Chapter 2
Metrics, Information Theory, Convergence, and Poisson Approximations

Sometimes it is technically convenient to prove a certain type of convergence by proving that, for some suitable metric d on the set of CDFs, $d(F_n, F) \to 0$ instead of proving the required convergence directly from the definition. Here F_n, F are CDFs on some space, say the real line. Metrics are also useful as statistical tools to assess errors in distribution estimation and to study convergence properties in such statistical problems. The metric, of course, will depend on the type of convergence desired.

The central limit theorem justifiably occupies a prominent place in all of statistics and probability theory. Fourier methods are most commonly used to prove the central limit theorem. This is technically efficient but fails to supply any intuition as to *why* the result should be true. It is interesting that proofs of the central limit theorem have been obtained that avoid Fourier methods and use instead much more intuitive information-theoretic methods. These proofs use convergence of entropies and Fisher information in order to conclude convergence in law to normality. It was then realized that such information-theoretic methods are useful also to establish convergence to Poisson limits in suitable paradigms; for example, convergence of appropriate Bernoulli sums to a Poisson limit. In any case, Poisson approximations are extremely useful in numerous complicated problems in both probability theory and statistics. In this chapter, we give an introduction to the use of metrics and information-theoretic tools for establishing convergences and also give an introduction to Poisson approximations.

Good references on metrics on distributions are Dudley (1989), Rachev (1991), and Reiss (1989). The role of information theory in establishing central limit theorems can be seen, among many references, in Linnik (1959), Brown (1982), and Barron (1986). Poisson approximations have a long history. There are first-generation methods and then there are the modern methods, often called the *Stein-Chen methods*. The literature is huge. A few references are LeCam (1960), Sevas'tyanov (1972), Stein (1972, 1986),

A. DasGupta, *Asymptotic Theory of Statistics and Probability*,

Chen (1975), and Barbour, Holst and Janson (1992). Two other references where interesting applications are given in an easily readable style are Arratia, Goldstein, and Gordon (1990) and Diaconis and Holmes (2004).

2.1 Some Common Metrics and Their Usefulness

There are numerous metrics and distances on probability distributions on Euclidean spaces. The choice depends on the exact purpose and on technical feasibility. We mention a few important ones only and give some information about their interrelationships, primarily in the form of inequalities. The inequalities are good to know in any case.

(i) Metric for convergence in probability
$d_E(X, Y) = E\left(\frac{|X-Y|}{1+|X-Y|}\right)$. This extends to the multidimensional case in the obvious way by using the Euclidean norm $||X - Y||$.
(ii) Kolmogorov metric
$d_K(F, G) = \sup_x |F(x) - G(x)|$. This definition includes the multidimensional case.
(iii) Lévy metric
$d_L(F, G) = \inf\{\epsilon > 0 : F(x - \epsilon) - \epsilon \le G(x) \le F(x + \epsilon) + \epsilon \quad \forall x\}$.
(iv) Total variation metric
$d_{TV}(P, Q) = \sup_{\text{Borel } A} |P(A) - Q(A)|$. This also includes the multidimensional case. If P, Q are both absolutely continuous with respect to some measure μ, then $d_{\text{TV}}(P, Q) = \frac{1}{2}\int |f(x) - g(x)|d\mu(x)$, where f is the density of P with respect to μ and g is the density of Q with respect to μ.
(v) Kullback-Leibler distance
$K(P, Q) = -\int(\log \frac{q}{p})dP = -\int(\log \frac{q}{p})pd\mu$, where $p = \frac{dP}{d\mu}$ and $q = \frac{dQ}{d\mu}$ for some μ. Again, the multidimensional case is included. Note that K is not symmetric in its arguments P, Q.
(vi) Hellinger distance
$H(P, Q) = \left[\int(\sqrt{p} - \sqrt{q})^2 d\mu\right]^{1/2}$, where again $p = \frac{dP}{d\mu}$ and $q = \frac{dQ}{d\mu}$ for some μ, and the multidimensional case is included.

Theorem 2.1

(i) $X_n \xrightarrow{\mathcal{P}} X$ iff $d_E(X_n, X) \to 0$.
(ii) $X_n \xrightarrow{\mathcal{L}} X$ iff $d_L(F_n, F) \to 0$, where $X_n \sim F_n$ and $X \sim F$.

(iii) $X_n \xrightarrow{\mathcal{L}} X$ if $d_K(F_n, F) \to 0$, the reverse being true only under additional conditions.
(iv) If $X \sim F$, where F is continuous and $X_n \sim F_n$, then $X_n \xrightarrow{\mathcal{L}} X$ iff $d_K(F_n, F) \to 0$ (Polyá's theorem).
(v) $X_n \xrightarrow{\mathcal{L}} X$ if $d_{TV}(P_n, P) \to 0$, where $X_n \sim P_n$, $X \sim P$ (the converse is not necessarily true).
(vi) $H(P, Q) \leq \sqrt{K(P, Q)}$.
(vii) $H(P, Q) \geq d_{\mathrm{TV}}(P, Q)$.
(viii) $H(P, Q)/\sqrt{2} \leq \sqrt{d_{\mathrm{TV}}(P, Q)}$.

Proofs of parts of Theorem 2.1 are available in Reiss (1989).

