What is Probability Theory?

K B Athreya

K B Athreya is a retired professor of mathematics and statistics at Iowa State University, Ames, Iowa, in the USA. He spends several months in India visiting schools, colleges and universities. He enjoys teaching mathematics and statistics at all levels. He loves Indian classical and folk music.

This issue of *Resonance* features Joseph Leonard Doob, who played a critical role in the development of probability theory in the world from 1935 onwards. The goal of the present article is to explain to the readers of *Resonance* what probability theory is all about.

Probability theory provides the mathematical basis for the study of random phenomena, that is, phenomena whose outcome is not predictable ahead of time. In this article, we try to provide a more detailed answer.

Introduction

Let us start with an example each of random and non-random (also called deterministic) phenomena:

i) What will be the temperature at 4pm a week from now at the 18th Cross and Margosa Road intersection in Bengaluru?

ii) We throw a stone up from a spot inside an open football ground and observe whether it falls down to the ground or not.

Which one of these can be termed random and which one non-random?

By and large, most physical and natural phenomena can be classified into one or other of these two categories. The readers are invited to construct their own examples of real world phenomena of both kinds (say, two each).

Over the last few centuries mathematical methods have been developed to study many deterministic (i.e. non-

Keywords
Random variables, distribution function, statistical inference, error function, law of large numbers.

random) phenomena, especially those evolving over time. The study of motion of physical objects over time by Newton led to his famous three laws of motion as well as many important developments in the theory of ordinary differential equations.

Similarly, the construction and study of buildings led to important results in geometry in many parts of the world such as India, China, Middle East, Greece. Also advances in quantum mechanics, relativity, etc., were based on deep results from the theory of ordinary and partial differential equations.

Early Beginnings

A mathematical study of random phenomena could be said to have originated in the calculations of odds in some gambling problems in the 18th century Europe. The principal models considered were binomial distributions and their Poisson approximations, and later on normal approximations. For example, if a coin is tossed n times independently (i.e., the outcome of the tosses of any subset of these n tosses has no effect on the outcomes of the remaining tosses) and the probability of 'heads' in any one toss is p, $0 \leq p \leq 1$, then it can be shown that the probability of getting r heads in n tosses is simply $p_{n,r} \equiv \binom{n}{r} p^r (1-p)^{n-r}$ for $r = 0, 1, 2, \cdots, n$, where $\binom{n}{r} = \frac{n!}{r!(n-r)!}$. This collection of $(n+1)$ numbers $\{p_{n,r}, r = 0, 1, 2, \cdots, n\}$ is called the *binomial probability distribution* $B(n,p)$, $0 \leq p \leq 1$, $n = 0, 1, 2, \cdots$. Note that $p_{n,r}$ is non-negative and $\sum_{r=0}^{n} p_{n,r} = (p + (1-p))^n$ by the binomial theorem and hence is equal to 1. Later on, it was shown by Poisson that this quantity $p_{n,r} \equiv \binom{n}{r} p^r (1-p)^{n-r}$ could be approximated for each r, $0 \leq r \leq n$, by $p_r \equiv e^{-\lambda} \frac{\lambda^r}{r!}$ if n is large and p is small but np is neither large nor small but close to some λ, $0 < \lambda < \infty$. This collection $\{p_r \equiv e^{-\lambda} \frac{\lambda^r}{r!},\ r = 0, 1, 2, \cdots\}$ of numbers is called the Poisson λ probability distribution, $0 < \lambda < \infty$. It

A mathematical study of random phenomena could be said to have originated in the calculations of odds in some gambling problems in the 18th century Europe.

may be noted that $p_r \geq 0$ for all $r = 0, 1, 2, \cdots$, and $\sum_{r=0}^{\infty} p_r = 1$.

A bit later, De Moivre and Laplace proved the following. Let n be large and let p_n be not necessarily small but such that $\sigma_n := \sqrt{np_n(1-p_n)}$ converge to a number in $(0, \infty)$. Then, the binomial probabilities can be approximated by a Gaussian distribution. More precisely, the sum of the binomial probabilities $\binom{n}{r} p_n^r (1-p_n)^{n-r}$ over all r in an interval of the form $(a\sigma_n + np_n, b\sigma_n + np_n)$ will converge to $\Phi(b) - \Phi(a)$ where $\Phi(y)$ is equal to the integral of the error function over $(-\infty, y)$. This can be translated into a probability statement:

As $n \to \infty$,

$$\text{Prob}\left(a < \frac{X_n - rp_n}{\sqrt{rp_n(1-p_n)}} < b\right) \to$$

$$\text{Prob}\left(a < Y < b\right) \equiv \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \mathrm{d}x,$$

where X_n is a random variable with the distribution binomial (n, p_n) and Y is a random variable that is normally distributed with mean $EY \equiv 0$ and variance $V(Y) = EY^2 - (EY)^2 = 1$. This is also referred to as an example of the Central Limit Theorem[1] (CLT) and could be thought of as a refinement of the weak law of large numbers which says: for each $\epsilon > 0$,

$$\text{Prob}\left(\left|\frac{X_n}{n} - p_n\right| > \epsilon\right) \to 0$$

as $n \to \infty$ provided $np_n(1-p_n) \to \infty$. Later, both these results were proved for a much larger class of distributions, than just the binomial (n, p_n) cases.

