be unknown. Here we estimate the unknown parameter θ with the help of samples. We study the theory of point estimation and particularly parametric point estimation [6].

Definition 8: (Estimator): If a statistic is used to estimate an unknown parameter θ of a distribution, then it is called an estimator and a particular value of the estimator say $T_n(X_1, X_2, \ldots, X_n)$ is called an estimate of θ. The process of estimating an unknown parameter is known as estimation.

II. Probability Distributions

A probability distribution is a mathematical function that assigns the probabilities of different outcomes to the possible values of a random variable. It provides a way of modeling the likelihood of each outcome in a random experiment [7]. There are two types of probability distributions namely—discrete probability distribution and continuous probability distribution.

1. Discrete Probability Distribution:

A discrete probability distribution describes the probability of occurrence of each value of a discrete random variable. A discrete random variable is a random variable that can only take finite or countably infinite specific values. For example, let X be a random variable representing the number of rainy days in a month at a location. It is obvious that X can take up values belonging to the set of non-negative integers only. Thus, X follows a discrete probability distribution. In this book, discrete probability distribution functions are referred to as probability mass function (pmf) and denoted as $p_X\,(x)$.

Binomial Distribution: Binomial distribution is a discrete probability distribution of the number of occurrences of an event in sequence ofn independent trials of a random experiment with a probability of occurrence of the event be p in each trial [8]. To obtain the probability concerning X we proceed as follows: If p and (1 − p) are the probability of occurrence and non-occurrence for each trial, respectively, then the probability of getting x occurrences (i.e., (n − x) non-occurrences), in any order is $p^x(1 - p)^{(n-x)}$. This is by the virtue of the generalized multiplication rule for more than two independent events. Now, the number of different orders in which x number of occurrences can happen is nC_x, i.e., the number of combinations of x objects selected from a set of n objects. Thus, the probability of x occurrences out of n trials can be expressed as:

$$p_X\,(x; n, p) = nC_x p^x (1 - p)^{(n-x)}$$

$$x = 0, 1, 2, \ldots, n$$

This is the pmf of binomial distribution for X with parameters n and p. The Cumulative Binomial Distribution is expressed as

$$F_X(x; n, p) = \sum_{i=0}^{x} p_X\,(i; n, p)$$

Negative Binomial Distribution: Negative binomial distribution is another discrete probability distribution of the random variable that denotes the number of trials in a Bernoulli process before a specific number (denoted by j) of occurrences [9]. The probability that the j^{th} occurrence happens at the X^{th} (X is the random variable here) trial can be calculated by noting that there must be (j − 1) occurrences in the x − 1 trials preceding the X^{th} trial. The probability of (j−1) occurrences in (x−1) trials can be computed from the binomial distribution (explained before) as $p_X\,(x;\, j\,-\,1, p) = x-1C_{j-1}p^{j-1}(1\,-\,p)^{(x-j)}$, where p is the probability of occurrence in each trial as defined in binomial distribution. Next, the probability of occurrence in X^{th} trial is p. As all the trials are independent, the joint probability distribution function is obtained by multiplying these probabilities ($x-1C_{j-1}p^{j-1}(1\,-\,p)^{(x-j)}$ and p). Thus, probability of X = x, i.e., pmf of the negative binomial distribution is given by

$$p_X\,(x; n, p)\;=\;\mathrm{x}-1C_{j-1}p^{j}(1\,-\,p)^{(x-j)}$$

$$x\;=\;j, j\,+1,\ldots$$

Thus, different functional forms will result for different values of j. The CDF (Cumulative Distribution Function) is expressed as

$$F_{\,X}(x; j, p)\;=\;\sum_{i=j}^{x} p_X\,(i; j, p)$$

Multinomial Distribution: Multinomial distribution is the generalized form of a binomial distribution by assuming each trial to have more than two (i.e., k) possible outcomes. Let us consider n independent trials, with each trial permitting k mutually exclusive outcomes whose respective probabilities are $p_1, p_2 \ldots, p_k$ such that $\sum_{i=1}^{k} p_i = 1$.Considering the outcomes of the first kind, second kind and so on, we are interested in the probability $p\,(x_1, \ldots, x_k)$ of getting x_1 outcomes of the first kind, x_2 outcomes of the second kind and so on. Using the arguments similar to the ones, probability mass function can be developed. The pmf can also be expressed using gamma function as

