DEF 23, 15 – 29 (2000)

DECISIONS IN ECONOMICS AND FINANCE

Normal approximations by Stein's method

Yosef Rinott[1], **Vladimir Rotar**[2]

[1] UCSD and Hebrew University
e-mail: rinott@mscc.huji.ac.il
[2] SDSU and Central Economic-Mathematical Institute, Russian Academy of Sciences
e-mail: vrotar@euclid.ucsd.edu

Received: 6 December 1999 / Accepted: 3 February 2000

Abstract. Stein's method for normal approximations is explained, with some examples and applications. In the study of the asymptotic distribution of the sum of dependent random variables, Stein's method may be a very useful tool. We have attempted to write an elementary introduction. For more advanced introductions to Stein's method, see Stein (1986), Barbour (1997) and Chen (1998).

Mathematics Subject Classification (2000): 60F05, 91A06; 62E20, 60J10

Journal of Economic Literature Classification: C00, C70

1. Introduction

Among the many well-known techniques for proving Central Limit Theorems and for studying normal approximations under various conditions, one might mention those that involve characteristic functions, the Lindeberg method which proves the proximity of a sum of random variables to a normal one by replacing the summands by normal variables one at a time, a related operator method, and more. A relatively recent method due to Stein (1972, 1986) is based on a simple differential equation which characterizes the normal distribution, and coupling, i.e., the construction of auxiliary random variables in the probability space of the variables under investigation. The main advantage of this method is that it works well and establishes

The first author was supported in part by NSF grant DMS-9803623. The second author was supported in part by the International Division of NSF grant DMS-9803623, and in part by a grant from the Russian Foundation for Basic Research 98-0100358.

rates, that is, bounds on the distance from normality for certain dependent variables as well.

In this article we present some of the main ideas and techniques of Stein's method for normal approximations; accordingly, proofs of technical lemmas and other details will be omitted. We hope that it will serve as an introduction to Stein's method and its applications.

1.1. An outline of the main ideas

In this section we try to indicate how bounds on the distance from normality are computed by Stein's method, deferring precise statements to later sections. A reader unfamiliar with Stein's method is advised to follow the general theme without getting absorbed in verifying any details at first reading.

We need the following simple lemmas from Stein (1986). The first is essentially proved by straightforward integration by parts, and the second is an elementary result on first order linear differential equations, while (4) requires some standard calculations related to the normal distribution. The proofs are omitted.

Lemma 1.1. *The random variable W has the standard normal distribution if and only if*

$$\mathrm{E}f'(W) = \mathrm{E}Wf(W) \tag{1}$$

for all continuous and piecewise continuously differentiable functions f for which the expectations in (1) exist.

Let Φ denote the standard normal cdf, and let $\Phi h = \mathrm{E}h(Z)$, where Z is standard normal and h is a function for which the expectation exists. Also, for a real valued h, let $||h||$ denote the sup norm, that is,

$$||h|| = \sup_x |h(x)|.$$

Lemma 1.2. *Let h be a bounded piecewise continuously differentiable real valued function. The function*

$$f(w) = e^{w^2/2} \int_{-\infty}^{w} [h(x) - \Phi h] e^{-x^2/2} dx \tag{2}$$

is a solution of the (first order linear) differential equation

$$f'(w) - wf(w) = h(w) - \Phi h, \tag{3}$$

and

$$\text{(a) } ||f|| \le \sqrt{2\pi}||h||, \quad \text{(b) } ||f'|| \le 2||h||, \quad \text{(c) } ||f''|| \le 2||h'||. \tag{4}$$

With these preliminaries we start with an informal description of Stein's approach. Our goal is to study the proximity of the distribution of a random variable W to the standard normal. For this purpose, assuming $\mathrm{E}W = 0$ and $\mathrm{Var}W = 1$, we evaluate, or provide an upper bound on, expressions of the form $|\mathrm{E}h(W) - \Phi h|$, for functions h from a suitable class. (In the classical CLT formulation, one considers h in the class of indicators of half intervals.)

Starting with a given h, and taking f as defined in (2), we have $|\mathrm{E}h(W) - \Phi h| = |\mathrm{E}f'(W) - \mathrm{E}Wf(W)|$, so our goal now is to bound the latter expression. If W were normal, it would vanish by Lemma 1.1. If W is close to normal, Lemma 1.1 suggests that we can hope for $|\mathrm{E}f'(W) - \mathrm{E}Wf(W)|$ to be small for a wide class of functions f, and not just for those that arise as solutions of the particular equation (3).

The next step was called "auxiliary randomization" by Stein. The idea appears under the name of "coupling" in other contexts (see, e.g., Lindvall (1992) and references therein).