Corollary 2.1 (a) The total variation distance and the Hellinger distance are equivalent in the sense $d_{TV}(P_n, P) \to 0 \Leftrightarrow H(P_n, P) \to 0$.

(b) If P_n, P are all absolutely continuous with unimodal densities, and if P_n converges to P in law, then $H(P_n, P) \to 0$.

(c) Convergence in Kullback-Leibler distance implies convergence in total variation and hence convergence in law.

Note that the proof of part (b) also uses Ibragimov's theorem stated below.

Remark. The Kullback-Leibler distance is very popular in statistics. Specifically, it is frequently used in problems of model selection, testing for goodness of fit, Bayesian modeling and Bayesian asymptotics, and in certain estimation methods known as minimum distance estimation. The Kolmogorov distance is one of the easier ones computationally and has been used in many problems, too, and notably so in the literature on robustness and Bayesian robustness. The Hellinger distance is a popular one in problems of density estimation and in time series problems. The Lévy metric is technically hard to work with but metrizes weak convergence, a very useful property. It, too, has been used in the robustness literature, but it is more common in probability theory. Convergence in total variation is extremely strong, and many statisticians seem to consider it unimportant. But it has a direct connection to $\mathcal{L}_1$ distance, which is intuitive. It has a transformation invariance property and, when it holds, convergence in total variation is extremely comforting.

Notice the last two parts in Theorem 2.1. We have inequalities in *both directions* relating the total variation distance to the Hellinger distance. Since computation of the total variation distance is usually difficult, Hellinger distances are useful in establishing useful bounds on total variation.

2.2 Convergence in Total Variation and Further Useful Formulas

Next, we state three important results on when convergence in total variation can be asserted; see Reiss (1989) for all three theorems and also almost any text on probability for a proof of Scheffé's theorem.

Theorem 2.2 (Scheffé) Let $f_n, n \geq 0$ be a sequence of densities with respect to some measure μ. If $f_n \to f_0$ a.e. (μ), then $d_{\mathrm{TV}}(f_n, f_0) \to 0$.

Remark. Certain converses to Scheffé's theorem are available, and the most recent results are due to Sweeting (1986) and Boos (1985). As we remarked before, convergence in total variation is very strong, and even for the simplest weak convergence problems, convergence in total variation should not be expected without some additional structure. The following theorem exemplifies what kind of structure may be necessary. This is a general theorem (i.e., no assumptions are made on the structural forms of the statistics). In the Theorem 2.4 below, convergence in total variation is considered for sample means of iid random variables (i.e., there is a restriction on the structural form of the underlying statistics). It is not surprising that this theorem needs fewer conditions than Theorem 2.3 to assert convergence in total variation.

Theorem 2.3 (Ibragimov) Suppose P_0 and (for large n) P_n are unimodal, with densities $f_0 = \frac{dP_0}{d\lambda}$ and $f_n = \frac{dP_n}{d\lambda}$, where λ denotes Lebesgue measure. Then $P_n \xrightarrow{\mathcal{L}} P_0$ iff $d_{\mathrm{TV}}(P_n, P_0) \to 0$.

Definition 2.1 A random variable X is said to have a lattice distribution if it is supported on a set of the form $\{a + nh : n \in \mathcal{Z}\}$, where a is a fixed real, h a fixed positive real, and $\mathcal{Z}$ the set of integers.

Theorem 2.4 Suppose $X_1, \ldots, X_n$ are iid nonlattice random variables with a finite variance and characteristic function $\psi(t)$. If, for some $p \geq 1$, $\psi \in L^p(\lambda)$, where λ denotes Lebesgue measure, then $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$ converges to $N(0, 1)$ in total variation.

Example 2.1 Suppose X_n is a sequence of random variables on $[0, 1]$ with density $f_n(x) = 1 + \cos(2\pi n x)$. Then, $X_n \overset{\mathcal{L}}{\Rightarrow} U[0, 1]$ by a direct verification of the definition using CDFs. However, note that the densities f_n *do not* converge to the uniform density 1 as $n \to \infty$. The limit distribution P_0 is unimodal, but the distribution P_n of X_n is not unimodal. The example

shows that the condition in Ibragimov's theorem above that the P_n need to be unimodal as well cannot be relaxed.

Example 2.2 Suppose $X_1, X_2, \ldots$ are iid $\chi^2(2)$ with density $\frac{1}{2}e^{-x/2}$. The characteristic function of X_1 is $\psi(t) = \frac{1}{1-2it}$, which is in $L^p(\lambda)$ for any $p > 1$. Hence, by Theorem 2.4, $\frac{\sqrt{n}(\overline{X}-2)}{2}$ converges in total variation to $N(0, 1)$. We now verify that in fact the density of $Z_n = \frac{\sqrt{n}(\overline{X}-2)}{2}$ converges pointwise to the $N(0, 1)$ density, which by Scheffé's theorem will also imply convergence in total variation. The pointwise convergence of the density is an interesting calculation.