[1] The great mathematician George Pólya coined the term 'central', meaning fundamental. An issue of *Resonance,* Vol.19, No.4, 2014 is devoted to Pólya.

Kolmogorov's Model

A mathematical theory as a basis for studying random phenomena was provided by the great Russian mathematician A N Kolmogorov[2] around 1930. About twenty

[2] *Resonance,* Vol.3, No.4, 1998.

years earlier, Henri Lebesgue of France extended the notion of length of intervals in $\mathbb{R}$, the real line, to a large class $\mathcal{M}$ of sets in $\mathbb{R}$, now called Lebesgue measurable sets. The extended function λ on $\mathcal{M}$ satisfied the condition that $(\mathbb{R}, \mathcal{M}, \lambda)$ is a measure space, i.e., $\mathcal{M}$ is known now as a σ-algebra of subsets of $\mathbb{R}$ that included all intervals and $\lambda : \mathcal{M} \to [0, \infty]$ was such that λ is a measure. (See precise definition later.)

Kolmogorov saw in Lebesgue's theory of measure on $\mathbb{R}$, an appropriate mathematical model for studying random phenomena.

Kolmogorov saw in Lebesgue's theory of measure on $\mathbb{R}$, an appropriate mathematical model for studying random phenomena.

First, one identifies the set Ω of possible outcomes associated with the given random phenomena. This set Ω is called the sample space and a typical individual element ω in Ω called a sample point. Even though the outcome of the experiment is not predictable ahead of time, one may be able to determine the 'chances' that some particular statement about the outcome is valid. The set of ω's for which a given statement is valid is called an event. Thus, an event is a subset of the sample space Ω. After identifying the sample space, one identifies a class $\mathcal{F}$ of subsets of Ω (not necessarily all of $\mathcal{P}(\Omega)$, the power set of Ω, i.e., the collection of all possible subsets of Ω) and then a set function P on $\mathcal{F}$ such that for an event A in $\mathcal{F}$, $P(A)$ will represent the chance of the event A happening. Thus, to a given random phenomenon, one associates a triplet $(\Omega, \mathcal{F}, P)$ where Ω is the set of all possible outcomes (called the sample space), a collection $\mathcal{F} \subset \mathcal{P}(\Omega)$ (called the events collection) and a function P on $\mathcal{F}$ to $[0, \infty]$ (called a probability distribution). It is reasonable to impose the following conditions on $\mathcal{F}$ and P.

i) $A \in \mathcal{F}$ implies $A^c \in \mathcal{F}$ (A^c is the complement of A, i.e., $A^c = \{\omega : \omega \notin A\}$), i.e., if A is an event then A not happening, i.e., A^c should also be an event.

ii) $A_1, A_2 \in \mathcal{F}$ should imply $A_1 \cup A_2 \in \mathcal{F}$, i.e., if A_1 and A_2 are events then at least one of the two events A_1 and A_2 happening should also be an event.

iii) For all A in $\mathcal{F}$, $P(A)$ should be in $[0,1]$ with $P(\Omega) = 1$ and $P(\emptyset) = 0$, where $\emptyset$ is the empty set.

iv) $A_1, A_2 \in \mathcal{F}$, $A_1 \cap A_2 = \emptyset$ should imply

$$P(A_1 \cup A_2) = P(A_1) + P(A_2)$$

i.e., if A_1 and A_2 are mutually exclusive events then the probability of at least one of them happening should simply be the sum of the probabilities of A_1 and A_2.

The above conditions (i−iv) imply that $\mathcal{F}$ is an algebra and P is a finitely additive set function on $\mathcal{F}$, i.e., $\mathcal{F}$ is closed under complementation and finite unions and

$$P\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} P(A_i)$$

for $k < \infty$ and $A_1, A_2, \cdots, A_k \in \mathcal{F}$ and $A_i \cap A_j = \emptyset$ for $i \neq j$.

Next, it is reasonable to require that $\mathcal{F}$ be closed under monotone increasing unions and P be monotone continuous from below. That is, if $\{A_n\}_{n\geq 1}$ is a sequence of events in $\mathcal{F}$ such that $A_n \subset A_{n+1}$ for each $n \geq 1$, then the 'event' $A \equiv \bigcup_{n=1}^{\infty} A_n$ of at least one of the A_n's happening should be in $\mathcal{F}$ and $P(A)$ should equal $\lim P(A_n)$.

This requirement is imposed by the practical idea that if A is a complicated subset of Ω but can be approximated by a sequence $\{A_n\}_{n\geq 1}$ of non-decreasing events such that the above holds then A should be an event and $P(A_n)$ should be close to $P(A)$ for large n. Thus, in addition to conditions (i−iv) on $\mathcal{F}$ and P, it is natural to require the following:

v) $A_n \in \mathcal{F}$, $A_n \subset A_{n+1}$ for all $n = 1, 2, \cdots$ should imply $A \equiv \bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$ and $P(A_n) \uparrow P(A)$ as $n \to \infty$.

