The central limit theorem and Poisson approximation

An introduction to Stein's method

Fraser Daly (Heriot–Watt University)

December 2013*

Contents

1 The central limit theorem

The central limit theorem is one of the most fundamental results in probability, and explains the appearance of the normal distribution in a whole host of diverse applications in mathematics, physics, biology and the social sciences. Results in this area date back to de Moivre in the 1730s, who used a normal distribution to approximate probabilities associated with binomial random variables.

*Minor corrections made in May 2018

A very readable account of the history of the central limit theorem is given by Le Cam (1986).

The name 'central limit theorem' was applied by Pólya in the 1920s to refer to results concerning sums of independent random variables (suitably scaled) converging to a normal distribution. The name now, however, applies to a much larger class of results concerning convergence in distribution to the normal.

The case of sums of independent random variables is treated by the Lindeberg–Lévy–Feller Theorem, which uses the Lindeberg conditions to give convergence to a normal distribution. Suppose we have independent random variables $X_1, X_2, \ldots$ with $\mathbb{E}X_i = \mu_i$ and $\mathrm{Var}(X_i) = \sigma_i^2 < \infty$ for each i. Let $s_n^2 = \sigma_1^2 + \cdots + \sigma_n^2$. We have the **Lindeberg conditions**

$$\max_k \frac{\sigma_k^2}{s_n^2} \to 0 \qquad \text{as } n \to \infty\,, \tag{1}$$

and

$$L_\varepsilon(n) = \frac{1}{s_n^2} \sum_{k=1}^{n} \mathbb{E}\left[|X_k - \mu_k|^2 I\left\{|X_k - \mu_k| > \varepsilon s_n\right\}\right] \to 0 \qquad \text{as } n \to \infty,\ \forall \varepsilon > 0\,. \tag{2}$$

Theorem 1 (Lindeberg–Lévy–Feller) *With $X_1, X_2, \ldots$ as above*

a). If (2) holds then so does (1) and

$$\frac{1}{s_n} \sum_{k=1}^{n} (X_k - \mu_k) \xrightarrow{d} N(0,1) \qquad \text{as } n \to \infty\,. \tag{3}$$

That is, $F_n(x) \to \Phi(x)$ as $n \to \infty$ for all $x \in \mathbb{R}$, where F_n is the distribution function of $s_n^{-1} \sum_{k=1}^{n} (X_k - \mu_k)$ and Φ is the distribution function of $Z \sim N(0,1)$.

b). If (1) and (3) hold then so does (2).

For the remainder of this section we will assume, without loss of generality, that $\mu_i = 0$ for all i.

Sketch Proof of Sufficiency 1: Characteristic Functions Let $S_n = X_1 + \cdots + X_n$ and for a random variable X let $\psi_X(t) = \mathbb{E}[e^{itX}]$ be the characteristic function of X. We need to show that

$$|\psi_{S_n/s_n}(t) - e^{-t^2/2}| \to 0 \qquad \text{as } n \to \infty\,.$$

Since $X_1, X_2, \ldots$ are independent, $\psi_{S_n/s_n}(t) = \prod_{k=1}^{n} \psi_{X_k}(t/s_n)$. Also,

$$e^{-t^2/2} = \prod_{k=1}^{n} \exp\left\{-\frac{\sigma_k^2 t^2}{2 s_n^2}\right\}.$$

Hence, we need to prove that

$$\left| \prod_{k=1}^{n} \psi_{X_k}(t/s_n) - \prod_{k=1}^{n} \exp\left\{-\frac{\sigma_k^2 t^2}{2 s_n^2}\right\} \right| \to 0 \qquad \text{as } n \to \infty\,.$$

For any $z_1, \ldots, z_n, w_1, \ldots, w_n \in \mathbb{C}$ with $|z_i| \leq 1$ and $|w_i| \leq 1$ for all i, it can be shown that

$$\left| \prod_{k=1}^{n} z_k - \prod_{k=1}^{n} w_k \right| \leq \sum_{k=1}^{n} |z_k - w_k| \,.$$

Hence, it is enough to prove that

$$\sum_{k=1}^{n} \left| \psi_{X_k}(t/s_n) - \exp\left\{ -\frac{\sigma_k^2 t^2}{2s_n^2} \right\} \right| \to 0 \qquad \text{as } n \to \infty \,.$$

We use the triangle inequality, and will show that

$$\sum_{k=1}^{n} \left| \psi_{X_k}(t/s_n) - \left(1 - \frac{\sigma_k^2 t^2}{2s_n^2} \right) \right| \;\to\; 0 \qquad \text{as } n \to \infty \,, \tag{4}$$

$$\sum_{k=1}^{n} \left| \left(1 - \frac{\sigma_k^2 t^2}{2s_n^2} \right) - \exp\left\{ -\frac{\sigma_k^2 t^2}{2s_n^2} \right\} \right| \;\to\; 0 \qquad \text{as } n \to \infty \,. \tag{5}$$

Since (5) is a special case of (4), the case where $X_k \sim \mathrm{N}(0, \sigma_k^2)$ for all k, if we can show (4) we are done.

To prove (4), we use properties of the characteristic function to write

$$\begin{aligned}
\sum_{k=1}^{n} \left| \psi_{X_k}(t/s_n) - \left(1 - \frac{\sigma_k^2 t^2}{2s_n^2} \right) \right| &\leq \sum_{k=1}^{n} \mathbb{E} \min\left\{ \frac{t^2 X_k^2}{s_n^2}, \frac{|t|^3 |X_k|^3}{6 s_n^3} \right\} \\
&\leq \sum_{k=1}^{n} \mathbb{E}\left[\frac{|t|^3 |X_k|^3}{6 s_n^3} I\left\{ |X_k| \leq \varepsilon s_n \right\} \right] \\
&\quad + \sum_{k=1}^{n} \mathbb{E}\left[\frac{t^2 X_k^2}{s_n^2} I\left\{ |X_k| > \varepsilon s_n \right\} \right] \\
&\leq \sum_{k=1}^{n} \frac{|t|^3 \varepsilon s_n}{6 s_n^3} \mathbb{E}\left[|X_k|^2 I\left\{ |X_k| \leq \varepsilon s_n \right\} \right] + t^2 L_\varepsilon(n) \\
&\leq \frac{|t|^3 \varepsilon}{6} + t^2 L_\varepsilon(n) \,,
\end{aligned} \tag{6}$$

for some $\varepsilon > 0$. Since $L_\varepsilon(n) \to 0$ as $n \to \infty$ for all ε, and ε was arbitrary, the result follows. □

Sketch Proof of Sufficiency 2: Replacement We show convergence in distribution by showing that $\mathbb{E}h(S_n/s_n) \to \mathbb{E}h(Z)$ as $n \to \infty$ for all $h : \mathbb{R} \mapsto \mathbb{R}$ with three bounded derivatives.