Since $S_n = \sum_{i=1}^n X_i$ has the $\chi^2(2n)$ distribution with density $\frac{e^{-x/2}x^{n-1}}{2^n\Gamma(n)}$, Z_n has density $f_n(z) = \frac{e^{-(z\sqrt{n}+n)}(1+\frac{z}{\sqrt{n}})^{n-1}n^{n-\frac{1}{2}}}{\Gamma(n)}$. Hence, $\log f_n(z) = -z\sqrt{n} - n + (n-1)(\frac{z}{\sqrt{n}} - \frac{z^2}{2n} + O(n^{-3/2})) + (n-\frac{1}{2})\log n - \log\Gamma(n) = -z\sqrt{n} - n + (n-1)(\frac{z}{\sqrt{n}} - \frac{z^2}{2n} + O(n^{-3/2})) + (n-\frac{1}{2})\log n - (n\log n - n - \frac{1}{2}\log n + \log\sqrt{2\pi} + O(n^{-1}))$ on using Stirling's approximation for $\log\Gamma(n)$.

On canceling terms, this gives $\log f_n(z) = -\frac{z}{\sqrt{n}} - \log\sqrt{2\pi} - \frac{(n-1)z^2}{2n} + O(n^{-1/2})$, implying that $\log f_n(z) \to -\log\sqrt{2\pi} - \frac{z^2}{2}$, and hence $f_n(z) \to \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$, establishing the pointwise density convergence.

Example 2.3 The Hellinger and the Kullback-Leibler distances are generally easier to calculate than the total variation distance. The normal case itself is a good example. For instance, the Kullback-Leibler distance $K(N_p(\mu, \mathbf{I}), N_p(0, \mathbf{I})) = \frac{1}{2}||\mu||^2$.

Many bounds on the total variation distance between two multivariate normal distributions are known; we mention a few below that are relatively neat.

$$d_{\mathrm{TV}}(N_p(\mu_1, \mathbf{I}), N_p(\mu_2, \mathbf{I})) \le \frac{1}{\sqrt{2}}||\mu_1 - \mu_2||,$$

$$d_{\mathrm{TV}}(N_p(0, \Sigma), N_p(0, \mathbf{I})) \le \min\left\{\begin{array}{l} \frac{1}{\sqrt{2}}\left(\sum_{i=1}^p(\sigma_i^2 - 1) - \log|\Sigma|\right)^{\frac{1}{2}} \\ p2^{p+1}||\Sigma - \mathbf{I}||_2 \end{array}\right.,$$

where $||A||_2$ denotes the usual Euclidean matrix norm $(\sum_i \sum_j a_{ij}^2)^{1/2}$. These and other bounds can be seen in Reiss (1989).

Example 2.4 Suppose $X_n \sim N(\mu_n, \sigma_n^2)$ and $X_0 \sim N(\mu, \sigma^2)$. Then X_n converges to X_0 in total variation if and only if $\mu_n \to \mu$ and $\sigma_n^2 \to \sigma^2$. This can be proved directly by calculation.

Remark. There is some interest in finding projections in total variation of a fixed distribution to a given class of distributions. This is a good problem but usually very hard, and even in simple one-dimensional cases, the projection can only be found by numerical means. Here is an example; the exercises at the end of the chapter offer some more cases.

Example 2.5 If $X_n \sim \text{Bin}(n, p_n)$ and $np_n \to \lambda, 0 < \lambda < \infty$, then X_n converges in law to the Poi(λ) distribution. In practice, this result is used to approximate a Bin(n, p) distribution for large n and small p by a Poisson distribution with mean np. One can ask what is the best Poisson approximation for a given Bin(n, p) distribution (e.g., what is the total variation projection of a given Bin(n, p) distribution onto the class of all Poisson distributions). An explicit description would not be possible. However, the total variation projection can be numerically computed.

For instance, if $n = 50$, $p = .01$, then the total variation projection is the Poisson distribution with mean .5025. If $n = 100$, $p = .05$, then the total variation projection is the Poisson distribution with mean 5.015. The best Poisson approximation seems to have a mean slightly off from np. In fact, if the total variation projection has mean λ_n, then $|\lambda_n - \lambda| \to 0$. We will come back to Poisson approximations to binomials later in this chapter.

2.3 Information-Theoretic Distances, de Bruijn's Identity, and Relations to Convergence

Entropy and Fisher information are two principal information-theoretic quantities. Statisticians, by means of well-known connections to inference such as the Cramér-Rao inequality and maximum likelihood estimates, are very familiar with the Fisher information. Probabilists, on the other hand, are very familiar with entropy. We first define them formally.

Definition 2.2 Let f be a density in $\mathcal{R}^d$. The entropy of f, or synonymously of a random variable $X \sim f$, is $H(X) = -\int f(x) \log f(x) dx = -E_f[\log f(X)]$.

For integer-valued variables, the definition is similar.

Definition 2.3 Let X be integer valued with $P(X = j) = p_j$. Then, the entropy of X is $H(X) = -\sum_j p(j) \log p(j)$.

Fisher information is defined only for smooth densities. Here is the definition.

Definition 2.4 Let f be a density in $\mathcal{R}^d$. Suppose f has one partial derivative with respect to each coordinate everywhere in its support $\{x : f(x) > 0\}$. The Fisher information of f, or synonymously of a random variable $X \sim f$, is $I(X) = \int_{x:f(x)>0} \frac{||\nabla f(x)||^2}{f(x)} dx = E_f[|| \nabla \log f(X)||^2]$, where $\nabla(.)$ denotes the gradient vector.

Remark. The function $\nabla \log f(x)$ is called *the score function* of f.