We start with the simplest (but not necessarily the most useful) type of coupling; see Goldstein and Reinert (1997) who called it

Zero bias coupling. Suppose that, on the same probability space onwhich W is defined, there exists another random variable W^*, whose marginal distribution is such that

$$\mathrm{E}Wf(W) = \mathrm{E}f'(W^*) \tag{5}$$

for all continuously differentiable f for which the above is well defined. We discuss this variable W^* and its existence shortly, but first observe that now

$$|\mathrm{E}h(W) - \Phi h| = |\mathrm{E}f'(W) - \mathrm{E}Wf(W)| = |\mathrm{E}f'(W) - \mathrm{E}f'(W^*)|. \tag{6}$$

By (4), if h is in a class of functions having a bounded derivative, then f'' is bounded. For such functions h (which do not include indicators of intervals, of course), (6) yields

$$|\mathrm{E}h(W) - \Phi h| \le ||f''|| \cdot \mathrm{E}|W^* - W|. \tag{7}$$

Thus, if we can construct W^* satisfying (5), which is a requirement only on its marginal distribution, and such that $\mathrm{E}|W^* - W|$ is small, a requirement on the joint distribution of W and W^*, we may obtain a bound on $|\mathrm{E}h(W) - \Phi h|$ for certain functions h, i.e., those with a bounded derivative.

In order to understand the construction of W^*, we consider the case of the classical CLT with $W = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i$, where X_i are iid, distributed as a random variable X satisfying $\mathrm{E}X = 0$ and $\mathrm{Var}X = 1$. Given such a r.v. X, let X^* denote a random variable satisfying $\mathrm{E}Xf(X) = \mathrm{E}f'(X^*)$ for any continuously differentiable f. Let X_n^* be independent of all other variables and distributed like X^*, and define $W^* = (S_{n-1} + X_n^*)/\sqrt{n}$, where

$S_n = \sum_{i=1}^n X_i$. We claim that (5) holds. Indeed $\mathrm{E}f'(W^*) = \mathrm{E}f'((S_{n-1} + X_n^*)/\sqrt{n}) = \mathrm{E}\sqrt{n}X_n f((S_{n-1} + X_n)/\sqrt{n}) = \mathrm{E}n\frac{X_n}{\sqrt{n}} f(W) = \mathrm{E}Wf(W)$, where in the second equality we applied the relation defining X^* with the function $\sqrt{n}f((S_{n-1} + \cdot)/\sqrt{n})$, fixing (or conditioning on) S_{n-1}, and in the last equality we used the obvious symmetry. Informally stated: we can change a normalized sum W to W^* by changing an individual summand.

Consider now, for example, the case of X taking the values 1 and -1 each with probability $1/2$. (We shall obtain the De Moivre–Laplace CLT.) Note that for such an X, the variable X^* satisfying $\mathrm{E}Xf(X) = \mathrm{E}f'(X^*)$ has $X^* \sim U(-1, 1)$, that is, the uniform distribution on the interval $(-1, 1)$. It follows that $|W - W^*| = |X_n - X_n^*|/\sqrt{n} \le 2/\sqrt{n}$, and so, applying (7) and (4), we obtain $|\mathrm{E}h(W) - \Phi h| \le 4||h'||/\sqrt{n}$.

Next, we discuss other examples of coupling variables. The next was proposed by Baldi, Rinott and Stein (1989) and Rinott and Goldstein (1996).

Size bias coupling. In our next example we assume $W \ge 0$, and set $\mathrm{E}W = \lambda$, and $\mathrm{Var}W = \sigma^2$. For f and h satisfying the relations (2)–(3), it is easy to see that $\mathrm{E}h(\frac{W-\lambda}{\sigma}) - \Phi h = \mathrm{E}\{f'(\frac{W-\lambda}{\sigma}) - (\frac{W-\lambda}{\sigma})f(\frac{W-\lambda}{\sigma})\}$. Setting $g(w) = f(\frac{w-\lambda}{\sigma})$, we have $\mathrm{E}h(\frac{W-\lambda}{\sigma}) - \Phi h = \mathrm{E}\{\sigma g'(W) - (\frac{W-\lambda}{\sigma})g(W)\}$.