To do this, introduce a sequence of independent random variables $Y_1, Y_2, \ldots$, where $Y_i \sim \mathrm{N}(0, \sigma_i^2)$ for each i. Write $Z_n = Y_1 + \cdots + Y_n$ so that $Z_n \sim \mathrm{N}(0, s_n^2)$. For $1 \leq j \leq n$, let

$$S_n^{(j)} = Y_1 + \cdots + Y_{j-1} + X_{j+1} + \cdots + X_n \,.$$

Note that $S_n = X_1 + S_n^{(1)}$ and $Z_n = S_n^{(n)} + Y_n$.

We have that

$$\begin{aligned} |\mathbb{E}h(S_n/s_n) - \mathbb{E}h(Z)| &= \left|\mathbb{E}h\left(\frac{S_n}{s_n}\right) - \mathbb{E}h\left(\frac{Z_n}{s_n}\right)\right| \\ &\leq \sum_{j=1}^{n}\left|\mathbb{E}h\left(\frac{S_n^{(j)} + X_j}{s_n}\right) - \mathbb{E}h\left(\frac{S_n^{(j)} + Y_j}{s_n}\right)\right| . \end{aligned}$$

Applying a Taylor expansion (noting that X_j and Y_j are independent and have the same first two moments for each j) we can show that this will go to zero if

$$\sum_{j=1}^{n}\mathbb{E}\min\left\{\left(\frac{X_j}{s_n}\right)^2, \left|\frac{X_j}{s_n}\right|^3\right\} \to 0 \qquad \text{as } n \to \infty\,, \tag{7}$$

$$\sum_{j=1}^{n}\mathbb{E}\min\left\{\left(\frac{Y_j}{s_n}\right)^2, \left|\frac{Y_j}{s_n}\right|^3\right\} \to 0 \qquad \text{as } n \to \infty\,. \tag{8}$$

Given that $Y_1, Y_2, \ldots$ satisfy the Lindeberg condition (2), which may be easily shown, it is enough to prove (7). A similar argument to that used in (6), splitting at εs_n for an arbitrary $\varepsilon > 0$, can be used to show that

$$\sum_{j=1}^{n}\mathbb{E}\min\left\{\left(\frac{X_j}{s_n}\right)^2, \left|\frac{X_j}{s_n}\right|^3\right\} \leq \varepsilon + L_\epsilon(n)\,,$$

from which the result follows. □

The Lindeberg condition in Theorem 1 may be difficult to verify in practice, but may be checked via the following slightly stronger sufficient condition, due to Lyapunov. With the random variables $X_1, X_2, \ldots$ as before, and with the additional assumption that the rth moment of X_i exists for each i and some $r > 2$, we assume that

$$\frac{1}{s_n^r}\sum_{k=1}^{n}\mathbb{E}|X_k - \mu_k|^r \to 0\,,$$

as $n \to \infty$. With this assumption, the Lindeberg condition, and hence the central limit theorem, holds.

There are several techniques for proving normal convergence in situations with dependence. Many of them, however, require the dependence to take a particular form, for example CLTs proved under mixing conditions or in a martingale setting. Stein's method for normal approximation, introduced by Stein (1972), may be applied in many different settings with dependence. Another advantage of Stein's method is that while proving convergence to a normal distribution it automatically gives a rate of convergence to accompany the limit theorem.

In the next part, we will present an argument, due to Stein (1972), that uses Stein's method to prove a central limit theorem in the independent case. Later (in Section 3) we will consider how these techniques may be used in settings with dependence. Before doing that, however, we consider Poisson approximation from both a coupling point of view and using Stein's method.

1.1 Stein's approach to normal approximation: the independent case

For use in this section and elsewhere, we will define the supremum norm of a function $f : \mathbb{R} \mapsto \mathbb{R}$ by $\|f\|_\infty = \sup_x |f(x)|$.

Suppose we have independent random variables $X_1, \ldots, X_n$ with $\mathbb{E}X_i = 0$ and $\mathrm{Var}(X_i) = \sigma_i^2$ for each i. Suppose $\sigma_1^2 + \cdots + \sigma_n^2 = 1$ and let $W = X_1 + \cdots + X_n$. We also write $W_i = W - X_i$.

Stein's approach relies on the observation that a random variable X has a standard normal distribution if and only if

$$\mathbb{E}\left[f'(X) - Xf(X)\right] = 0\,,$$

for all absolutely continuous $f : \mathbb{R} \mapsto \mathbb{R}$. (To see this: in one direction use integration by parts, for the other direction solve a differential equation).

Thus, if $W \approx \mathrm{N}(0,1)$ then $\mathbb{E}\left[f'(W) - Wf(W)\right] \approx 0$. How can we bound this (for a function f with bounded second derivative)?