$$p\,(x_1, x_2 \ldots, x_k) = \frac{\Gamma \sum_i x_i + 1}{\prod_i \Gamma x_i + 1} \prod_{i=1}^{k} {p_i}^{x_i}$$

$$\text{for } x_i\;=\;0,1,\ldots,n$$

where $\Gamma(\bullet)$ is the gamma function. The CDF is expressed as

$$F_X\,(x_i)\;=\;\sum_{X_i<x_i} p\,(x_1, x_2 \ldots, x_k)$$

Hypergeometric Distribution: Hypergeometric distribution is a discrete probability distribution. Let us consider a sample of size n selected from a population of size N. The total possible outcome of the selection is NC_n. The number of ways x occurrences may happen is kC_x, k being the specific possibilities of occurrences. The number of ways (n − x) non-occurrences may happen is $N-kC_{n-x}$ where $(N\,-\,k)$is the total number of possible non-occurrences. Thus considering all the possibilities to be equally likely and for sampling without replacement, the probability of getting 'x occurrences in a sample size of n'is

$$p_X\,(x; N, n, k)\;=\frac{kC_x \times N-kC_{n-x}}{NC_n}$$

For,
$x\;=\;max(0, n\,+k\,-\,N), \ldots, min(n, k)$

where x cannot exceed k and (n − x) cannot exceed (N − k). The CDF is expressed as

$$F_X\ (x; N, n, k) = \sum_{x=\max\ (0, n+k-N)} p_X\ (i; n, p)$$

Geometric Distribution: Geometric distribution is another discrete probability distribution of a random variable that defines the number of trials to get the first occurrence of a particular event in a Bernoulli process. The probability that the first success of a Bernoulli trial occurs on the x^{th} trial can be found using the geometric distribution. In order to attain the first occurrence on the x^{th} trial there must be (x −1) preceding trials whose outcome is non-occurrence. Since the successive outcomes in the Bernoulli process are independent, the desired probability distribution is given by

$$p_X\ (x{:}\, p) = p(1-p)^{x-1}$$

$$for\ x\ =\ 1, 2, \ldots, n$$

The CDF is expressed as

$$F_X\ (x; p) = \sum_{i=1}^{x} p_X\ (i; p)$$

Poisson Distribution: Poisson distribution is a discrete probability distribution of a random variable that describes the probability of a particular number of events occurring within a fixed time interval. Let us consider a Bernoulli process defined over an interval of time and let p be the probability of occurrence of an event in a particular interval of time. If the time interval becomes shorter, the probability of occurrence of the event (p) in the interval also becomes smaller; on the other hand, the number of trials (n) increases. As a result, np (denoted by λ) remains constant, i.e., the expected number of occurrences in a time interval remains the same. In such cases, the binomial distribution approaches to a Poisson distribution and is given by

$$p_X\ (x;\ \lambda) = \lambda^x \frac{e^{-\lambda}}{x!}$$

$$for\ x\ =\ 0, 1, \ldots;\ \lambda\ >\ 0$$

2. Continuous Probability Distributions:
If a random variable can take any possible real value from the range of real numbers, its probability distribution is called a continuous probability distribution. Let X be a random variable representing the annual stream flow at a particular station. It can take any possible value from 0 to ∞. Such random variables (X) will follow a continuous probability distribution. In this book, continuous probability distribution functions are referred to as probability density function (pdf) and denoted as $f_X(x)$. In the following section, we will explain some of the most commonly used continuous probability distributions [10].