Entropy and Fisher information each satisfy certain suitable subadditivity properties. We record their most basic properties below. Johnson (2004) can be consulted for proofs of the theorems in this section apart from the specific references given for particular theorems below.

Theorem 2.5 (a) For jointly distributed random variables X, Y, $H(X, Y) \leq H(X) + H(Y)$ with equality iff X, Y are independent:

(b) For any $\sigma > 0$, $H(\mu + \sigma X) = \log \sigma + H(X)$.

(c) For independent random variables X, Y, $H(X+Y) \geq \max\{H(X), H(Y)\}$.

(d) For jointly distributed random variables X, Y, $I(X, Y) \geq \max\{I(X), I(Y)\}$.

(e) For any σ, $I(\mu + \sigma X) = \frac{I(X)}{\sigma^2}$.

(f) For independent random variables $X, Y, I(X + Y) \leq \alpha^2 I(X) + (1 - \alpha)^2 I(Y) \forall 0 \leq \alpha \leq 1$ with equality iff X, Y are each normal.

(g) For independent random variables $X, Y, I(X + Y) \leq \left(\frac{1}{I(X)} + \frac{1}{I(Y)}\right)^{-1}$ with equality iff X, Y are each normal.

Example 2.6 For some common distributions, we give expressions for the entropy and Fisher information when available.

Distribution	$H(X)$	$I(X)$
Exponential(1)	1	1
$N(0, 1)$	$\frac{1}{2}\log(2\pi) + \frac{1}{2}$	1
Gamma$(\alpha, 1)$	$\alpha + \log\Gamma(\alpha) + (\alpha - 1)\psi(\alpha)$	$\frac{1}{\alpha-2}(\alpha > 2)$
$C(0, 1)$	—	$\frac{1}{2}$
$N_d(0, \Sigma)$	$\frac{d}{2}\log(2\pi) + \log\lvert\Sigma\rvert + \frac{d}{2}$	$\mathrm{tr}\Sigma^{-1}$

Remark. In the table above, ψ is the di-Gamma function (i.e., the derivative of $\log\Gamma$).

Entropy and Fisher information, interestingly, are connected to each other. They are connected by a link to the normal distribution and also through an

algebraic relation known as *de Bruijn's identity*. We mention the link through the normal distribution first.

Theorem 2.6 Among all densities with mean 0 and variance $\sigma^2 < \infty$, the entropy is maximized by the $N(0, \sigma^2)$ density. On the other hand, among all densities with mean 0 and variance $\sigma^2 < \infty$ such that the Fisher information is defined, Fisher information is minimized by the $N(0, \sigma^2)$ density.

Remark. The theorem says that normal distributions are extremals in two optimization problems with a variance constraint, namely the maximum entropy and the minimum Fisher information problems. Actually, although we state the theorem for $N(0, \sigma^2)$, the mean is irrelevant. This theorem establishes an indirect connection between H and I inherited from a connection of each to normal distributions.

We can use H and I to define distances between two different distributions. These are defined as follows.

Definition 2.5 Let $X \sim f, Y \sim g$, and assume that $g(x) = 0 \Rightarrow f(x) = 0$. The *entropy divergence or differential entropy between* f *and* g is defined as

$$D(f||g) = \int f(x) \log\left(\frac{f(x)}{g(x)}\right).$$

The *Fisher information distance between* f *and* g is defined as

$$I(f||g) = I(X||Y) = \int [|| \bigtriangledown \log f - \bigtriangledown \log(g)||^2] f(x) dx.$$

Using the normal distribution as a benchmark, we can define a standardized Fisher information as follows.

Definition 2.6 Let $X \sim f$ have finite variance σ^2. The *standardized Fisher information of* f is defined as $I_s(f) = I_s(X) = \sigma^2 I(f||N(0, \sigma^2))$.

The advantage of the standardization is that $I_s(f)$ can be zero only when f itself is a normal density. Similarly, the entropy divergence of a density f with a normal density can be zero only if f is that same normal density.

We state the elegant algebraic connection between entropy divergence and standardized Fisher information next.

Theorem 2.7 (**De Bruijn's Identity**) Let $X \sim f$ have variance 1. Let Z be a standard normal variable independent of X. For $t > 0$, let f_t denote the density of $X + \sqrt{t}Z$. Then, $I(f_t) = 2\frac{d}{dt}[H(f_t)]$.

Remark. De Bruijn's identity (which extends to higher dimensions) is a consequence of the heat equation of partial differential equations; see Johnson (2004). A large number of such $\frac{d}{dt}$ identities of use in statistics (although not de Bruijn's identity itself) are proved in Brown et al. (2006). That such a neat algebraic identity links entropy with Fisher information is a pleasant surprise.

We now describe how convergence in entropy divergence is a very strong form of convergence.

Theorem 2.8 Let f_n, f be densities in $\mathcal{R}^d$. Suppose $D(f_n||f) \to 0$. Then f_n converges to f in total variation; in particular, convergence in distribution follows.

This theorem has completely general densities f_n, f. In statistics, often one is interested in densities of normalized convolutions. Calculating their entropies or entropy distances from the density of the limiting $N(0, 1)$ distribution could be hard because convolution densities are difficult to write. In a remarkable result, Barron (1986) proved the following.