Let us call this last condition: $P(\cdot)$ is *monotone continuous from below* (*mcfb*). This last condition (v) looks very natural but along with (i−iv) forces that $(\Omega, \mathcal{F}, P)$ be a measure space, i.e., the following holds:

vi) $\mathcal{F}$ is a σ-algebra (i.e., $\mathcal{F}$ is closed under complementation and countable unions) and $P : \mathcal{F} \to [0, 1]$ is a measure, i.e., P is countably additive, i.e.,

$$P\left(\bigcup_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty} P(B_n)$$

for any $\{B_n\}_{n\geq 1} \subset \mathcal{F}$ such that $B_n \cap B_m = \emptyset$ for $n \neq m$.

Since we demand $P(\Omega) = 1$, $(\Omega, \mathcal{F}, P)$ is called a probability space. Thus, Kolmogorov's model for the study of a random phenomena $\mathcal{E}$ is to determine Ω, the sample space, the sets of all possible outcomes of $\mathcal{E}$, a collection $\mathcal{F}$ of events and a probability set function P mapping $\mathcal{F}$ to $[0, 1]$ so that the triplet $(\Omega, \mathcal{F}, P)$ is a measure space, i.e., the condition (vi) holds with $P(\Omega) = 1$.

Some Examples.

Example 1 (*Finite Sample Space*). Let $\Omega \equiv \{\omega_1, \omega_2, \cdots, \omega_k\}$, $k < \infty$, $\mathcal{F} \equiv \mathcal{P}(\Omega)$, the power set of Ω, i.e., the collection of all possible subsets of Ω (show that there are exactly 2^k of them). Now every probability set function P on $\mathcal{F}$ is necessarily of the form:

$$P(A) = \sum_{i=1}^{k} p_i I_A(\omega_i)\,,$$

where $\{p_i\}_{i=1}^{k}$ are such that $p_i \geq 0$ for all i and $\sum_{i=1}^{k} p_i = 1$ and $I_A(\omega) = 1$ if ω is in A and 0 if ω is not in A.

This is a probability model for random experiments with finitely many possible outcomes. An important example of this is in finite population sampling, used extensively by the National Sample Survey Organization of the Government of India as well as many market research groups.

Let $\{U_1, U_2, \cdots, U_N\}$ be a finite population of N units or objects. These could be individuals in a city, districts in a state, acreage under cultivation of some crops, etc. In a typical sample survey procedure, one chooses a subset of size n (n usually small compared to N) and makes measurements on the chosen subset and uses this data to make inferences about the big population. Here, each sample point is a subset of size n. Thus, the sample space Ω consists of $k = \binom{N}{n} = \frac{N!}{n!(N-n)!}$ sample points and the probabilities p_i of selecting the ith sample are determined by a given sampling scheme. In the so-called simple random sampling without replacement (SRSWOR), each $p_i = \frac{1}{k}$, $i = 1, 2, \cdots, k$, where $k = \binom{N}{n}$. Other examples include coin tossing (finite number of times), roll of dice, card games such as Bridge.

Another important example with a finite sample space is from statistical mechanics in particle physics. Suppose $S \equiv \{s = (i_1, i_2, i_3), i_j \in \{0, 1, -1\}, j = 1, 2, 3\}$ is a set of sites. Note that there are $3 \times 3 \times 3 = 27$ sites in S. Suppose at each site s in S, there is a spin $\omega(s)$ that could be $+1$ or -1. Consider the collection Ω of all spin functions ω mapping S to $\{+1, -1\}$. Then the size of Ω is 2^{27}, a finite, but large number. Call a typical element ω in Ω a configuration. Physicists assign probabilities to any configuration ω by using a parameter β, temperature T, and a function $V(\omega)$, called the potential function. It is of the form

$$p(\omega) = \frac{e^{-\frac{\beta}{T}V(\omega)}}{z_{\beta,T}},$$

where $Z_{\beta,T} \equiv \sum_{\omega' \in \Omega} e^{-\frac{\beta}{T}V(\omega')}$ is called the partition func-

tion. Here $0 < \beta < \infty$, $0 < T < \infty$.

The probability distribution $\{p(\omega) : \omega \in \Omega\}$ is called Gibbs distribution. Computing $\{p(\omega)\}$ is a very challenging task since computing the partition function $Z_{\beta,T}$ is quite difficult. Even more so is computing the mean and variance of some function $g : \Omega \to \mathbb{R}$ with respect to Gibbs distribution. That is, computing λ_1 and $\lambda_2 - \lambda_1^2$ where $\lambda_k = \Sigma_\omega (g(\omega))^k p(\omega)$, k a positive integer. For this, the physicists Metropolis *et al* [2] invented a method in the early 1950's. Statisticians discovered this paper in the early 1990's and coined the term Markov Chain Monte Carlo (MCMC) and since then this subject, i.e., MCMC, has seen some rapid growth. (see [1] Section 9.3, [2])

The physicists Metropolis *et al* [2] invented a method in the early 1950's. Statisticians discovered this paper in the early 1990's and coined the term Markov Chain Monte Carlo (MCMC) and since then this subject, i.e., MCMC, has seen some rapid growth.