Uniform Distribution: Uniform distribution is the simplest and symmetric continuous probability distribution function. It is defined over a range (known as support) such that its occurrence is equally possible (equiprobable) over any subinterval of same length within the support. Let us consider a continuous random process restricted to a finite interval $[\alpha, \beta]$ and the probability of an outcome lying with in a subinterval of $[\alpha, \beta]$ is proportional to the length of the subinterval. Such processes are said to be uniformly distributed over the interval $\alpha\ to\ \beta$. The probability density function for the uniform distribution is as follows:

$$f_X(x) = \frac{1}{\beta - \alpha} \qquad \alpha \le x \le \beta$$

The cumulative density function for the continuous uniform distribution is as follows:

$$F_X\,(x) = \frac{\mathrm{x} - \alpha}{\beta\ - \alpha} \qquad \alpha \le \mathrm{x} \le \beta$$

Exponential Distribution: Exponential distribution is a continuous probability distribution that may take any value between $0\ and\ \infty$, with higher probability of occurrence for lower values. It is an asymmetric distribution. Let us assume that the inter-arrival times of an event is being noted. Since the probability that the event occurs during a certain time interval is proportional to the length of that time interval, it follows an exponential distribution. The continuous probability distribution of the inter-arrival time, i.e., the time between the occurrences of two successive events, can be evaluated by noting the $P\,(X\ \le\ t)$ is equal to $1 - P\,(X\ >\ t)$. Thus, the CDF is

$$F_X\,(x) = 1 - e^{-\lambda x} \qquad for\ x > 0$$

and the corresponding probability density function is given by

$$f\ _X(x) = \frac{\mathrm{d}}{\mathrm{dx}} F_X\,(x) = \lambda e^{-\lambda x}$$

$$\text{for x} \ \ge\ 0, \lambda\ >\ 0$$

Normal Distribution: Normal distribution, also known as Gaussian distribution or bell curve, is a continuous probability distribution. Normal distribution is the most frequently used continuous probability distribution function. When mean is zero and variance is 1, the distribution is called standard normal distribution. It can be noticed that it is symmetrical with respect to mean and the typical shape is known as a bell-shaped curve. The line of symmetry and the shape will change depending on the values of mean and variance respectively. The pdf of the Normal Distribution is given by

$$f\ _X(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad -\infty < x < \infty$$

The CDF of the Normal Distribution is given by

$$F_X\,(x; \mu, \sigma^2) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\ dx \qquad -\infty < x < \infty$$

As stated before the mean and the variance of the distribution are μ and σ^2 respectively, and the coefficient of skewness is 0, as it is a symmetric distribution.

Gamma Distribution: Gamma distribution is a continuous probability distribution that is positively skewed over the positive side of the real line. The gamma distribution can be treated as the sum of exponentially distributed random variables each with the same parameter. The parameter α is the number of random variables following exponential distribution and β is the inverse of parameter of the exponential distributions. Gamma distribution has the probability density function as follows:

$$f\ _X(x) = \begin{cases} \dfrac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/3} & \text{for x} \ \ge\ 0, \alpha\ >\ 0, \beta\ >\ 0 \\ 0 & elsewhere \end{cases}$$

Where $\Gamma(\alpha)$ is the value of the gamma function defined by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\ dx$$

Beta Distribution: Beta distribution is a continuous probability distribution that represents out comes for percentages or proportions over an interval, parameterized by two shape parameters. Beta distribution has both upper and lower bounds. Thus, if a random variable takes values specifically in the interval $(0,1)$, one choice of probability density can be beta distribution. However, the beta distribution can also be transformed to any interval (a, b). The shape parameters of the distribution vary with the nature of the distribution. Considering the usual case of limits as 0 and 1, the density function is as follows:

$$f_X(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} & \text{for } 0 < x < 1, \\ & \alpha > 0, \beta > 0 \\ 0 & elsewhere \end{cases}$$

III. Characteristics of Estimators

Various statistical properties of estimators can be used to decide which estimator is most appropriate in a given situation

Definition 1: (Unbiasedness): A statistic T is an unbiased estimator of of the parameter θ if $iff\ E(T) = \theta$.

Definition 2: (Consistency): Let $X_1, X_2, ...$be a sequence of iid random variables with common distribution function $F_\theta(x), \theta \in \Theta$. A sequence of point estimators $T_n(X_1, X_2, ... X_n) = T_n$ will be called consistent for $\psi(\theta)$ if Tn converges to $\psi(\theta)$ in probability that is

$$T_n \to \psi(\theta), as\ n \to \infty$$

Definition 3: (Efficiency): In general there exists more than one consistent estimators. Thus it is necessary to find some criteria to choose between the estimators. Such a criterion which is based on the variances of sampling distributions of estimators is known as efficiency.