Theorem 2.9 Let $X_1, X_2, \cdots$ be iid zero-mean, unit-variance random variables and let f_n denote the density (assuming it exists) of $\sqrt{n}\bar{X}$. If, for some m, $D(f_m||N(0, 1)) < \infty$, then $D(f_n||N(0, 1)) \to 0$.

Analogously, one can use Fisher information in order to establish weak convergence. The intuition is that if the Fisher information of $\sqrt{n}\bar{X}$ is converging to 1, which is the Fisher information of the $N(0, 1)$ distribution, then by virtue of the unique Fisher information minimizing property of the $N(0, 1)$ subject to a fixed variance of 1 (stated above), it ought to be the case that $\sqrt{n}\bar{X}$ is converging to $N(0, 1)$ in distribution. The intuition is pushed to a proof in Brown (1982), as stated below.

Theorem 2.10 Let $X_1, X_2, \cdots$ be iid zero-mean, unit-variance random variables and $Z_1, Z_2, \cdots$ be an iid $N(0, 1)$ sequence independent of the $\{X_i\}$. Let $v > 0$ and $Y_n(v) = \sqrt{n}\bar{X} + \sqrt{v}Z_n$. Then, for any v, $I_s(Y_n(v)) \to 0$ and hence $\sqrt{n}\bar{X} \stackrel{\mathcal{L}}{\Rightarrow} N(0, 1)$.

Remark. It had been suspected for a long time that there should be such a proof of the central limit theorem by using Fisher information. It was later found that Brown's technique was so powerful that it extended to central limit theorems for many kinds of non-iid variables. These results amounted to a triumph of information theory tools and provided much more intuitive proofs of the central limit results than proofs based on Fourier methods.

An interesting question to ask is what can be said about the rates of convergence of the entropy divergence and the standardized Fisher information

in the canonical CLT situation (i.e., for $\sqrt{n}\bar{X}$ when X_i are iid with mean 0 and variance 1). This is a difficult question. In general, one can hope for convergence at the rate of $\frac{1}{n}$. The following is true.

Theorem 2.11 Let $X_1, X_2, \cdots$ be iid zero-mean, unit-variance random variables. Then, each of $D(\sqrt{n}\bar{X}||N(0,1))$ and $I_s(\sqrt{n}\bar{X})$ is $O(\frac{1}{n})$.

Remark. This is quite a bit weaker than the best results that are now known. In fact, one can get bounds valid for *all* n, although they involve constants that usually cannot be computed. Johnson and Barron (2003) may be consulted to see the details.

2.4 Poisson Approximations

Exercise 1.5 in Chapter 1 asks to show that the sequence of $\text{Bin}(n, \frac{1}{n})$ distributions converges in law to the Poisson distribution with mean 1. The $\text{Bin}(n, \frac{1}{n})$ is a sum of n independent Bernoullis but with a success probability that is small and also depends on n. The $\text{Bin}(n, \frac{1}{n})$ is a count of the total number of occurrences among n independent rare events. It turns out that convergence to a Poisson distribution can occur even if the individual success probabilities are small but not the same, and even if the Bernoulli variables are not independent. Indeed, approximations by Poisson distributions are extremely useful and accurate in many problems. The problems arise in diverse areas. Poisson approximation is a huge area, with an enormous body of literature, and there are many book-length treatments. We provide here a glimpse into the area with some examples.

Definition 2.7 Let p, q be two mass functions on the integers. The total variation distance between p and q is defined as $d_{\text{TV}}(p,q) = \sup_{A \subseteq \mathcal{Z}} |P_p(X \in A) - P_q(X \in A)|$, which equals $\frac{1}{2}\sum_j |p(j) - q(j)|$.

A simple and classic result is the following.

Theorem 2.12 (**LeCam (1960)**) (a) $d_{\text{TV}}(\text{Bin}(n, \frac{\lambda}{n}), \text{Poi}(\lambda)) \le \frac{8\lambda}{n}$. (b) For $n \ge 1$, let $\{X_{in}\}_{i=1}^n$ be a triangular array of independent $\text{Ber}(p_{in})$ variables. Let $S_n = \sum_{i=1}^n X_{in}$ and $\lambda_n = \sum_{i=1}^n p_{in}$. Then, $d_{\text{TV}}(S_n, \text{Poi}(\lambda_n)) \le \frac{8}{\lambda_n}\sum_{i=1}^n p_{in}^2$, if $\max\{p_{in}, 1 \le i \le n\} \le \frac{1}{4}$.

A neat corollary of LeCam's theorem is the following.

Corollary 2.2 If X_{in} is a triangular array of independent $\text{Ber}(p_{in})$ variables such that $\max\{p_{in}, 1 \le i \le n\} \to 0$, and $\lambda_n = \sum_{i=1}^n p_{in} \to \lambda, 0 < \lambda < \infty$, then $d_{\text{TV}}(S_n, \text{Poi}(\lambda)) \to 0$ and hence $S_n \stackrel{\mathcal{L}}{\Rightarrow} \text{Poi}(\lambda)$.

The Poisson distribution has the property of having equal mean and variance, so intuition would suggest that if a sum of independent Bernoulli variables had, asymptotically, an equal mean and variance, then it should converge to a Poisson distribution. That, too, is true.