Example 2 (*Countably Infinite Sample Space*). Here, $\Omega \equiv \{\omega_1, \omega_2, \cdots\}$ is a countably infinite set, $\mathcal{F} = \mathcal{P}(\Omega)$, the power set of Ω,

$$P(A) = \sum_{i=1}^{\infty} p_i I_A(\omega_i)\,,$$

where $p_i \geq 0$, $\sum_{i=1}^{\infty} p_i = 1$.

An example of this is the experiment of recording the number of radioactive emissions during a given period $[0, T]$ from a specified radioactive source. Here, $\Omega = \{0, 1, 2, \cdots\}$ and $\{p_i\}_{i \geq 0}$ is typically a Poisson distribution of the form

$$p_i = \frac{e^{-\lambda}\lambda^i}{i!}, \quad i = 0, 1, 2, \cdots, 0 < \lambda < \infty.$$

Example 3 (*Real-Valued Random Variables*). Let $\Omega \equiv \mathbb{R}$, $\mathcal{F} \equiv \mathcal{B}(\mathbb{R})$, the Borel σ-algebra in $\mathbb{R}$, i.e., the smallest σ-algebra containing all intervals. (See the definition of a o-algebra given earlier.) Let $F : \mathbb{R} \to [0, 1]$ be a cumulative distribution function (CDF), i.e.,

i) $x_1 \leq x_2 \Rightarrow F(x_1) \leq F(x_2)$,

ii) $\lim_{x\to-\infty} F(x) = 0,$

iii) $\lim_{x\to+\infty} F(x) = 1.$

Then it was shown by Stieltjes [1] that there is a probability measure μ_F on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that

$$\mu_F(a,b] = F(b+) - F(a+) \quad \forall -\infty < a < b < \infty,$$

where $F(x+) \equiv \lim_{y\downarrow x} F(y)$. Let $X : \Omega \to \Omega$ be the identity map, i.e. $X(\omega) = \omega$. This serves as a model for a single real-valued random variable X. We give below a number of examples of F's that are probability distribution functions on $\mathbb{R} = (-\infty, \infty)$.

i) Normal or Gaussian (μ, σ^2): Here,

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{(u-\mu)^2}{2\sigma^2}} \mathrm{d}u;$$
$$-\infty < \mu < \infty, 0 < \sigma < \infty.$$

ii) Gamma (α, p): $0 < \alpha, p < \infty$. Here

$$F(x) = \begin{cases} 0, & x \le 0 \\ \left(\int_0^x e^{-\alpha u} u^{p-1} \mathrm{d}u \right) \frac{1}{\alpha^p \Gamma(p)}, & x > 0 \end{cases}$$

where $\Gamma(p) = \int_0^\infty e^{-u} u^{p-1} \mathrm{d}u.$

iii) Beta (α, β): $0 < \alpha < \infty$, $0 < \beta < \infty$,

$$F(x) = \begin{cases} 0, & x \le 0 \\ \left(\int_0^x y^{\alpha-1}(1-y)^{\beta-1} \mathrm{d}y \right) \frac{1}{B(\alpha,\beta)}, & 0 \le x \le 1 \\ 1, & x > 1 \end{cases}$$

where $B(\alpha, \beta) = \int_0^1 y^{\alpha-1}(1-y)^{\beta-1} \mathrm{d}y.$

iv) Uniform $(0,1)$ is Beta $(1,1)$.

v) Cauchy (γ,σ): $-\infty<\gamma<\infty$, $0<\sigma<\infty$,

$$F(x)=\frac{1}{\pi}\int_{-\infty}^{x}\frac{1}{\sigma}\frac{1}{(\frac{y-\gamma}{\sigma})^2+1}\mathrm{d}y, \quad -\infty<x<\infty.$$

vi) Binomial (n,p): n a positive integer, $[x]=k$, $k\leq x<k+1$,

$$F(x)=\begin{cases}0, & x<0\\ \displaystyle\sum_{r=0}^{[x]}\binom{n}{r}p^r(1-p)^{n-r}, & 0\leq x\leq n\\ 1, & x>n\end{cases}$$

vii) Geometric (p): $0<p<1$

$$F(x)=\begin{cases}0, & x<0\\ (1-p)^{[x]}p, & x>0\end{cases}$$

viii) Poisson (λ):

$$F(x)=\begin{cases}0, & x<0\\ \displaystyle\sum_{r=0}^{[x]}\frac{e^{-\lambda}\lambda^r}{r!}, & x\geq 0\end{cases}$$

Stieltjes's result is more general [1]. It shows that, given a function $F:\mathbb{R}\to\mathbb{R}$ that is non-decreasing, i.e., $x_1\leq x_2\Rightarrow F(x_1)\leq F(x_2)$, there is a measure μ_F defined on the Borel σ-algebra $\mathcal{B}(\mathbb{R})$ of $\mathbb{R}$ such that

$$\mu_F(a,b]=F(b+)-F(a+)\,,$$

where $F(x+)\equiv\lim_{y\downarrow x}F(y)$.

Example 4 (*Random Vectors*). Let $\Omega\equiv\mathbb{R}^k$, $\mathcal{F}\equiv\mathcal{B}(\mathbb{R}^k)$, the Borel σ-algebra in $\mathbb{R}^k$, i.e., the smallest σ-algebra containing all sets of the form

$$(a_1,b_2)\times(a_2,b_2)\times\cdots\times(a_n,b_n),$$

where $a_i < b_i$, $i = 1, 2, \cdots, k$. Let μ be a measure on $\mathcal{B}(\mathbb{R}^k)$ such that $\mu(\mathbb{R}^k) = 1$.