Definition 4: (Sufficiency): An estimator is said to be sufficient for a parameter, if it contains all the information in the sample regarding the parameter. Let $X = (X_1, X_2, ... X_n)$ be a sample from a family of distributions $F_\theta(x), \theta \in \Theta$. A statistic T is sufficient for θ if and only if the conditional distribution of X given $T = t$, does not depend upon θ.

Definition 5: (Completeness): A statistic is said to be complete if the family of distributions of T is is complete. Let $F_\theta(x), \theta \in \Theta$ be a family of pdf's we say the family is complete, if $E_\theta g(X) = 0 \forall \theta \in \Theta$.

IV. Methods of Estimation

Normally there are two different approaches for obtaining point estimators for parameter is known. Namely classical method and decision theoretic approach. Now we outline some of the most important methods for obtaining estimators. Most commonly used methods under classical estimation are as follows.

Method of Moments:

Suppose X is a continuous random variable with probability density function (PDF) $f(x; \theta_1, \theta_2, ..., \theta_K)$ characterized by k unknown parameters. Let $X_1, X_2, ... X_n$ be a random sample of size n from X. Defining the first k sample moments about origin as $\acute{m}_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r, r = 1,2, ... k$. The first k population moments about origin are given by $\acute{\mu}_r = E(X^r)$, which are in general functions of k unknown parameters. Equating the sample moments and population moments yields k simultaneous equations in k unknowns. $\acute{\mu}_r = \acute{m}_r, r = 1,2, ..., k$. The solutions to the above equations denoted by $\widehat{\theta_1}, \widehat{\theta_2}, \ldots, \widehat{\theta_K}$ yields the moment estimators of $\theta_1, \theta_2, \ldots, \theta_K$.

Method of Maximum Likelihood Estimation:

Suppose $(X_1, X_2, \ldots . X_n)$ be a random vector with PDF $f_\theta(x_1, x_2, \ldots . x_n)$, $\theta_n \in \Theta$,where is a multidimensional vector valued unknown parameter. Then the likelihood function is given by $L(\theta\,;\, x_1, x_2, \ldots . x_n) = f_\theta\,(x_1, x_2, \ldots . x_n)$ which is nothing but a function of unknown parameter θ. If $X_1, X_2, \ldots . X_n$ are iid with PDF $f_\theta(x)$, then the likelihood function is $L = (\theta\,;\, x_1, x_2, \ldots . x_n) = \prod_{i=1}^{n} f_\theta(x_i)$. The maximum likelihood estimator (MLE) of θ is the value of θ say $\hat{\theta}$ that maximizes the likelihood function $L = (\theta\,;\, x_1, x_2, \ldots . x_n)$. Note that in many cases, the likelihood function can be infinitesimal and it is much easier to deal with the log-likelihood function that is $log\, L = (\theta\,;\, x_1, x_2, \ldots . x_n)$. Since log is a monotone function, when likelihood function is maximized, log-likelihood function is also maximized, and vice versa.

Definition 1: (Loss Function): Loss function represents the loss incurred when the true value of the parameter is θ and we are estimating θ by $\delta(x)$. Throughout the discussion the loss function $L(\theta, \delta(x))$ is taken as nonnegative and real valued in both its arguments. When the correct estimate is chosen the loss becomes zero. Depending on the loss function Bayes estimators are different. Different types of loss functions are discussed below.

Definition 2: (Linear Loss Function): The linear loss function is defined as

$$L\big(\theta, \delta(x)\big) = c_1\,(\delta(x) - \theta)\,, \delta(x) \geq \theta$$

$$= c_2(\theta - \delta(x))\,, \delta(x) < \theta$$

The constants c_1 and c_2 reflect the effect over and under estimating θ. If c_1 and c_2 are functions of θ, the above loss function is called weighted linear loss function.