Corollary 2.3 If X_{in} is a triangular array of independent Ber(p_{in}) variables such that $\sum_{i=1}^{n} p_{in}$ and $\sum_{i=1}^{n} p_{in}(1 - p_{in})$ each converge to λ, $0 < \lambda < \infty$, then $S_n \stackrel{\mathcal{L}}{\Rightarrow} \text{Poi}(\lambda)$.

It is a fact that, in many applications, although the variable can be represented as a sum of Bernoulli variables, they are not independent. The question arises if a Poisson limit can still be proved. The question is rather old. Techniques that we call *first-generation techniques*, using combinatorial methods, are successful in some interesting problems. These methods typically use generating functions or sharp Bonferroni inequalities. Two very good references for looking at those techniques are Kolchin, Sevas'tyanov, and Chistyakov (1978) and Galambos and Simonelli (1996). Here is perhaps the most basic result of that type.

Theorem 2.13 For $N \geq 1$, let $X_{in}, i = 1, 2, \cdots, n = n(N)$ be a triangular array of Bernoulli random variables, and let $A_i = A_{in}$ denote the event where $X_{in} = 1$. For a given k, let $M_k = M_{kn}$ be the kth binomial moment of S_n; i.e., $M_k = \sum_{j=k}^{n} \binom{j}{k} P(S_n = j)$. If there exists $0 < \lambda < \infty$ such that, for every fixed k, $M_k \to \frac{\lambda^k}{k!}$ as $N \to \infty$, then $S_n \stackrel{\mathcal{L}}{\Rightarrow} \text{Poi}(\lambda)$.

Remark. In some problems, typically of a combinatorial nature, careful counting lets one apply this basic theorem and establish convergence to a Poisson distribution.

Example 2.7 (**The Matching Problem**) Cards are drawn one at a time from a well-shuffled deck containing N cards, and a match occurs if the card bearing a number, say j, is drawn at precisely the jth draw from the deck. Let S_N be the total number of matches. Theorem 2.13 can be used in this example. The binomial moment M_k can be shown to be $M_k = \binom{N}{k} \frac{1}{N(N-1)\ldots(N-k+1)}$, and from here, by Stirling's approximation, for every fixed k, $M_k \to \frac{1}{k!}$, establishing that the total number of matches converges to a Poisson distribution with mean 1 as the deck size $N \to \infty$. Note that the mean value of S_N is exactly 1 for any N. Convergence to a Poisson distribution is extremely fast in this problem; even for $N = 5$, the Poisson approximation is quite good. For $N = 10$, it is almost exact!

For information, we note the following superexponential bound on the error of the Poisson approximation in this problem; this is proved in DasGupta (1999).

Theorem 2.14 $d_{\mathrm{TV}}(S_N, \mathrm{Poi}(1)) \le \frac{2^N}{(N+1)!}\ \forall N$.

Example 2.8 (**The Committee Problem**) From n people, $N = N(n)$ committees are formed, each committee of a fixed size m. We let $N, n \to \infty$, holding m fixed. The Bernoulli variable X_{in} is the indicator of the event that the ith person is not included in any committee. Under the usual assumptions of independence and also the assumption of random selection, the binomial moment M_k can be shown to be $M_k = \binom{n}{k}[\frac{\binom{n-k}{m}}{\binom{n}{m}}]^N$.

Stirling's approximation shows that $M_k \sim \frac{n^k}{k!}e^{-kN(\frac{m}{n}+O(n^{-2}))}$ as $n \to \infty$. One now sees on inspection that if N, n are related as $N = \frac{n\log n}{m} - n\log\lambda + o(n^{-1})$ for some $0 < \lambda < \infty$, then $M_k \to \frac{\lambda^{km}}{k!}$ and so, from the basic convergence theorem above, the number of people who are left out of *all* committees converges to $\mathrm{Poi}(\lambda^m)$.

Example 2.9 (**The Birthday Problem**) This is one of the most colorful examples in probability theory. Suppose each person in a group of n people has, mutually independently, a probability $\frac{1}{N}$ of being born on any given day of a year with N calendar days. Let S_n be the total number of pairs of people (i, j) such that they have the same birthday. $P(S_n > 0)$ is the probability that there is at least one pair of people in the group who share the same birthday. It turns out that if n, N are related as $n^2 = 2N\lambda + o(N)$, for some $0 < \lambda < \infty$, then $S_n \stackrel{\mathcal{L}}{\Rightarrow} \mathrm{Poi}(\lambda)$. For example, if $N = 365, n = 30$, then S_n is roughly Poisson with mean 1.233.

A review of the birthday and matching problems is given in DasGupta (2005). Many of the references given at the beginning of this chapter also discuss Poisson approximation in these problems.