Let $F(\vec{x}) \equiv \mu(-\vec{\infty}, \vec{x}]$

where,

$\vec{x} = (x_1, x_2, \cdots, x_k) \in \mathbb{R}^k, \vec{\infty} = (\infty, \infty, \cdots, \infty)$, and $(-\vec{\infty}, \vec{x}] \equiv \{\vec{y} : \vec{y} = (y_1, y_2, \cdots, y_k), -\infty < y_i \leq x_i, i \leq k\}$.

Then, F is called a k-variate CDF. It satisfies some well-known conditions. Conversely, given such a F, there exists a unique probability measure μ_F on $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$. For details see [1], Section 1.3. The identity map $X(\omega) = \omega$ is a model for the notion of a random vector of k-dimensions.

Example 5 (*Random Sequences*). Let $\Omega \equiv \mathbb{R}^\infty \equiv \{\omega : \omega : \mathbb{N} \to \mathbb{R}\}$, $\mathbb{N} = \{1, 2, 3, \cdots\}$. Let $\forall\ k \in \mathbb{N}$, μ_k be a probability measure on $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}))$ as in Example 4. Suppose $\{\mu_k\}_{k\geq 1}$ satisfies $\mu_{k+1}(A \times \mathbb{R}) = \mu_k(A)$ for all $A \in \mathcal{B}(\mathbb{R}^k)$. Let $\mathcal{F}$ be a σ-algebra generated by the class $\mathcal{C}$ of finite dimensional sets of the form $A \times \mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \cdots$, where $A \in \mathcal{B}(\mathbb{R}^k)$ for some $1 \leq k < \infty$. Then by Kolmogorov's consistency theorem [1], there exists a probability measure μ on $(\mathbb{R}^\infty, \mathcal{F})$ such that for $\forall\ k < \infty$, $A \in \mathcal{B}(\mathbb{R}^k)$, $\mu(A \times \mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \cdots) = \mu_k(A)$. This is a model for a sequence of random variables $\{X_k\}_{k\geq 1}$ such that for every $1 \leq k < \infty$, the probability distribution of $(X_1, X_2, \cdots, X_k)$ under μ will coincide with μ_k.

Example 6 (*Random Functions*). Let T be a nonempty set. For example, T could be a finite set or a countable set or an interval or a subset of some Euclidean space. Let $\Omega \equiv \{f : T \to \mathbb{R}\}$ be the set of all real-valued functions ω on T. Suppose we want to model the choice of an element ω from Ω by a random mechanism. Kolmogorov proved a result known as the consistency theorem to make this precise. Suppose for every finite vector $(t_1, t_2, \cdots, t_k)$, $t < \infty$, of elements from T there

is a probability measure $\mu_{(t_1,t_2,\cdots,t_k)}(\cdot)$ on $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$. Suppose this family of probability measures satisfy:

(i) $\mu_{(t_1,t_2,\cdots,t_k)}(A_1 \times A_2 \times \cdots \times A_k) =$

$$\mu_{(t_{\pi(1)},t_{\pi(2)},\cdots,t_{\pi(k)})}(A_{\pi(1)} \times A_{\pi(2)} \times \cdots \times A_{\pi(k)})$$

for every permutation π of $(1, 2, \cdots, k)$, where $A_1, A_2, \cdots, A_k$ are Borel sets in $\mathbb{R}$.

(ii) $\mu_{(t_1,t_2,\cdots,t_k,t_{k+1})}(A_1 \times A_2 \times \cdots \times A_k \times \mathbb{R}) =$

$$\mu_{(t_1,t_2,\cdots,t_k)}(A_1 \times A_2 \times \cdots \times A_k).$$

Then, there exists a σ-algebra $\mathcal{B}_T$ of subsets of Ω and a probability measure μ_T on $\mathcal{B}_T$ such that for any $(t_1, t_2, \cdots, t_k)$ and $A_1, A_2, \cdots, A_k$ in $\mathcal{B}(\mathbb{R})$, the Borel σ-algebra of $\mathbb{R}$,

$$\mu_T(\omega(t_1) \in A_1, \cdots, \omega(t_k) \in A_k) = \mu_{(t_1,t_2,\cdots,t_k)}(A_1 \times A_2 \times \cdots \times A_k).$$

See [1], Section 1.3 for a proof and further details.

An example of this when $T = [0, \infty)$ is the standard Brownian motion. Here, for every $t_1, t_2, \cdots, t_k \in T = [0, \infty)$, the probability distribution $\mu_{t_1,t_2,\cdots,t_k}(\cdot)$ is that of a k variate normal distribution with mean vector $(0, 0, \cdots, 0)$ and covariance matrix $\sigma_{ij} \equiv \min(t_i, t_j)$ ([1], Section 10.2).