Firstly, note that $\mathbb{E}[Wf(W)] = \mathbb{E}\sum_{i=1}^n X_i f(W_i + X_i)$. Then, using a Taylor expansion,

$$X_i f(W_i + X_i) = X_i f(W_i) + X_i^2 \int_0^1 f'(W_i + uX_i)\, du\,.$$

By independence, the first term vanishes on taking expectation. Hence

$$\mathbb{E}[Wf(W)] = \mathbb{E}\sum_{i=1}^n X_i^2 \int_0^1 f'(W_i + uX_i)\, du\,.$$

Also,

$$\begin{aligned}
\mathbb{E}[f'(W)] &= \mathbb{E}\sum_{i=1}^n \sigma_i^2 f'(W) \\
&= \mathbb{E}\sum_{i=1}^n \sigma_i^2 f'(W_i) + \mathbb{E}\sum_{i=1}^n \sigma_i^2 \left(f'(W) - f'(W_i)\right) \\
&= \mathbb{E}\sum_{i=1}^n X_i^2 f'(W_i) + \mathbb{E}\sum_{i=1}^n \sigma_i^2 \left(f'(W) - f'(W_i)\right)\,.
\end{aligned}$$

Combining these,

$$\begin{aligned}
&\mathbb{E}[f'(W) - Wf(W)] \\
&\qquad = \mathbb{E}\sum_{i=1}^n X_i^2 \int_0^1 \left(f'(W_i) - f'(W_i + uX_i)\right)\, du + \mathbb{E}\sum_{i=1}^n \sigma_i^2 \left(f'(W) - f'(W_i)\right)\,.
\end{aligned}$$

By the mean value theorem, $|f'(W_i) - f'(W_i + uX_i)| \le |X_i| \|f''\|_\infty$. The same bound may also be applied in the second term of the above (with $u = 1$). Hence

$$|\mathbb{E}[f'(W) - Wf(W)]| \le \|f''\|_\infty \sum_{i=1}^n \left(\mathbb{E}|X_i^3| + \sigma_i^2 \mathbb{E}|X_i|\right) \le 2\|f''\|_\infty \sum_{i=1}^n \mathbb{E}|X_i^3|\,.$$

So, if $\sum_{i=1}^n \mathbb{E}|X_i^3|$ is small, we would expect a CLT to hold. A crucial step in Stein's method is to relate the work we have just done in bounding $|\mathbb{E}[f'(W) - Wf(W)]|$ to an explicit result in normal approximation. Suppose we have an absolutely continuous function $h : \mathbb{R} \mapsto \mathbb{R}$. In assessing a normal approximation for W, we may be interested in bounding $|\mathbb{E}h(W) - \mathbb{E}h(Z)|$, where $Z \sim \mathrm{N}(0,1)$. We can relate this to our characterization of Z through the differential equation

$$h(x) - \mathbb{E}h(Z) = f'(x) - xf(x)\,,$$

whose solution $f = f_h$ depends on h. It can be shown that for h absolutely continuous, $\|f''\|_\infty \leq 2\|h'\|_\infty$. Hence, from the above,

$$|\mathbb{E}h(W) - \mathbb{E}h(Z)| = |\mathbb{E}[f'(W) - Wf(W)]| \leq 2\|f''\|_\infty \sum_{i=1}^n \mathbb{E}|X_i^3| \leq 4\|h'\|_\infty \sum_{i=1}^n \mathbb{E}|X_i^3|\,.$$

2 Poisson approximation

Another of the key limit theorem in probability is that of Poisson convergence (often called "the law of small numbers"). The convergence of the binomial distribution $\mathrm{Bin}(n, \lambda/n)$ to the Poisson distribution $\mathrm{Po}(\lambda)$ as $n \to \infty$ was first established by Poisson in 1837. A famous early statistical application, by von Bortkewitsch, was to the number of deaths of Prussian soldiers resulting from being kicked by a horse. More recently, applications of Poisson limits are found in biology, communications and social sciences.

Throughout this section we will let $X_1, \ldots, X_n$ be (possibly dependent) Bernoulli random variables. We will write $p_i = \mathbb{E}X_i$ and $\lambda = \sum_{i=1}^n p_i$. We are interested in the approximation of $W = X_1 + \cdots + X_n$ by a Poisson random variable $Z \sim \mathrm{Po}(\lambda)$.

For the most part, we will assess closeness of non–negative, integer–valued random variables using the total variation distance:

$$d_{TV}(W, Z) = \frac{1}{2}\sum_{j=0}^{\infty} |\mathbb{P}(W = j) - \mathbb{P}(Z = j)| = \sup_{\|f\|_\infty \leq 1} |\mathbb{E}f(W) - \mathbb{E}f(Z)|\,.$$

2.1 A coupling approach

We begin with a Poisson approximation bound derived using coupling techniques.

Definition 1 *A **coupling** of random variables X and Y is a bivariate random variable $(\widehat{X}, \widehat{Y})$ such that $\widehat{X} \stackrel{d}{=} X$ and $\widehat{Y} \stackrel{d}{=} Y$. Such a coupling is **maximal** if*

$$\mathbb{P}(\widehat{X} = \widehat{Y}) = \sup\left\{\mathbb{P}(\widetilde{X} = \widetilde{Y}) : (\widetilde{X}, \widetilde{Y}) \textit{ is a coupling of } (X, Y)\right\}\,.$$

Before we give a Poisson approximation bound, we note some properties of maximal couplings (which we state without proof).

Lemma 1 *Let $(\widehat{X}, \widehat{Y})$ be a maximal coupling of the non–negative, integer–valued random variables X and Y. Then*

$$\mathbb{P}(\widehat{X} = \widehat{Y}) = \sum_{j=0}^{\infty} \min\{\mathbb{P}(X = j), \mathbb{P}(Y = j)\}\,.$$

Lemma 2 *If $(\widehat{X}, \widehat{Y})$ is a maximal coupling of X and Y*

$$d_{TV}(X, Y) = \mathbb{P}(\widehat{X} \neq \widehat{Y})\,.$$

We are now in a position to state and prove the following well–known Poisson approximation result.