Definition 3: (Absolute Error Loss Function): The absolute error loss function is defined as

$$L(\theta, \delta(x)) = |\delta(x) - \theta|.$$

Definition 4: (Squared Error Loss Function): The squared error loss function is defined as

$$L(\theta, \delta(x)) = k(\delta(x) - \theta)^2$$

It is also called as quadratic loss function.

Bayes Estimation:

In Bayesian Principle the unknown parameter θ which is treated as random variable assumes a probability distribution known as a priori of θ denoted by $\Pi(\theta)$.

1. Noninformative Prior: A pdf $\Pi(\theta)$ is said to be a noninformative prior if it contains no information about θ. Some simple examples of noninformative priors are $\Pi(\theta) = 1, \Pi(\theta) = \frac{1}{\theta}$.

Natural conjugate prior: To avoid problem of integration, Statisticians use natural conjugate prior distributions. Usually there is a natural parameter family of distributions such that the posterior distributions also belong to the same family. These priors make the computations much easier. Conjugate priors are usually associated with the exponential family of distributions. Some example of natural conjugate priors are: with sampling from pdf $N(\theta, \sigma^2)$ we take prior distribution $N(\mu, \tau^2)$, the posterior distribution is

$$N\left(\frac{\sigma^2\mu + x\tau^2}{\sigma^2 + \tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}\right)$$

With sampling distribution Binomial and prior distribution Beta the posterior distribution is Beta.

Definition 4: (posterior distribution): The posterior distribution of θ given $X = x$ is obtained by dividing the joint density of θ and X by the marginal distribution of X. Mathematically

$$\frac{\Pi(\theta) f_\theta(x)}{\int_{\Theta} \Pi(\theta) f_\theta(x) d\theta}$$

where Θ is the parameter space.

Definition 5: (Bayes Estimator): A Bayes estimator is that which minimizes the Bayes risk defined above. Accordingly if δ_0 is Bayes estimator of θ with prior distribution $\Pi(\theta)$, then we must have

$$R^*(\theta, \delta_0) = inf\ R^*(\theta, \delta).$$

V. Conclusion

In this paper, Probability Distributions and Estimation of Parameters of Some Continuous Distribution Functions is described. However, a variety of approaches exist for estimating these distributions, drawing on algorithms for estimating probability distributions that are used in computer science and statistics. A probability distribution is a mathematical function that assigns the probabilities of different outcomes to the possible values of a random variable. This paper explains types of probability distributions namely—discrete probability distribution and continuous probability distribution. It also discusses some techniques for estimating parameters of a distribution function such as maximum likelihood, method of moments and the Bayesian approach.

VI. References

[1] Tripathy, M. R.(2009): Estimation of Parameters Under Equality Restrictions, Ph.d. Thesis, IIT Kharagpur

[2] Rohatgi, V. K. and Saleh, A. K. Md. E.(2009): An Introduction to Probability and Statistics, Wiley-interscience, New York

[3] Tripathy, M. R.(2009): Estimation of Parameters Under Equality Restrictions, Ph.d. Thesis, IIT Kharagpur

[4] Sricharan, Kumar, Dennis Wei, and Alfred O Hero (2013). “Ensemble estimators for multivariate entropy estimation”. In: IEEE Transactions on Information Theory 59.7, pp. 4374–4388.

[5] Uppal, Ananya, Shashank Singh, and Barnaás Póczos (2019). “Nonparametric Den sity Estimation under Besov IPM Losses”. In: arXiv preprint arXiv:1902.03511.

[6] Vapnik, Vladimir N and A Ya Chervonenkis (2015). “On the uniform convergence of relative frequencies of events to their probabilities”. In: Measures of complexity. Springer, pp. 11–30.

[7] Weed, Jonathan and Quentin Berthet (2019). “Estimation of smooth densitiesin Wasser stein distance”. In: arXiv preprint arXiv:1902.01778.

[8] Wisler, Alan, Kevin Moon, and Visar Berisha (2017). “Direct ensemble estimation of density functionals”. In: arXiv preprint arXiv:1705.06315

[9] Johnson, N. L., Kotz, S. and Balakrishnan, N.(2004): Continuous Univariate Distributions Vol-1, Wiley-interscience, New York.

[10] Sinha, S. K.(2004): Bayesian Estimation, New Age International, New Delhi.