If T is a singleton and $\Omega = \mathbb{R}$, then the random element ω is called a random variable. If T is a finite set the random element ω of $\Omega_T \equiv \mathbb{R}^T$, the set of all functions from Ω to $\mathbb{R}$ is called a random vector. If T is a countable set it is called a random sequence. If T is an interval, it is called a random function. If T is a subset of $\mathbb{R}^k$, it is called a random field. Typically, $\Omega \equiv \mathbb{R}^T$, the collection of all real-valued functions on T is very large. But the σ-algebra $\mathcal{B}_T$ in Kolmogorov's construction is not very large. This makes many interesting quantities $M = \sup\{|\omega(t)| : t \in T\}$ not $\mathcal{B}_T$-measurable, i.e., $\{\omega : M(\omega) \leq a\}$ need not be in $\mathcal{B}_T$ for all a in $\mathbb{R}$. Since the

Kolmogorov construction gives probabilities only for sets in B_T, the probability that $M \leq m$ where m is a given real number, can not be discussed. J L Doob devised a method called separability to take care of this problem [3].

Mean, Variance, Moments of a Random Variable

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Then a function $X : \Omega \to \mathbb{R}$ is called a random variable on $(\Omega, \mathcal{F}, P)$, if sets of the form $\{\omega : X(\omega) \leq a\} \in \mathcal{F}$ for each $a \in \mathbb{R}$ and hence are events and one can talk about the probability distribution of X, i.e., $F_X(a) \equiv P(\omega : X(\omega) \leq a)$. This $F_X(\cdot)$ is called the cumulative distribution function of X. It satisfies:

i) $x_1 \leq x_2 \Rightarrow F_X(x_1) \leq F_X(x_2)$,

ii) $F_X(x+) \equiv \lim_{y \downarrow x} F_X(y) = F_X(x)$ (right continuity), and

iii) $F_X(-\infty) \equiv \lim_{x \downarrow -\infty} F_X(x) = 0$,
$F_X(\infty) \equiv \lim_{x \uparrow \infty} F_X(x) = 1$.

For any Borel set A in $\mathbb{R}$, $P\{\omega : X(\omega) \in A\} \equiv P(X^{-1}(A))$ will coincide with $\mu_{F_X}(A)$, where $\mu_{F_X}(\cdot)$ is the Stieltjes measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ induced by $F_X(\cdot)$.

If X is a simple random variable, i.e., it takes only finitely many distinct real values $\{a_1, a_2, \cdots, a_k\}$ then the expected value EX of X or the mean value of X is defined as

$$EX \equiv \sum_{i=1}^{k} a_i p_i.$$

If X is a non-negative random variable, then X can be approximated by a sequence $\{X_n\}_{n \geq 1}$ of simple random variables such that for each sample point ω in Ω, $X_n(\omega) \geq 0$, $X_n(\omega) \leq X_{n+1}(\omega) \; \forall \; n \geq 1$, $\lim_n X_n(\omega) = X(\omega)$.

Call such a sequence admissible for X. It is natural to define the mean value of X, i.e., EX by setting it equal to $\lim_n EX_n$. It can be shown [1] that $\{EX_n\}_{n\geq 1}$ is a non-decreasing sequence in n and that $\lim_n EX_n$ will be the same for all admissible sequences. It could be $+\infty$. Here the properties that $\mathcal{F}$ is a σ-algebra and P is countably additive are crucially used.

Next, for any real-valued random variable X on $(\Omega, \mathcal{F}, P)$ to $\mathbb{R}$, let

$$X^+(\omega) \equiv \max\{X(\omega), 0\},$$

$$X^-(\omega) \equiv \max\{-X(\omega), 0\}.$$

Then it can be shown that both X^+ and X^- are non-negative random variables on $(\Omega, \mathcal{F}, P)$ and for every ω, $X(\omega) = X^+(\omega) - X^-(\omega)$. So, it is natural to define EX, the expected value of X as $EX = EX^+ - EX^-$ provided at least one of the two quantities EX^+, EX^- is finite. Typically one requires both EX^+ and EX^- to be finite. This renders $E|X| < \infty$. So we see that EX is well defined for any random variable X such that $E|X| < \infty$.

For any random variable X, the kth moment of X for a positive integer k is defined as EX^k provided $E|X^k| < \infty$. The variance of a random variable X is defined as $V(X) \equiv E(X - EX)^2$ provided $EX^2 < \infty$. It can be seen that if $EX^2 < \infty$, then $V(X) = EX^2 - (EX)^2$.

The reader is invited to compute the mean EX and the variance VX for random variables X with probability distributions $F(\cdot)$ mentioned earlier in Example 3.

Laws of Large Numbers and CLT.