We are ready to introduce the coupling variable W^*, as a variable whose marginal distribution is defined by the relation

$$\mathrm{E}Wg(W) = \lambda \mathrm{E}g(W^*) \tag{8}$$

holding for any g such that the expectations exist. W^* has the familiar size-biased distribution, which is well-known, e.g., in statistics and renewal theory. The construction of such a variable on a joint space with the original W depends on the particular case at hand, and will be discussed briefly later. We obtain

$$\mathrm{E}h\left(\frac{W-\lambda}{\sigma}\right) - \Phi h = \mathrm{E}\{\sigma g'(W) - \frac{\lambda}{\sigma}[g(W^*) - g(W)]\}. \tag{9}$$

We continue with a rough calculation (a formal statement of the bound will be given later), replacing $g(W^*) - g(W)$ by $g'(W)(W^* - W)$, and now neglecting the remainder term in the Taylor expansion. Applying the relation (8) to the function $g(w) = w - \lambda$, we see that $\mathrm{E}(W^* - W) = \sigma^2/\lambda$. If $(W^* - W)$ is small, and has a small variance, it should be close to its expectation with high probability. Then $g'(W)(W^* - W) \approx g'(W)\sigma^2/\lambda$, and we see that the r.h.s. of (9) is small. More precisely, the first term in the expansion of the r.h.s. of (9) equals

$$\frac{\lambda}{\sigma}\mathrm{E}\{g'(W)[\sigma^2/\lambda - (W^* - W)]\} \le \frac{2\lambda||h||}{\sigma^2}\sqrt{\mathrm{Var}\{\mathrm{E}[(W^* - W)|W]\}},$$

where the latter bound is obtained by applying the Cauchy–Schwarz inequality and the relation $||g'|| = \frac{1}{\sigma}||f'|| \le 2||h||/\sigma$.

Clearly, we need to construct W^* jointly with W so that $\mathrm{Var}(W^* - W)$ (or rather the smaller quantity $\mathrm{Var}\{\mathrm{E}[(W^* - W)|W]\}$) is small. In the above two cases we provided the core of the argument, and the details involve only simple technical calculations.

The construction of W^* in this case is somewhat similar to the previous case in the sense that for iid variables it suffices to change one of the summands. An example for dependent variables will be discussed in Subsection 3.1. For further details see Rinott and Goldstein (1996).

We briefly describe two more coupling possibilities.

The exchangeable pair. This approach is highlighted in Stein (1986). For applications and refinements, see also Rinott and Rotar (1997). Again let W be a random variable satisfying $\mathrm{E}W = 0$, $\mathrm{Var}W = 1$. We immediately introduce the coupling variable: suppose there exists, on the same probability space as W, another variable W' such that the pair (W, W') is exchangeable (i.e., the pairs (W, W') and (W', W) have the same distribution) and

$$\mathrm{E}(W'|W) = (1 - \lambda)W \tag{10}$$

for some positive $\lambda < 1$. We shall discuss these conditions and provide examples and modifications later. A direct calculation shows that $\mathrm{E}\{Wf(W)\} = \frac{\mathrm{E}\{(W'-W)[f(W')-f(W)]\}}{2\lambda}$. Together with (3) this implies that

$$\mathrm{E}h(W) - \Phi h = \mathrm{E}f'(W) - \frac{\mathrm{E}\{(W' - W)[f(W') - f(W)]\}}{2\lambda}. \tag{11}$$

Again, we hope to construct W' to be close to W, and expand $f(W') - f(W)$. Replacing $f(W') - f(W)$ by $(W' - W)f'(W)$, (the first term in the Taylor expansion), we see that the r.h.s. of (11) is bounded by $\frac{1}{2\lambda}\mathrm{E}\{f'(W)[2\lambda - (W' - W)^2]\}$ plus a remainder term which we now ignore. By (10), $\mathrm{E}(W' - W)^2 = 2\lambda$, and the Cauchy–Schwarz inequality readily yields

$$\mathrm{E}\{f'(W)[2\lambda - (W' - W)^2]\} \le \frac{||h||}{\lambda}\sqrt{\mathrm{Var}\{\mathrm{E}[(W' - W)^2|W]\}}.$$

It follows that the r.h.s. of (11) is small provided W' and W are close. With a straightforward calculation of the remainder, one indeed obtains a bound for $|\mathrm{E}h(W) - \Phi h|$ (see Theorem 2.4 below).

Here, and in the previous cases, the coupled variable appears in the bound, and in applications one has to construct this variable and use the construction to compute the bound. Some ideas for such constructions will be provided later.

Dependency neighborhoods structure. The final example in this overview is more specific, and deals with the case in which W is a sum of variables

which exhibit some local dependence. Here we describe a simple version. In this example, the coupling is shown explicitly, and the final bound does not contain a coupling variable.