Theorem 2 (Le Cam) *Let $X_1, \ldots, X_n$ be independent Bernoulli random variables, with $\mathbb{E}X_i = p_i$. Let $W = X_1 + \cdots + X_n$ and $\lambda = \mathbb{E}W = p_1 + \cdots + p_n$. If $Z \sim$ Po(λ)*

$$d_{TV}(W, Z) \leq \sum_{i=1}^{n} p_i^2\,.$$

Proof Write $Z = \sum_{i=1}^{n} Z_i$, where $Z_i \sim \text{Po}(p_i)$. We can couple X_i and Z_i maximally for each i (using Lemma 1) to get $(\widehat{X}_i, \widehat{Z}_i)$ with

$$\begin{aligned}\mathbb{P}(\widehat{X}_i = \widehat{Z}_i) &= \sum_{j=0}^{\infty} \min\{\mathbb{P}(X_i = j), \mathbb{P}(Z_i = j)\} \\ &= \min\{1 - p_i, e^{-p_i}\} + \min\{p_i, p_i e^{-p_i}\} = 1 - p_i + p_i e^{-p_i} \geq 1 - p_i^2\,.\end{aligned}$$

Then, since $\left(\sum_{i=1}^{n} \widehat{X}_i, \sum_{i=1}^{n} \widehat{Z}_i\right)$ is a coupling of W and Z,

$$d_{TV}(W, Z) \leq \mathbb{P}\left(\sum_{i=1}^{n} \widehat{X}_i \neq \sum_{i=1}^{n} \widehat{Z}_i\right) \leq \mathbb{P}\left(\bigcup_{i=1}^{n} \left\{\widehat{X}_i \neq \widehat{Z}_i\right\}\right) \leq \sum_{i=1}^{n} \mathbb{P}(\widehat{X}_i \neq \widehat{Z}_i) \leq \sum_{i=1}^{n} p_i^2\,.$$

□

This is an elegant result, but there is much room for improvement. To see this, consider the following results, established by Le Cam (1960) using operator techniques:

$$\begin{aligned} d_{TV}(W, Z) &\leq 4.5 \max_i p_i\,, \\ d_{TV}(W, Z) &\leq 8\lambda^{-1} \sum_{i=1}^{n} p_i^2\,, \end{aligned} \tag{9}$$

this last inequality proved under the assumption $\max_i p_i \leq 1/4$. We are most interested in the second of these inequalities, which can represent a substantial improvement over the bound of Theorem 2 when λ is large, achieved by the inclusion of the "magic factor" of λ^{-1}

Many other techniques have been employed to tackle the Poisson approximation problem of Theorem 2. The vast majority, however, rely on the independence of the summands $X_1, \ldots, X_n$.

2.2 Stein's method for Poisson approximation

Stein's method for Poisson approximation was first developed by Chen (1975). There have been numerous developments since then: see Barbour *et al.* (1992) and Barbour and Chen (2005) for surveys. As in the case of normal approximation, Stein's method has the advantage of being able to handle dependence between the random variables X_i. We will also see that results here include "magic factors" akin to that in Le Cam's result (9) and that were missing in the coupling argument of Theorem 2.

Before proceeding further, we need a definition. For any function $g : \mathbb{Z}^+ \mapsto \mathbb{R}$ we let Δ be the forward difference operator, so that $\Delta g(j) = g(j+1) - g(j)$.

As in the normal case, the starting point of Stein's method is a characterization of the Poisson distribution. We note that for X a non-negative, integer–valued random variable, $X \sim \text{Po}(\lambda)$ if and only if

$$\lambda \mathbb{E}\left[g(X+1)\right] = \mathbb{E}\left[Xg(X)\right], \qquad \forall \text{ bounded } g : \mathbb{Z}^+ \mapsto \mathbb{R}.$$

From this we can define the **characterising operator** A for the $\text{Po}(\lambda)$ distribution: $Ag(j) = \lambda g(j+1) - jg(j)$. So,

$$X \sim \text{Po}(\lambda) \quad \Longleftrightarrow \quad \mathbb{E}\left[Ag(X)\right] = 0 \quad \forall \text{ bounded } g : \mathbb{Z}^+ \mapsto \mathbb{R}.$$

The next step is to solve the **Stein equation**. For a given function $h : \mathbb{Z}^+ \mapsto \mathbb{R}$, we solve

$$h(j) - \mathbb{E}h(Z) = \lambda f(j+1) - jf(j), \tag{10}$$

to find the function $f = f_h$. Replacing j with W and taking expectations we have that

$$\mathbb{E}h(W) - \mathbb{E}h(Z) = \mathbb{E}\left[\lambda f(W+1) - Wf(W)\right].$$

If $W \approx \text{Po}(\lambda)$, then the LHS should be small for a suitably large class of functions h. So, the RHS should also be small. We can 'measure how close W is to Poisson' by looking at how large the RHS can become for h in some suitable class. To make this idea precise (in the case of total variation distance; other metrics may be treated similarly) let

$$\mathcal{H} = \mathcal{H}_{TV} = \{h : \mathbb{Z}^+ \mapsto \mathbb{R} \mid \|h\|_\infty \leq 1\}$$

Then

$$d_{TV}(W, Z) = \sup_{h \in \mathcal{H}} |\mathbb{E}h(W) - \mathbb{E}h(Z)| = \sup_{h \in \mathcal{H}} |\mathbb{E}\left[\lambda f(W+1) - Wf(W)\right]|. \tag{11}$$

It is not clear that bounding the RHS of this is any more straightforward than the problem we started with. However, it turns out that it is.

In order to effectively bound $\mathbb{E}\left[\lambda f(W+1) - Wf(W)\right]$ we will need some properties of the solution $f = f_h$ of the Stein equation for $h \in \mathcal{H}$. The properties that we need are stated (without proof) in Lemma 3 below.

Lemma 3 *Let $f = f_h$ solve the Stein equation (10). Then*

$$\sup_{h\in\mathcal{H}} \|f\|_\infty \leq \min\left\{1, \sqrt{\frac{2}{e\lambda}}\right\}, \qquad \sup_{h\in\mathcal{H}} \|\Delta f\|_\infty \leq \min\left\{1, \frac{1-e^{-\lambda}}{\lambda}\right\}.$$

These bounds are known as **Stein factors**, or magic factors. Note that such bounds depend only on the Stein equation (10) and not on the random variable W of interest.

To see how we may bound (11) in practice, we first consider the case where W is a sum of independent Bernoulli random variables.