There are two results in probability theory that make the subject very useful in applications. This area of application of probability theory to the real world is often termed as the field of statistics. It involves collecting data (i.e., generating random variables) according to

The application of probability theory to the real world (often termed the field of statistics) involves collection of data. That is, it involves generating random variables according to well-defined rules of probability theory and then making inferences about the underlying population based on the data – this is called statistical inference.

well-defined rules of probability theory and then making inferences about the underlying population based on the data (referred to as statistical inference). A fundamental notion needed for these two results is that of the independence of random variables. Let $\mathcal{E}$ be a random experiment, $(\Omega, \mathcal{F}, P)$ be a probability space associated with $\mathcal{E}$ and $X_1, X_2, \cdots, X_k$ be k real-valued random variables $(k < \infty)$ defined on $(\Omega, \mathcal{F}, P)$. Recall that a real-valued random variable X on a probability space $(\Omega, \mathcal{F}, P)$ is simply a function X from Ω to $\mathbb{R}$ such that for each a in $\mathcal{R}$ the set $\{\omega : X(\omega) \leq a\}$ is in $\mathcal{F}$. This is often referred to as X is a measurable function on $(\Omega, \mathcal{F})$ to $\mathbb{R}$. Note that X being measurable depends on $\mathcal{F}$ and not on P. It can also be verified that $X : \Omega \to \mathbb{R}$ is a random variable on $(\Omega, \mathcal{F})$, if and only if, $\{\omega : X(\omega) \in B\} \in \mathcal{F}$ for all B in $\mathcal{B}(\mathbb{R})$, the Borel σ-algebra of $\mathbb{R}$, i.e., the smallest σ-algebra containing intervals of the form (α, β), $\alpha, \beta \in \mathbb{R}$. This property is called X is $(\mathcal{F}, \mathcal{B}(\mathbb{R}))$ measurable [1].

A finite collection $X_1, X_2, \cdots, X_k$, $k < \infty$, of real-valued random variables on a space $(\Omega, \mathcal{F})$ are said to be independent with respect to the probability measure (distribution) P if for any $a_1, a_2, \cdots, a_k$ in $\mathbb{R}$,

$$\begin{aligned} &P\{\omega : X_1(\omega) \leq a_1, X_2(\omega) \leq a_2, \cdots, X_k(\omega) \leq a_k\} \\ &\quad = P\{\omega : X_1(\omega) \leq a_1\} \cdots P\{\omega : X_k(\omega) \leq a_k\}, \end{aligned}$$

i.e., the joint distribution function

$$\begin{aligned} &F_{(X_1, X_2, \cdots, X_k)}(a_1, a_2, \cdots, a_k) \equiv \\ &P\{\omega : X_1(\omega) \leq a_1, X_2(\omega) \leq a_2, \cdots, X_k(\omega) \leq a_k\} \end{aligned}$$

is equal to the product of all the marginal distribution functions, i.e., it equals $\prod_{i=1}^{k} F_{X_i}(a_i)$, where $F_{X_i}(a_i) \equiv P\{\omega : X_i(\omega) \leq a_i\}$.

A family $\{X_t(\omega) : t \in T\}$ of real-valued random variables on a probability space $(\Omega, \mathcal{F}, P)$, where T is an arbitrary index set is said to be independent with respect to P if $\forall$ finite set $\{t_1, t_2, \cdots, t_k\} \subset T$, $k < \infty$,

$\{X_{t_1}(\omega), X_{t_2}(\omega), \cdots, X_{t_k}(\omega)\}$ are independent with respect to P.

An example of an infinite sequence of independent random variables is the following. Let $\Omega = [0,1]$, $\mathcal{F} = \mathcal{B}[0,1]$, the Borel σ-algebra of $[0,1]$, $P =$ Lebesgue measure. For each ω, let $\omega \equiv \sum_{i=1}^{\infty} \frac{\delta_i(\omega)}{2^i}$ be the binary expansion of ω in base 2. Then, it can be shown that for each $k < \infty$, the functions $\delta_1(\omega), \cdots, \delta_k(\omega)$ are independent on this $(\Omega, \mathcal{F}, P)$ with each δ_i having distribution

$$P\{\omega : \delta_i(\omega) = 0\} = \frac{1}{2} = P\{\omega : \delta_i(\omega) = 1\},$$

called the Bernoulli($\frac{1}{2}$) distribution. One just needs to verify that $\{\omega : \delta_1(\omega) = s_1, \cdots, \delta_k(\omega) = s_k\}$ for any given $s_1, \cdots, s_k \in \{0,1\}$ is simply an interval of length $\frac{1}{2^k}$ in $[0,1]$. A similar result holds for expansion to base p, where p is an integer > 1.

The following results known as the 'laws of large numbers' are consequences of slightly more general results due to Kolmogorov.

Theorem 1 (Weak) Law of Large Numbers. *Let $X_1, X_2, X_3, \cdots, X_n$ be independent random variables on some probability space $(\Omega, \mathcal{F}, P)$ such that*

i) they are identically distributed, i.e., $P\{\omega : X_i(\omega) \leq a\} \equiv F(a)$, $a \in \mathbb{R}$ is the same for all $i = 1, 2, \cdots, n$

ii) the mean value of X_1 is well defined, i.e., $E|X_1| < \infty$ (see definition given earlier).

Then, for each $\epsilon > 0$

$$P\{|\overline{X}_n - EX_1| > \epsilon\} \to 0 \qquad \text{as } n \to \infty,$$

where $\overline{X}_n \equiv \frac{X_1 + X_2 + \cdots + X_n}{n}$ (called the sample mean), and EX_1 is as defined earlier.

