

# ON THE DEFINITION OF CONDITIONAL PROBABILITY

D.A.S. Fraser  
University of Toronto, Toronto, Canada

Amir Naderi  
Isfahan University of Technology, Isfahan, Iran.

## **Abstract**

The definition of conditional probability is discussed in detail. The definition is based on the derivative of measures and covers conditioning with respect to events of probability zero. It defines conditional probability as a limit that has physical interpretation and avoids arbitrariness.

Keywords: Probability measures, Derivatives of probability measures, Conditioning.

## 1. Introduction

The notion of the conditional probability of an event given the value of a random variable is basic to the theory of random processes. It is widely used in the statistical literature and is basic to the definition of sufficiency and to Bayesian analysis. Let  $(\Omega, \mathcal{F}, P)$  be a probability space and  $T : (\Omega, \mathcal{F}) \rightarrow (X, \chi)$  be a measurable statistic. If  $A \in \mathcal{F}$ , the customary definition of the conditional probability of  $A$  given  $\{T = t\}$  is  $P(A|T = t) = P(A \cap B)/P(B)$  with  $B = \{T = t\}$ . If  $P(B) = 0$ , this ratio becomes indeterminate and the definition is no longer directly applicable. There are, however, many natural and intuitive ways in which conditioning with respect to events of probability zero can arise. The attempts to solve the difficulty range from an elementary approach using the events  $\{t - h < T < t + h\}$  with  $h \downarrow 0$  or the events  $X \in H_n(t)$  with  $H_n(t) \downarrow \{t\}$  as  $n \rightarrow \infty$  (for example, Parzen, 1960) through to the more advanced treatment involving conditional probability as a Radon-Nikodym derivative (for example, Ash, 1972), or by working with a collection of events with respect to  $\delta$ -fields (for example, Billingsley, 1979). This paper gives a natural and precise definition of the conditional probability  $P(A|T = t)$  even though the event  $\{T = t\}$  may have probability zero for some perhaps all  $t$ . The approach is based on the derivative of measures and defines the conditional probability as a limit that has physical interpretation and thus avoids arbitrariness. The use of a derivative of measures to give a precise definition of probability density and an appropriate definition of likelihood is discussed in Naderi (1988) and Fraser et al (1995).

## 2. Notation And Assumptions

Let  $(\Omega, \mathcal{F}, P)$  be a probability space,  $X$  a metric space,  $\chi$  the smallest  $\sigma$ -field of subsets of  $X$  that contains the open subsets of  $X$  and let  $T : (\Omega, \mathcal{F}) \rightarrow (X, \chi)$  be a measurable statistic. Define  $\mu = PT^{-1}$ . Then since  $\mu$  is finite and  $X$  is a metric space we have that  $\mu$  is regular. Denote the completion of  $(X, \chi, \mu)$  by  $(X, \bar{\chi}, \mu)$  and the completion of  $(\Omega, \mathcal{F}, P)$  by  $(\Omega, \bar{\mathcal{F}}, P)$ . Clearly  $T$  is measurable with respect to  $\bar{\chi}$  and  $\bar{\mathcal{F}}$ . A sequence  $\{E_n\}$  of sets of  $\bar{\chi}$  is said to converge regularly to  $x \in X$  if there is a number  $\alpha > 0$  and a sequence  $\{S_n\}$  of closed spheres of radius  $r_n$  such that  $x \in E_n \subset S_n, \mu(E_n) \geq \alpha\mu(S_n)$ , for all  $n \geq 1$  and  $r_n \rightarrow 0$  as  $n \rightarrow \infty$ . A class  $W \subset \bar{\chi}$  is said to cover a set  $E \subset X$  in the sense of Vitali if for every  $a \in E$ , there is a sequence of sets in  $W$  that converges regularly to  $a$ . A class  $V \subset \bar{\chi}$

is said to possess the Vitali covering property  $[\mu]$  if, for any set  $E \subset X$  and class  $W \subset V$  that covers  $E$  in the sense of Vitali, there is a sequence  $\{E_n\}$  of disjoint sets in  $W$  satisfying  $\mu(E - \cup_{n=1}^{\infty} E_n) = 0$ . Let  $V$  possess the Vitali covering property  $[\mu]$  and let  $Q$  be a finite measure absolutely continuous with respect to  $\mu$ . Let  $Q(E_n)/\mu(E_n) = 0$  if  $\mu(E_n) = 0$ , and define

$$\begin{aligned}\overline{D}(Q, \mu, V)(x) &= \sup\{\limsup_n Q(E_n)/\mu(E_n)\}, \\ \underline{D}(Q, \mu, V)(x) &= \inf\{\liminf_n Q(E_n)/\mu(E_n)\},\end{aligned}$$

where the expression in braces denotes the limit superior (limit inferior) for a sequence  $\{E_n\}$  of sets in  $V$  converging regularly to  $x$ , and the supremum (infimum) is taken among all sequences of sets in  $V$  converging regularly to  $x$ . Then  $Q$  is said to be differentiable at  $x$  with respect to  $V$  and  $\mu$  if

$$\underline{D}(Q, \mu, V)(x) = \overline{D}(Q, \mu, V)(x) < \infty.$$

The common value is called the derivative of  $Q$  at  $x$  with respect to  $V$  and  $\mu$  and is denoted by  $D(Q, \mu, V)(x)$ . It is not difficult to see that  $Q$  is differentiable at  $x$  with respect to  $V$  and  $\mu$  if and only if

$$\lim_{n \rightarrow \infty} Q(E_n)/\mu(E_n) = D(Q, \mu, V)(x)$$

for all sequences  $\{E_n\}$  of sets in  $V$  converging regularly to  $x$ . The following results hold:

**Theorem 1.**  $D(Q, \mu, V)(x)$  exists a.e.  $[\mu]$  and is  $\overline{\chi}$ -measurable and for every  $E \in \overline{\chi}$ ,

$$Q(E) = \int_E D(Q, \mu, V)(x) d\mu(x)$$

**Proof:** A proof may be given by showing that  $\overline{D}(Q, \mu, V)(x)$  and  $\underline{D}(Q, \mu, V)(x)$  are both  $\overline{\chi}$ -measurable and finite, that  $\overline{D}(Q, \mu, V)(x) = \underline{D}(Q, \mu, V)(x)$  giving that  $D(Q, \mu, V)$  exists and is  $\overline{\chi}$ -measurable, and that  $D(Q, \mu, V)$  equals a Radon-Nikodym derivative  $g(x)$  a.e.  $[\mu]$  giving

$$Q(E) = \int_E g(x) d\mu(x) = \int_E D(Q, \mu, V)(x) d\mu(x)$$

For the details see, for example, Naderi (1988)

### 3. The Definition

We begin with the following Theorem:

**Theorem 2.** (i) For each  $A \in \mathcal{F}$  and for all sequences  $\{E_n\}$  of sets  $V$  converging regularly to  $x$ , the  $\lim_{n \rightarrow \infty} P(A \cap T^{-1}(E_n))/P(T^{-1}(E_n))$  exists a.e.  $[\mu]$

(ii) Define  $P(A|T = x)$  equal to the limit in (i) if it exists and equal to zero otherwise. Then  $P(A|T = x)$  is  $\bar{\mathcal{X}}$ -measurable and for each  $E \in \bar{\mathcal{X}}$

$$P(A \cap T^{-1}(E)) = \int_E P(A|T = x)d\mu(x)$$

**Proof:** Define  $P_A(E) = P(A \cap E)$ . Then  $P_A \ll P$ , that is,  $P_A$  is absolutely continuous with respect to  $P$ . Hence  $Q = P_A T^{-1} \ll P T^{-1} = \mu$ . Since  $Q$  is finite then by Theorem 1  $D(Q, \mu, V)(x)$  exists a.e.  $[\mu]$ . But

$$\begin{aligned} D(Q, \mu, V)(x) &= \lim_{n \rightarrow \infty} \frac{Q(E_n)}{\mu(E_n)} = \lim_{n \rightarrow \infty} \frac{P_A T^{-1}(E_n)}{P T^{-1}(E_n)} \\ &= \lim_{n \rightarrow \infty} \frac{P_A(T^{-1}(E_n))}{P(T^{-1}(E_n))} = \lim_{n \rightarrow \infty} \frac{P(A \cap T^{-1}(E_n))}{P(T^{-1}(E_n))} \end{aligned}$$

for all sequences  $\{E_n\}$  of sets in  $V$  converging regularly to  $x$ . The measurability of  $P(A|T = x)$  follows from that of  $D(Q, \mu, V)$ . Also for each  $E \in \bar{\mathcal{X}}$ ,

$$\begin{aligned} P(A \cap T^{-1}(E)) &= P_A T^{-1}(E) = Q(E) = \int_E D(Q, \mu, V)(x)d\mu(x) \\ &= \int_E P(A|T = x)d\mu(x) \cdot \end{aligned}$$

Then  $P(A|T = x)$  defined in Theorem 2 is called the conditional probability of the event  $A$  given  $T$ . If  $\{E_n\}$  shrinks to  $x$  then  $T^{-1}(E_n)$  shrinks to  $\{T = x\}$  and thus the above definition has quite natural properties. The result of Theorem 2 is general and may be used to derive many important and known formulas regarding conditioning with respect to a random variable or a random vector. For example if  $X = R^n$  then the conditional distribution of the random vector  $Y = (Y_1, Y_2, \dots, Y_n) : \Omega \rightarrow R^n$  given the random variable  $Z : \Omega \rightarrow R$  is

$$P(Y \in A|Z = x) = P(Y^{-1}(A)|Z = x), A \in \bar{\mathcal{X}}, x \in R$$

and

$$P(Y \in A, Z \in B) = \int_B P(Y^{-1}(A)|Z = x)d\mu(x)$$

where  $B$  is a one-dimensional Borel set. This then leads to equations defining conditional distribution and probability density functions.

#### 4. References

Ash, Robert B. (1972), Real Analysis and Probability. Academic press, New york.

Billingsley, P. (1979), Probability and Measure. Wiley and sons, New york.

Fraser, D.A.S.; Mc Dunnough, P.; Naderi, A; Plante, A (1995), On the definition of probability densities and sufficiency of the likelihood map; Journal of Probability and Mathematical Statistics. To appear.

Naderi, A. (1988) Measurability and other properties of likelihood map. Ph.D. Thesis, University of Toronto.

Parzen, E. (1960) Modern Probability and its Applications, Wiley and Sons, New York.