2.2.1 Independent summands

In the following theorem, note the improvement over previous results. We retain the magic factor of λ^{-1} appearing in (9), but without the restriction on the p_i.

Theorem 3 *Let $X_1, \ldots, X_n$ be independent Bernoulli random variables, with $\mathbb{E}X_i = p_i$. Let $W = X_1 + \cdots + X_n$ and $\lambda = \mathbb{E}W = p_1 + \cdots + p_n$. If $Z \sim Po(\lambda)$*

$$d_{TV}(W, Z) \leq \left(\frac{1-e^{-\lambda}}{\lambda}\right) \sum_{i=1}^n p_i^2 .$$

Proof For each i we write $W_i = W - X_i$. We begin by noting that

$$\mathbb{E}\left[\lambda f(W+1) - Wf(W)\right] = \sum_{i=1}^n \mathbb{E}\left[p_i f(W+1) - X_i f(W)\right] .$$

For each i, $\mathbb{E}[X_i f(W)] = p_i \mathbb{E}[f(W_i + 1)]$ and so

$$\mathbb{E}\left[\lambda f(W+1) - Wf(W)\right] = \sum_{i=1}^n p_i \mathbb{E}\left[f(W+1) - f(W_i+1)\right] .$$

Since

$$\begin{aligned} \left|\mathbb{E}\left[f(W+1) - f(W_i+1)\right]\right| &\leq \sup_{h\in\mathcal{H}} \|\Delta f\|_\infty \mathbb{E}|W - W_i| \\ &\leq \left(\frac{1-e^{-\lambda}}{\lambda}\right) \mathbb{E}X_i \\ &= \left(\frac{1-e^{-\lambda}}{\lambda}\right) p_i , \end{aligned}$$

(using Lemma 3) we have that

$$\left|\mathbb{E}\left[\lambda f(W+1) - Wf(W)\right]\right| \leq \left(\frac{1-e^{-\lambda}}{\lambda}\right) \sum_{i=1}^n p_i^2 . \qquad \square$$

It is worth noting that the bound given by Theorem 3 is of the right order. We have the corresponding lower bound

$$d_{TV}(W,Z) \geq \frac{1}{32}\min\left\{1,\frac{1}{\lambda}\right\}\sum_{i=1}^{n} p_i^2 .$$

2.2.2 Dependent summands: the local approach

One of the advantages to Stein's approach is that the argument of Theorem 3 may be easily adapted to cover the case where the X_i are no longer independent. In the Poisson case, there are two widely used approaches to doing this: the 'local approach' and the 'coupling approach'. We will spend some time on each, beginning with the local approach.

Theorem 4 *Let $X_1,\ldots,X_n$ be Bernoulli random variables, with $\mathbb{E}X_i = p_i$. Let $W = X_1 + \cdots + X_n$ and $\lambda = \mathbb{E}W = p_1 + \cdots + p_n$. For each i, divide $\{1,\ldots,i-1,i+1,\ldots,n\}$ into two subsets Γ_i and Θ_i so that, informally,*

$$\Gamma_i = \{j : X_j \textit{ is strongly dependent on } X_i\} .$$

Let $Z_i = \sum_{j\in\Gamma_i} X_j$ and $W_i = \sum_{j\in\Theta_i} X_j$. If $Z \sim Po(\lambda)$

$$d_{TV}(W,Z) \leq \left(\frac{1-e^{-\lambda}}{\lambda}\right)\sum_{i=1}^{n}\left(p_i\mathbb{E}[X_i+Z_i] + \mathbb{E}[X_iZ_i]\right) + \sqrt{\frac{2}{e\lambda}}\sum_{i=1}^{n}\mathbb{E}\left|p_i - \mathbb{E}[X_i|W_i]\right| .$$

Proof We write

$$\begin{aligned}
\mathbb{E}\left[\lambda f(W+1) - Wf(W)\right] &= \sum_{i=1}^{n}\mathbb{E}\left[p_i f(W+1) - X_i f(W)\right] \\
&= \sum_{i=1}^{n}\mathbb{E}\left[p_i f(W+1) - p_i f(W_i+1)\right] \\
&\quad + \sum_{i=1}^{n}\mathbb{E}\left[p_i f(W_i+1) - X_i f(W_i+1)\right] \\
&\quad + \sum_{i=1}^{n}\mathbb{E}\left[X_i f(W_i+1) - X_i f(W)\right] .
\end{aligned}$$

For each i we have the bounds

$$\begin{aligned}
|f(W+1) - f(W_i+1)| &\leq \|\Delta f\|_\infty (X_i+Z_i) , \\
|X_i f(W_i+1) - X_i f(W)| &\leq \|\Delta f\|_\infty X_iZ_i , \\
|\mathbb{E}\left[p_i f(W_i+1) - X_i f(W_i+1)\right]| &\leq \|f\|_\infty \mathbb{E}\left|p_i - \mathbb{E}[X_i|W_i]\right| .
\end{aligned}$$

Combining all these we have

$$|\mathbb{E}\left[\lambda f(W+1) - Wf(W)\right]|$$
$$\leq \|\Delta f\|_\infty \sum_{i=1}^{n} \left(p_i \mathbb{E}[X_i + Z_i] + \mathbb{E}[X_i Z_i]\right) + \|f\|_\infty \sum_{i=1}^{n} \mathbb{E}\left|p_i - \mathbb{E}[X_i | W_i]\right| ,$$

from which the result follows using Lemma 3. □

In the case where X_i are independent, we choose $\Gamma_i = \emptyset$ for each i and recover the bound of Theorem 3.

Example (the birthday problem)

Suppose m balls (people) are thrown independently and equiprobably into d boxes (days of the year). Let W be the number of pairs that go into the same box. How close is W to Poisson?

Let Γ be the set of all 2–subsets of $\{1, \ldots, m\}$. That is, $\Gamma = \{i \subset \{1, \ldots, m\} : |i| = 2\}$. If $i = \{i_1, i_2\}$, we write X_i for the indicator that balls i_1 and i_2 land in the same box. So, $W = \sum_{i \in \Gamma} X_i$.