Theorem 2 (Strong) Law of Large Numbers. *Let* $X_1, X_2, \cdots$ *be a sequence of random variables on some probability space* $(\Omega, \mathcal{F}, P)$ *such that for each* $n < \infty$, $X_1, X_2, \cdots, X_n$ *satisfy the hypothesis of Theorem 1. Then*

$$P\{\omega : \overline{X}_n(\omega) \to EX_1 \text{ as } n \to \infty\} = 1.$$

These two laws of large numbers are what makes the subject of statistics very useful. If the mean value λ of a random variable X is not known it can be estimated from a sample data. More precisely, let $X_1, X_2, \cdots, X_n$ be a sample of n independent copies of X, then by the law of large numbers i.e., Theorem 1, the sample mean $\overline{X_n}$ converges to λ as n tends to infinity. This is called the IID Monte Carlo IIDMC) method.

An example of this is opinion polls in election surveys. Suppose there are two candidates A and B contesting for a position in a city with a large electorate. Suppose the organizers of the candidate A want to estimate the support for A in that city. They choose a small sample of people from that city. Find out the support that A has in that sample. Use that to estimate the support A enjoys in the whole city.

The estimate $\overline{X}_n \equiv \frac{X_1 + X_2 + \cdots + X_n}{n}$ based on n independent observations $X_1, X_2, \cdots, X_n$ of a random variable X is often referred to as a point estimate for the quantity $\lambda \equiv EX$. Another kind of estimate called interval estimate or a confidence interval I_n for $\lambda \equiv EX$ based on the observation $X_1, X_2, \cdots, X_n$ is generated by the use CLT in probability theory (referred to earlier in this article). We give this below.

Central Limit Theorem: Let $X_1, X_2, \cdots, X_n, \cdots$ be independent identically distributed real-valued random variables. Let $EX_1^2 < \infty$. Let $EX_1 = \mu$ and $0 < VX_1 \equiv EX_1^2 - (EX_1)^2 \equiv \sigma^2 < \infty$. Let $\overline{X}_n \equiv \frac{1}{n}\sum_{i=1}^{n} X_i$ for $n = 1, 2, \cdots$. Then, for any $-\infty < a < b < \infty$,

$$\text{i)} \lim_{n\to\infty} P\left(a \le \sqrt{n}\frac{(\overline{X}_n - \mu)}{\sigma} \le b\right) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \mathrm{d}x$$

$$\text{ii)} \lim_{n\to\infty} P\left(a \le \sqrt{n}\frac{(\overline{X}_n - \mu)}{\sigma_n} \le b\right) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \mathrm{d}x$$

where

$$\sigma_n^2 \equiv \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2, n \ge 1.$$

The function $\Phi(y) \equiv \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \mathrm{d}x$, $-\infty < y < \infty$ is called the standard normal distribution function or Gaussian distribution named after the great German mathematician Carl F Gauss[3]. The function $\phi(y) = \frac{d\Phi(y)}{dy} = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}$ is called the standard normal probability density function. The graph of the curve $(x, \phi(x))$ as x varies over $(-\infty, \infty)$ looks like a bell and is referred to as bell curve.

[3] *Resonance*, Vol.2, No.6, 1997.

Suppose given $0 < \alpha < 1$ one wants to produce an interval I_n based on observations $X_1, X_2, \cdots, X_n$ such that

$$P(\mu \in I_n) \to (1-\alpha) \quad \text{as } n \to \infty,$$

where $\mu = EX_1$. For this, one first chooses $a \in (0, \infty)$ such that

$$\Phi(a) - \Phi(-a) = \int_{-a}^{+a} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \mathrm{d}u = (1-\alpha)$$

Note that if

$$I_n \equiv \left[\overline{X}_n - \frac{a\sigma_n}{\sqrt{n}}, \overline{X}_n + \frac{a\sigma_n}{\sqrt{n}}\right]$$

then the CLT (ii) above assures us that

$$P(\mu \in I_n) = P\left(-a \le \frac{(\overline{X}_n - \mu)\sqrt{n}}{\sigma_n} \le b\right) \to (1-\alpha.)$$

This random interval I_n is referred to an approximate confidence interval of level $(1-\alpha)$ for the parameter $\mu = EX_1$. Typically, one chooses α to be 0.05 and the corresponding interval I_n is called a 95% level confidence interval.

It may be noted that CLT is a refinement of the law of large numbers which says that if $E|X_1| < \infty$ then $\overline{X}_n - \mu \to 0$ as $n \to 0$ with probability one. CLT says that if $EX_1^2 < \infty$ while $(\overline{X}_n - \mu) \to 0$, then $(\overline{X}_n - \mu)$ decays at the rate of $\frac{1}{\sqrt{n}}$.

There are similar results when $EX_1^2 = \infty$, but $EX_1 < \infty$. This requires the study of what is called stable distributions [1].

Address for Correspondence
K B Athreya
Department of Mathematics
Iowa State University
Ames, Iowa, USA
Email:
kbathreya@gmail.com

Suggested Reading

[1] **K B Athreya and S N Lahiri, *Measure Theory and Probability Theory*, Springer, New York. (see also TRIM Series Vol.36 and 41, 2006).**

[2] **K B Athreya, M Delampady and T Krishnan, MCMC Methods, *Resonance*, April, July, October, December, 2003.**

[3] **J L Doob, *Stochastic Processes*, John Wiley, New York, 1953.**