We earlier described the binomial moment method as a first-generation method for establishing Poisson convergence. The modern method, which has been fantastically successful in hard problems, is known as the *Stein-Chen method*. It has a very interesting history. In 1972, Stein gave a novel method of obtaining error bounds in the central limit theorem. Stein (1972) gave a technique that allowed him to have dependent summands and also allowed him to use non-Fourier methods, which are the classical methods in that problem. We go into those results, generally called Berry-Esseen bounds, later in the book (see Chapter 11). Stein's method was based on a very simple identity, now universally known as Stein's iden-

tity (published later in Stein (1981)), which says that if $Z \sim N(0,1)$, then for *nice* functions f, $E[Zf(Z)] = E[f'(Z)]$. It was later found that if Stein's identity holds for *many* nice functions, then the underlying variable Z *must* be $N(0,1)$. So, the intuition is that if for some random variable $Z = Z_n$, $E[Zf(Z) - f'(Z)] \approx 0$, then Z should be close to $N(0,1)$ in distribution. In a manner that many still find mysterious, Stein reduced this to a comparison of the mean of a suitable function h, related to f by a differential equation, under the true distribution of Z and the $N(0,1)$ distribution. From here, he was able to obtain non-Fourier bounds on errors in the CLT for dependent random variables. A Stein type identity was later found for the Poisson case in the decision theory literature; see Hwang (1982). Stein's method for the normal case was successfully adapted to the Poisson case in Chen (1975). The Stein-Chen method is now regarded as the principal tool in establishing Poisson limits for sums of dependent Bernoulli variables. Roughly speaking, the dependence should be weak, and for any single Bernoulli variable, the number of other Bernoulli variables with which it shares a dependence relation should not be very large. The Stein-Chen method has undergone a lot of evolution with increasing sophistication since Chen (1975). The references given in the first section of this chapter contain a wealth of techniques, results, and, most of all, numerous new applications. Specifically, we recommend Arratia, Goldstein, and Gordon (1990), Barbour, Holst and Janson (1992), Dembo and Rinott (1996), and the recent monograph by Diaconis and Holmes (2004). See Barbour, Chen, and Loh (1992) for use of the Stein-Chen technique for compound Poisson approximations.

2.5 Exercises

Exercise 2.1 * Let $X \sim F$ with density $\frac{1}{\pi(1+x^2)}$, $-\infty < x < \infty$. Find the total variation projection of F onto the family of all normal distributions.

Exercise 2.2 For each of the following cases, evaluate the indicated distances.

(i) $d_{\mathrm{TV}}(P, Q)$ when $P = \mathrm{Bin}(20, .05)$ and $Q = \mathrm{Poisson}(1)$.
(ii) $d_K(F, G)$ when $F = N(0, \sigma^2)$ and $G = \mathrm{Cauchy}(0, \tau^2)$.
(iii) $H(P, Q)$ when $P = N(\mu, \sigma^2)$ and $Q = N(\nu, \tau^2)$.

Exercise 2.3 * Write an expansion in powers of ϵ for $d_{\mathrm{TV}}(P, Q)$ when $P = N(0,1)$ and $Q = N(\epsilon, 1)$.

Exercise 2.4 Calculate and plot (as a function of μ) $H(P, Q)$ and $d_{\mathrm{TV}}(P, Q)$ when $P = N(0, 1)$ and $Q = N(\mu, 1)$.

Exercise 2.5 * Suppose $P_n = \mathrm{Bin}(n, p_n)$ and $P = \mathrm{Poi}(\lambda)$. Give a sufficient condition for $d_{\mathrm{TV}}(P_n, P) \to 0$. Can you give a nontrivial necessary condition?

Exercise 2.6 Show that if $X \sim P$, $Y \sim Q$, then $d_{\mathrm{TV}}(P, Q) \leq P(X \neq Y)$.

Exercise 2.7 Suppose $X_i \overset{\text{indep.}}{\sim} P_i$, $Y_i \overset{\text{indep.}}{\sim} Q_i$. Then $d_{\mathrm{TV}}(P_1 * P_2 * \cdots * P_n, Q_1 * Q_2 * \cdots * Q_n) \leq \sum_{i=1}^{n} d_{\mathrm{TV}}(P_i, Q_i)$, where $*$ denotes convolution.

Exercise 2.8 Suppose X_n is a Poisson variable with mean $\frac{n}{n+1}$ and X is Poisson with mean 1.

(a) Show that the total variation distance between the distributions of X_n and X converges to zero.

(b) * (Harder) Find the rate of convergence to zero in part (a).

Exercise 2.9 * Let $P = N(0, 1)$ and $Q = N(\mu, \sigma^2)$. Plot the set $S = \{(\mu, \sigma) : d_{\mathrm{TV}}(P, Q) \leq \epsilon\}$ for some selected values of ϵ.

Exercise 2.10 Suppose $X_1, X_2, \ldots$ are iid Exp(1). Does $\sqrt{n}(\bar{X} - 1)$ converge to standard normal in total variation?

Exercise 2.11 If X_i are iid, show that $\bar{X}_n \xrightarrow{\mathcal{P}} 0$ iff $E\left(\frac{\bar{X}_n^2}{1+\bar{X}_n^2}\right) \to 0$.

Exercise 2.12 * Let $X \sim U[-1, 1]$. Find the total variation projection of X onto the class of all normal distributions.

Exercise 2.13 * Consider the family of densities with mean equal to a specified μ. Find the density in this family that maximizes the entropy.

Exercise 2.14 * (**Projection in Entropy Distance**) Suppose X has a density with mean μ and variance σ^2. Show that the projection of X onto the class of all normal distributions has the same mean and variance as X.

Exercise 2.15 * (**Projection in Entropy Distance Continued**) Suppose X is an integer-valued random variable with mean μ. Show that the projection of X onto the class of all Poisson distributions has the same mean as X.

Exercise 2.16 * First write the exact formula for the entropy of a Poisson distribution, and then prove that the entropy grows at the rate of $\log \lambda$ as the mean $\lambda \to \infty$.