Note that $\mathbb{E}X_i = d^{-1}$ for all $i \in \Gamma$, and so $\lambda = \mathbb{E}W = \binom{m}{2} d^{-1}$. Also, $\mathbb{E}[X_i X_j] = d^{-2}$ for all $i \neq j$.

We choose $\Gamma_i = \{j \in \Gamma \setminus \{i\} : i \cap j \neq \emptyset\}$. Then X_i is independent of X_j for all $j \notin \Gamma_i \cup \{i\}$ and so the final term of the bound in Theorem 4 vanishes.

We obtain

$$\begin{aligned} d_{TV}(W, Z) &\leq \left(\frac{1 - e^{-\lambda}}{\lambda}\right) \sum_{i \in \Gamma} \left(p_i \mathbb{E}[X_i + Z_i] + \mathbb{E}[X_i Z_i]\right) \\ &= \left(\frac{1 - e^{-\lambda}}{\lambda}\right) \binom{m}{2} \left(\frac{2(m-1)+1}{d^2} + \frac{2(m-1)}{d^2}\right) \\ &= \left(\frac{1 - e^{-\lambda}}{\lambda}\right) \binom{m}{2} \frac{4m-3}{d^2} \\ &\leq \frac{8\lambda(1 - e^{-\lambda})}{m-1} . \end{aligned}$$

2.2.3 Size biasing and coupling

As well as the local approach discussed in the previous section, the other common approach to Stein's method for Poisson approximation uses coupling. This approach is discussed in great detail by Barbour *et al.* (1992). As an example of the type of result that can be obtained, we have the following. The proof is very similar to that of Theorem 4.

Theorem 5 *With notation as in Theorem 4, let* $(\widetilde{W_i^1}, W_i^1)$ *be a coupling of* $(W_i|X_i = 1)$ *and* W_i. *Then*

$$d_{TV}(W, Z) \leq \left(\frac{1 - e^{-\lambda}}{\lambda}\right) \sum_{i=1}^{n} \left(p_i \mathbb{E}[X_i + Z_i] + \mathbb{E}[X_i Z_i] + p_i \mathbb{E}\left|W_i^1 - \widetilde{W_i^1}\right|\right) .$$

Coupling results simplify considerably if the random variable W of interest is such that a **monotone coupling** exists. In these cases, computation of the bound in Theorem 5 reduces to knowing the first two moments of W. Such simplifications are a great advantage of the coupling method, as monotone couplings do indeed exist in many interesting examples.

In this section we will need the concept of size–biasing. If W is a non–negative, integer–valued random variable with mean $\lambda > 0$, we let $W^\star$ have the W–size biased distribution, given by

$$\mathbb{P}(W^\star = j) = \frac{j\mathbb{P}(W = j)}{\lambda} .$$

Equivalently, we may define $W^\star$ by letting

$$\lambda \mathbb{E}[g(W^\star)] = \mathbb{E}[W g(W)] , \tag{12}$$

for all $g : \mathbb{Z}^+ \mapsto \mathbb{R}$ for which the expectations above exist.

With this definition, it is clear that we may write

$$\mathbb{E}[Af(W)] = \mathbb{E}\left[\lambda f(W + 1) - W f(W)\right] = \lambda \mathbb{E}\left[f(W + 1) - f(W^\star)\right] .$$

Note that we may also rewrite our characterization of the Poisson distribution by saying that X has a Poisson distribution if and only if $X + 1$ is equal in distribution to $X^\star$.

It can be shown that if

$$\mathbb{E}\left[g(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)|X_i = 1\right] \leq \mathbb{E}\left[g(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)\right] , \tag{13}$$

for all increasing $g\{0, 1\}^{n-1} \mapsto \mathbb{R}$, then there is a coupling $(\widehat{W^{(1)}}, \widehat{W^\star})$ of $W + 1$ and $W^\star$ such that $\widehat{W^{(1)}} \geq \widehat{W^\star}$ almost surely. The property (13) is called **negative relation**.

We may write

$$\begin{aligned} \mathbb{E}\left[f(W + 1) - f(W^\star)\right] &= \sum_{j=0}^{\infty} f(j) \left(\mathbb{P}(W + 1 = j) - \mathbb{P}(W^\star = j)\right) \\ &= \sum_{j=0}^{\infty} \Delta f(j) \left(\mathbb{P}(W + 1 > j) - \mathbb{P}(W^\star > j)\right) . \end{aligned}$$

Then, if $X_1, \ldots, X_n$ are negatively related we may use our monotone coupling (or stochastic ordering) to get that

$$\begin{aligned} |\mathbb{E}[Af(W)]| &\leq \lambda \|\Delta f\|_\infty \sum_{j=0}^{\infty} |\mathbb{P}(W + 1 > j) - \mathbb{P}(W^\star > j)| \\ &= \lambda \|\Delta f\|_\infty \mathbb{E}[W + 1 - W^\star] \\ &= \|\Delta f\|_\infty (\lambda - \text{Var}(W)) . \end{aligned}$$

Hence, using Lemma 3, we have the following.

Theorem 6 *Let $X_1, \ldots, X_n$ be negatively related Bernoulli random variables. Let $W = X_1 + \cdots + X_n$ and $\lambda = \mathbb{E}W$. If $Z \sim Po(\lambda)$*

$$d_{TV}(W, Z) \leq \left(\frac{1 - e^{-\lambda}}{\lambda}\right) [\lambda - Var(W)] \ .$$

We illustrate the applicability of this result with some examples.

- Suppose we distribute m balls uniformly into N urns (each with capacity for up to one ball) and let W count the number of the first n urns which are occupied. Then W has a hypergeometric distribution with

  $$\lambda = \frac{nm}{N}\,, \qquad \text{and} \qquad \text{Var}(W) = \frac{nm(N-n)}{N(N-1)}\left(1 - \frac{m}{N}\right) .$$

  Writing X_i for an indicator that the ith urn is occupied, $W = X_1 + \cdots + X_n$ and $X_1, \ldots, X_n$ are negatively related.

