Chapter 15
Probability Metrics

The central limit theorem is a very good example of approximating a potentially complicated exact distribution by a simpler and easily computable approximate distribution. In mathematics, whenever we do an approximation, we like to quantify the error of the approximation. Common sense tells us that an error should be measured by some notion of distance between the exact and the approximate. Therefore, when we approximate one probability distribution (measure) by another, we need a notion of distances between probability measures. Fortunately, we have an abundant supply of distances between probability measures. Some of them are for probability measures on the real line, and others for probability measures on a general Euclidean space. Still others work in more general spaces. These distances on probability measures have other independent uses besides quantifying the error of an approximation. We provide a basic treatment of some common distances on probability measures in this chapter. Some of the distances have the so-called *metric property*, and they are called *probability metrics*, whereas some others satisfy only the weaker notion of being a *distance*. Our choice of which metrics and distances to include was necessarily subjective.

The main references for this chapter are Rachev (1991), Reiss (1989), Zolotarev (1983), Leise and Vajda (1987), Dudley (2002), DasGupta (2008), Rao (1987), and Gibbs and Su (2002). Diaconis and Saloff-Coste (2006) illustrate some concrete uses of probability metrics. Additional references are given in the sections.

15.1 Standard Probability Metrics Useful in Statistics

As we said above, there are numerous metrics and distances on probability measures. The choice of the metric depends on the need in a specific situation. No single metric or distance is the best or the most preferable. There is also the very important issue of analytic tractability and ease of computing. Some of the metrics are more easily bound, and some less so. Some of them are hard to compute. Our choice of metrics and distances to cover in this chapter is guided by all these factors, and also personal preferences. The definitions of the metrics and distances are given below. However, we must first precisely draw the distinction between metrics and distances.

A. DasGupta, *Probability for Statistics and Machine Learning: Fundamentals and Advanced Topics*, Springer Texts in Statistics, DOI 10.1007/978-1-4419-9634-3_15,

Definition 15.1. Let $\mathcal{M}$ be a class of probability measures on a sample space Ω. A function $d : \mathcal{M} \bigotimes \mathcal{M} \to \mathcal{R}$ is called a *distance* on $\mathcal{M}$ if

(a) $d(P, Q) \geq 0 \; \forall \; P, Q, \in \mathcal{M}$ and $d(P, Q) = 0 \Leftrightarrow P = Q$;
(b) $d(P_1, P_3) \leq d(P_1, P_2) + d(P_2, P_3) \; \forall \; P_1, P_2, P_3 \in \mathcal{M}$ (Triangular inequality). d is called a *metric* on $\mathcal{M}$ if, moreover.
(c) $d(P, Q) = d(Q, P) \; \forall \; P, Q, \in \mathcal{M}$ (Symmetry).

Here now are the probability metrics and distances that we mention in this chapter.

Definition 15.2 (Kolmogorov Metric). Let P, Q be probability measures on $\mathcal{R}^d$, $d \geq 1$, with corresponding CDFs F, G. The *Kolmogorov metric* is defined as

$$d(P, Q) = \sup_{x \in \mathcal{R}^d} |F(x) - G(x)|.$$

Wasserstein Metric. Let P, Q be probability measures on $\mathcal{R}$ with corresponding CDFs F, G. The *Wasserstein metric* is defined as

$$W(P, Q) = \int_{-\infty}^{\infty} |F(x) - G(x)| dx.$$

Total Variation Metric. Let P, Q be absolutely continuous probability measures on $\mathcal{R}^d, d \geq 1$, with corresponding densities f, g. The *total variation metric* is defined as

$$\rho(P, Q) = \rho(f, g) = \frac{1}{2} \int |f(x) - g(x)| dx.$$

If P, Q are discrete, with corresponding mass functions p, q on the set of values $\{x_1, x_2, \ldots, \}$, then the total variation metric is defined as

$$\rho(P, Q) = \rho(p, q) = \frac{1}{2} \sum_i |p(i) - q(i)|,$$

where $p(i), q(i)$ are the probabilities at x_i under P and Q, respectively.

Separation Distance. Let P, Q be discrete, with corresponding mass functions p, q, The *separation distance* is defined as

$$D(P, Q) = \sup_i \left(1 - \frac{p(i)}{q(i)}\right).$$

Note that the order of P, Q matters in defining $D(P, Q)$.

Hellinger Metric. Let P, Q be absolutely continuous probability measures on $\mathcal{R}^d, d \geq 1$, with corresponding densities f, g. The *Hellinger metric* is defined as

$$H(P, Q) = \left[\int \left(\sqrt{f(x)} - \sqrt{g(x)} \right)^2 dx \right]^{1/2}.$$

If P, Q are discrete, with corresponding mass functions p, q, the Hellinger metric is defined as

$$H(P,Q) = \left[\sum_i \left(\sqrt{p_i} - \sqrt{q_i} \right)^2 \right]^{1/2}.$$

Kullback–Leibler Distance. Let P, Q be absolutely continuous probability measures on $\mathcal{R}^d, d \geq 1$, with corresponding densities f, g. The *Kullback–Leibler distance* is defined as

$$K(P,Q) = K(f,g) = \int f(x) \log \frac{f(x)}{g(x)} dx.$$

If P, Q are discrete, with corresponding mass functions p, q, then the Kullback–Leibler distance is defined as

$$K(P,Q) = K(p,q) = \sum_i p(i) \log \frac{p(i)}{q(i)}.$$

Note that the order of P, Q matters in defining $K(P, Q)$.

Lévy–Prokhorov Metric. Let P, Q be probability measures on $\mathcal{R}^d, d \geq 1$. The *Lévy–Prokhorov metric* is defined as

$$L(P,Q) = \inf\{\epsilon > 0 : \forall\, B \text{ Borel}, \; P(B) \leq Q(B^\epsilon) + \epsilon\},$$

where B^ϵ is, *the outer ϵ-parallel body of B*; that is,

$$B^\epsilon = \{x \in \mathcal{R}^d : \inf_{\{y \in B\}} ||x - y|| \leq \epsilon\}.$$

If $d = 1$, then $L(P, Q)$ equals

$$L(P,Q) = \inf\{\epsilon > 0 : \forall\, x, \; F(x) \leq G(x + \epsilon) + \epsilon\},$$

where F, G are the CDFs of P, Q.

***f*-Divergences.** Let P, Q be absolutely continuous probability measures on $\mathcal{R}^d$, $d \geq 1$, with densities p, q, and f any real-valued convex function on $\mathcal{R}^+$,with $f(1) = 0$. The *f-divergence between P, Q* is defined as

$$d_f(P,Q) = \int q(x) f\left(\frac{p(x)}{q(x)} \right) dx.$$

If P, Q are discrete, with corresponding mass functions p, q, then the f-divergence is defined as

$$d_f(P,Q) = \sum_i q(i) f\left(\frac{p(i)}{q(i)} \right).$$

f-divergences have the *finite partition property* that $d_f(P,Q) = \sup_{\{A_j\}} \sum_j Q(A_j)$ $f\left(\frac{P(A_j)}{Q(A_j)}\right)$, where the supremum is taken over all possible finite partitions $\{A_j\}$ of $\mathcal{R}^d$.