Exercise 2.17 What can you say about the existence of entropy and Fisher information for Beta densities? What about the double exponential density?

Exercise 2.18 Prove that the standardized Fisher information of a Gamma(α, 1) density converges to zero at the rate $\frac{1}{\alpha}$, α being the shape parameter.

Exercise 2.19 * Consider the Le Cam bound $d_{\mathrm{TV}}(\mathrm{Bin}(n,\, p), \mathrm{Poi}(np)) \leq 8p$. Compute the ratio $\frac{d_{\mathrm{TV}}(\mathrm{Bin}(n,p),\mathrm{Poi}(np))}{p}$ for a grid of $(n,\, p)$ pairs and investigate the best constant in Le Cam's inequality.

Exercise 2.20 * For $N = 5, 10, 20, 30$, compute the distribution of the total number of matches in the matching problem, and verify that the distribution in each case is unimodal.

Exercise 2.21 Give an example of a sequence of binomial distributions that converge neither to a normal (on centering and norming) nor to a Poisson distribution.

References

Arratia, R., Goldstein, L., and Gordon, L. (1990). Poisson approximation and the Chen-Stein method, Stat. Sci., 5(4), 403–434.

Barbour, A., Chen, L., and Loh, W-L. (1992). Compound Poisson approximation for non-negative random variables via Stein's method, Ann. Prob., 20, 1843–1866.

Barbour, A., Holst, L., and Janson, S. (1992). *Poisson Approximation*, Clarendon Press, New York.

Barron, A. (1986). Entropy and the central limit theorem, Ann. Prob., 14(1), 336–342.

Boos, D. (1985). A converse to Scheffe's theorem, Ann. Stat., 1, 423–427.

Brown, L. (1982). A proof of the central limit theorem motivated by the Cramér-Rao inequality, G. Kallianpur, P.R. Krishnaiah, and J.K. Ghosh *Statistics and Probability, Essays in Honor of C.R. Rao*, North-Holland, Amsterdam, 141–148.

Brown, L., DasGupta, A., Haff, L.R., and Strawderman, W.E. (2006). The heat equation and Stein's identity: Connections, applications, J. Stat. Planning Infer, Special Issue in Memory of Shanti Gupta, 136, 2254–2278.

Chen, L.H.Y. (1975). Poisson approximation for dependent trials, Ann. Prob., 3, 534–545.

DasGupta, A. (1999). The matching problem and the Poisson approximation, Technical Report, Purdue University.

DasGupta, A. (2005). The matching, birthday, and the strong birthday problems: A contemporary review, J. Stat. Planning Infer, Special Issue in Honor of Herman Chernoff, 130, 377–389.

Dembo, A. and Rinott, Y. (1996). Some examples of normal approximations by Stein's method, in *Random Discrete Structures*, D. Aldous and Pemantle R. IMA Volumes in Mathematics and Its Applications, Vol. 76, Springer, New York, 25–44.

Diaconis, P. and Holmes, S. (2004). *Stein's Method: Expository Lectures and Applications*, IMS Lecture Notes Monograph Series, vol. 46, Institute of Mathematical Statistics, Beachwood, OH.

Dudley, R. (1989). *Real Analysis and Probability*, Wadsworth, Pacific Grove, CA.

Galambos, J. and Simonelli, I. (1996). *Bonferroni-Type Inequalities with Applications*, Springer, New York.

Hwang, J.T. (1982). Improving upon standard estimators in discrete exponential families with applications to Poisson and negative binomial cases, Ann. Stat., 10(3), 857–867.

Johnson, O. (2004). *Information Theory and the Central Limit Theorem*, Imperial College Press, Yale University London.

Johnson, O. and Barron, A. (2003). Fisher information inequalities and the central limit theorem, Technical Report.

Kolchin, V., Sevas'tyanov, B., and Chistyakov, V. (1978). *Random Allocations*, V.H. Winston & Sons, Washington, distributed by Halsted Press, New York.

LeCam, L. (1960). An approximation theorem for the Poisson binomial distribution, Pac. J. Math., 10, 1181–1197.

Linnik, Y. (1959). An information theoretic proof of the central limit theorem, Theory Prob. Appl., 4, 288–299.

Rachev, S. (1991). *Probability Metrics and the Stability of Stochastic Models*, John Wiley, Chichester.

Reiss, R. (1989). *Approximate Distributions of Order Statistics, with Applications to Nonparametric Statistics*, Springer-Verlag, New York.

Sevas'tyanov, B.A. (1972). A limiting Poisson law in a scheme of sums of dependent random variables, Teor. Veroyatni. Primen., 17, 733–738.

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables, L. Le Cam, J. Neyman, and E. Scott in *Proceedings of the Sixth Berkeley Symposium*, Vol. 2, University of California Press, Berkeley, 583–602.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution, Ann. Stat., 9, 1135–1151.

Stein, C. (1986). *Approximate Computations of Expectations*, Institute of Mathematical Statistics, Hayward, CA.

Sweeting, T. (1986). On a converse to Scheffe's theorem, Ann. Stat., 3, 1252–1256.

http://www.springer.com/978-0-387-75970-8

Asymptotic Theory of Statistics and Probability
DasGupta, A.
2008, XXVII, 722 p., Hardcover
ISBN: 978-0-387-75970-8