- Distribute n points uniformly on the circumference of a circle. Let $S_1, \ldots, S_n$ be the arc–length distances between adjacent points and $X_i = I(S_i < a)$, the indicator that S_i falls below some threshold a. Then $X_1, \ldots, X_n$ are negatively related and their sum W counts the number of small spacings on our circle.

- Let ν be a permutation of $\{1, \ldots, n\}$ drawn uniformly from the group of such permutations. Let $X_i = I(\nu(i) \leq a_i)$ for some given $a_1, \ldots, a_n$ and $W = X_1 + \cdots + X_n$. We have that $X_1, \ldots, X_n$ are negatively related.

To conclude this section, we note that we say that $X_1, \ldots, X_n$ are **positively related** if (13) holds with the inequality reversed for all increasing $g : \{0, 1\}^{n-1} \mapsto \mathbb{R}$. In this case, there is the following analogue of Theorem 6.

Theorem 7 *Let $X_1, \ldots, X_n$ be positively related Bernoulli random variables with $p_i = \mathbb{E}X_i$. Let $W = X_1 + \cdots + X_n$ and $\lambda = \mathbb{E}W$. If $Z \sim Po(\lambda)$*

$$d_{TV}(W, Z) \leq \left(\frac{1 - e^{-\lambda}}{\lambda}\right) \left[Var(W) - \lambda + 2\sum_{i=1}^{n} p_i^2\right] .$$

This is proved by showing that, under positive relation, there is a coupling $(\widehat{W^{(2)}}, \widehat{W^\star})$ of $W + 1 - X_V$ and $W^\star$ such that $\widehat{W^\star} \geq \widehat{W^{(2)}}$ almost surely, where V is a random index, independent of all else, such that $\mathbb{P}(V = j) = p_j/\lambda$ for $j = 1, \ldots, n$.

3 Stein's method for normal approximation

We have already seen the basics of Stein's method for normal approximation in Section 1.1 where we explored Stein's proof of the central limit theorem (for independent summands) using

the characterising operator A for the standard normal distribution given by $Ag(x) = g'(x) - xg(x)$. We spend this section discussing briefly three approaches to Stein's method for normal approximation in situations with dependence. Much of this discussion is based on the recent survey by Chen *et al.* (2011).

Exchangeable pairs

The approach used in the monograph by Stein (1986) is based on the construction of a random variable W' such that (W, W') is an exchangeable pair and $\mathbb{E}[W'|W] = (1-\eta)W$ for some $\eta \in (0,1)$. The following theorem (given without proof) is typical of the results one can obtain using the exchangeable pairs approach.

Theorem 8 *Let W and W' be mean zero, variance 1 exchangeable random variables satisfying $\mathbb{E}[W'|W] = (1-\eta)W$ for some $\eta \in (0,1)$. Then*

$$\sup_{z\in\mathbb{R}} |\mathbb{P}(W \leq z) - \mathbb{P}(Z \leq z)| \leq \frac{\sqrt{Var(\mathbb{E}[(W'-W)^2|W])}}{2\eta} + (2\pi)^{-1/4}\sqrt{\frac{\mathbb{E}|W'-W|^3}{\eta}},$$

where $Z \sim N(0,1)$.

The starting point of the proof of such results is the observation that $\mathbb{E}g(W, W') = 0$ for every antisymmetric function $g : \mathbb{R}^2 \mapsto \mathbb{R}$.

Size biasing

Size biasing, as defined in the previous section, may also be used to give normal approximation results. Suppose we have a non–negative, integer–valued random variable Y with mean μ and variance σ^2. Let $Y^\star$ be the size–biased version. Letting

$$W = \frac{Y-\mu}{\sigma}, \qquad \text{and} \qquad \widetilde{W} = \frac{Y^\star - \mu}{\sigma},$$

we have that, using (12),

$$\begin{aligned}
\mathbb{E}\left[Af(W)\right] &= \mathbb{E}\left[f'(W) - Wf(W)\right] \\
&= \mathbb{E}\left[f'(W) - \frac{\mu}{\sigma}\left(f(\widetilde{W}) - f(W)\right)\right] \\
&= \mathbb{E}\left[f'(W)\left(1 - \frac{\mu}{\sigma}(\widetilde{W} - W)\right) - \frac{\mu}{\sigma}\int_0^{\widetilde{W}-W} \left(f'(W+t) - f'(W)\right)\, dt\right].
\end{aligned}$$

Bounding these terms (using Stein factors for normal approximation to bound f and its derivatives) gives us a bound on the distance of W from the standard normal distribution. The bounds simplify somewhat if we assume the coupling of Y and $Y^\star$ is bounded. In this case we obtain the following.

Theorem 9 *Let Y and $Y^\star$ be as above and assume that they are coupled such that $|Y - Y^\star| \leq \delta$ almost surely for some $\delta \geq 0$. Then if $Z \sim N(0,1)$*

$$\sup_{z\in\mathbb{R}} |\mathbb{P}(W \leq z) - \mathbb{P}(Z \leq z)| \leq \frac{6\mu\delta^2}{\sigma^3} + \frac{2\mu}{\sigma^2}\sqrt{Var(\mathbb{E}[Y^\star - Y|Y])}\,.$$

To illustrate a typical application, consider the lightbulb process, in which n lightbulbs (which each have two states: on and off) are all switched off at time zero. At time r (for $r = 1, \ldots, n$), exactly r lightbulbs are chosen uniformly at random to have their states changed. The random variable of interest Y is the number of lightbulbs switched on at time n. By coupling Y to its size–biased version it can be shown, for example, that when $n \geq 6$ is even and $\sigma^2 = \mathrm{Var}(Y)$

$$\sup_{z\in\mathbb{R}} |\mathbb{P}(W \leq z) - \mathbb{P}(Z \leq z)| \leq \frac{n}{2\sigma^2}\left(\frac{1}{2\sqrt{n}} + \frac{1}{2n} + e^{-n/2}\right) + \frac{1.64n}{\sigma^3} + \frac{2}{\sigma}\,,$$

where W is again the standardized version of Y and $Z \sim \mathrm{N}(0,1)$.

Zero biasing

Size biasing is just one member of a whole family of such transformations which may be defined. Another transformation which has proved useful in conjunction with Stein's method for normal approximation is zero biasing. If W is a random variable with $\mathbb{E}W = 0$ and $\mathrm{Var}(W) = \sigma^2$, we say that $W^\dagger$ has the W–zero biased distribution if

$$\sigma^2\mathbb{E}[g'(W^\dagger)] = \mathbb{E}[Wg(W)]\,,$$

for all absolutely continuous $g : \mathbb{R} \mapsto \mathbb{R}$ for which the expectations exist. Then it is clear that

$$\mathbb{E}\left[f'(W) - Wf(W)\right] = \mathbb{E}\left[f'(W) - f'(W^\dagger)\right]\,.$$

As with size–biasing, results simplify if we assume a bounded coupling. The following result may then be proved.

Theorem 10 *Let W be a mean zero, variance 1 random variable, coupled to its zero–biased version $W^\dagger$ such that $|W - W^\dagger| \leq \delta$ almost surely. Then*

$$\sup_{z\in\mathbb{R}} |\mathbb{P}(W \leq z) - \mathbb{P}(Z \leq z)| \leq 2.03\delta\,,$$

where $Z \sim N(0,1)$.

To illustrate one result which may be obtained from this, we present without proof a combinatorial central limit theorem.

Theorem 11 *Let $\{a_{i,j}\}_{i,j=1}^n$ be an array of real numbers and let ν be a permutation of $\{1,\ldots,n\}$ chosen uniformly from the group of such permutations. Let $Y = \sum_{i=1}^n a_{i,\nu(i)}$ and $W = (Y-\mu)/\sigma$, where*

$$\mu = \mathbb{E}Y = na_{\bullet,\bullet}\,, \qquad \sigma^2 = Var(Y) = \frac{1}{n-1}\sum_{i,j}\left(a_{i,j} - a_{i,\bullet} - a_{\bullet,j} + a_{\bullet,\bullet}\right)^2\,,$$

and $a_{i,\bullet}$, $a_{\bullet,j}$ and $a_{\bullet,\bullet}$ represent the row, column and array averages, respectively. Then, if $Z \sim N(0,1)$,

$$\sup_{z\in\mathbb{R}} |\mathbb{P}(W \le z) - \mathbb{P}(Z \le z)| \le \frac{16.3}{\sigma}\max_{i,j}|a_{i,j} - a_{i,\bullet} + a_{\bullet,j} - a_{\bullet,\bullet}|\,.$$

Finally, it is also worth noting that the $\mathrm{N}(0,\sigma^2)$ distribution is the unique fixed–point of the zero–biasing transformation.

4 Concluding remarks

Stein's method applies well beyond the normal and Poisson approximation problems we have considered here. It has been employed in approximation problems relating to the binomial, geometric, negative binomial, compound Poisson, discretized normal, exponential, gamma, beta and Fréchet distributions, among others. See Barbour and Chen (2005) and Chen *et al.* (2011) for recent surveys which mention many of these.

In each of these cases, the basic ingredients of Stein's method are the same:

- A characterising operator,
- A solution to the corresponding Stein equation, and
- Stein factors giving bounds on the solution.

Many of the arguments then used to bound the resulting expression have a similar flavour to those we have used here: a coupling or local dependence approach, for example.

Some classes of distribution may be treated more generally. For example, Brown and Xia (2001) consider the problem of approximation by the stationary distribution of a birth–death process on $\mathbb{Z}^+$ with birth rates α_j and death rates β_j for $j \ge 0$. One possible characterising operator in this situation is given by $Ag(j) = \alpha_j g(j+1) - \beta_j g(j)$. Brown and Xia (2001) show that if $\beta_0 = 0$ and $\alpha_j - \alpha_{j-1} \le \beta_j - \beta_{j-1}$ for $j \ge 1$ then the solution of the corresponding Stein equation satisfies

$$\sup_{h\in\mathcal{H}} |f_h(j+1) - f_h(j)| \le \min\left\{\frac{1}{\alpha_j}, \frac{1}{\beta_j}\right\}\,.$$

Stein's method also extends well beyond the univariate setting. Approximation theorems for multivariate distributions (such as the multivariate normal) and processes (such as Poisson and compound Poisson processes) have been developed in this same framework. Again, see the recent surveys by Barbour and Chen (2005) and Chen *et al.* (2011).

One recent and exciting development is the combination of Stein's method with the techniques of Malliavin calculus to give normal approximation theorems. See the recent book by Nourdin and Peccati (2012).

References

A. D. Barbour and L. Y. H. Chen (eds.) (2005). *An Introduction to Stein's Method*. Lecture Notes Series 4, Institute for Mathematical Sciences, Singapore University Press, Singapore.

A. D. Barbour, L. Holst and S. Janson (1992). *Poisson Approximation*. Oxford University Press, Oxford.

T. C. Brown and A. Xia (2001). Stein's method and birth–death processes. *Ann. Probab.* 29: 1373-1403.

L. Le Cam (1960). An approximation theorem for the Poisson binomial distribution. *Pacific J. Math.* 10: 1181–1197.

L. Le Cam (1986). The Central Limit Theorem around 1935. *Statistical Science* 1: 78–96.

L. H. Y. Chen (1975). Poisson approximation for dependent trials. *Ann. Probab.* 3: 534–545.

L. H. Y. Chen, L. Goldstein and Q.–M. Shao (2011). *Normal Approximation by Stein's Method*. Springer, Berlin.

I. Nourdin and G. Peccati (2012) *Normal Approximations with Malliavin Calculus: From Stein's Method to Universality*. Cambridge University Press, Cambridge.

C. Stein (1972). A bound for the error in normal approximation to the distribution of a sum of dependent random variables. *Proc. Sixth Berkeley Symp. Math. Statis. Probab.* 2:583–602.

C. Stein (1986). *Approximate Computation of Expectations*. IMS Lect. Notes Monogr. Ser. 7, Hayward, California.