Let $X_1, \ldots, X_n$ be random variables, and $W = \sum_{i=1}^n X_i$, with $\mathrm{E}X_i = \mathrm{E}W = 0$ and $\mathrm{Var}W = 1$. Let $M_i \subset \{1, \ldots, n\}$ be such that $j \in M_i$ if and only if $i \in M_j$ and (X_i, X_j) is independent of $\{X_k\}_{k \notin M_i \cup M_j}$ for $i, j = 1, \ldots, n$. In particular, X_i is independent of $\{X_k\}_{k \notin M_i}$ (take $i = j$ in the previous condition), so that M_i should be interpreted as "a neighborhood of dependence". We remark that the size of these neighborhoods may depend on n, and also that they may be random in some cases. Let $|M_i|$ denote the cardinality of M_i.

We now indicate how to construct a bound for $|\mathrm{E}h(W) - \Phi h|$. The bound itself will depend on $|M_i|$, and some parameters of the X_i's. The details will be given in Theorem 2.1 . In order to define the coupling variable, let I denote a random index uniformly distributed over $\{1, \ldots, n\}$ independent of the X_i's. Now set

$$W^* = W - \sum_{j \in M_I} X_j. \tag{12}$$

Note that here, for the first time, W^* is constructed explicitly, and not just in terms of conditions as, e.g., in (8) or (10).

With f and h related as in (3), we have

$$\begin{aligned} \mathrm{E}h(W) - \Phi h &= \mathrm{E}\{f'(W) - nX_I f(W)\} \\ &= \mathrm{E}\{f'(W) - nX_I[f(W) - f(W^*)]\}, \end{aligned} \tag{13}$$

where the latter equality follows from the fact that the dependence structure and the assumption $\mathrm{E}X_i = 0$ imply that $\mathrm{E}\{nX_I f(W^*)\} = 0$. Observe that, under the present dependence structure, $\mathrm{E}\{nX_I[W - W^*]\} = \mathrm{Var}W = 1$, and moreover, if the sets M_i are small, $\mathrm{E}(nX_I[W - W^*] \mid X_1, \ldots, X_n)$ is close to its expectation, that is, to one, by the strong law of large numbers (for simplicity, think about the case $M_i = \{i\}$ first). A Taylor expansion of $f(W) - f(W^*)$ in (13), neglecting the remainder for the time being, yields $\mathrm{E}\{nX_I[f(W) - f(W^*)]\} \approx \mathrm{E}\{f'(W)nX_I[W - W^*]\} \approx \mathrm{E}f'(W)$, and we can hope for the r.h.s. of (13) to be small. A more detailed calculation will be pursued immediately in the next section.

2. Some theorems

In this section we state some theorems and provide further discussion of the proofs, relying on the calculations already shown. References are given to articles in which complete proofs can be found.

2.1. Local dependence

Consider a sum $W = \sum_{i=1}^{n} Y_i$ of random variables whose dependence can roughly be described as follows: for each i, there is a "dependency neighborhood" of indices M_i such that Y_i is allowed to depend on $\{Y_j : j \in M_i\}$, but Y_i is independent of $\{Y_j : j \notin M_i\}$. The precise dependence structure is specified in Theorem 2.2 below. Note that this structure does not depend on a linear ordering of the variables, unlike many well-known models such as markov chains, martingales, etc. We start with a relatively simple version of Corollary 2 of Stein (1986), which leads to a useful result and demonstrates some of the calculations. In this first version, we do not deal with approximating the distribution function, but rather with expectations of smooth functions. This issue will be discussed later.

Theorem 2.1. *Let $Y_1, \ldots, Y_n$ be random variables, and let $M_i \subset \{1, \ldots, n\}$ be such that $j \in M_i$ if and only if $i \in M_j$ and (Y_i, Y_j) is independent of $\{Y_k\}_{k \notin M_i \cup M_j}$. Assume $\mathrm{E}Y_i = 0$, and $\mathrm{Var} \sum_{i=1}^{n} Y_i = \sigma^2 > 0$. Then, for any function h, continuous, and piecewise continuously differentiable,*

$$|\mathrm{E}h\left(\frac{\sum_{i=1}^{n} Y_i}{\sigma}\right) - \Phi h| \le \frac{2}{\sigma^2}||h|| \sqrt{\mathrm{E}\left\{\sum_{i=1}^{n}\sum_{j\in M_i}(Y_iY_j - \mathrm{E}Y_iY_j)\right\}^2} + \frac{1}{\sigma^3}||h'||\mathrm{E}\left\{\sum_{i=1}^{n}|Y_i|(\sum_{j\in M_i} Y_j)^2\right\}. \tag{14}$$

Proof of Theorem 2.1. Given h, let f denote the function defined by (2). Set $X_i = Y_i/\sigma$, $W = \sum_{i=1}^{n} X_i$, and $X = \{X_i : i = 1, \ldots, n\}$, and let W^* be defined by (12). Note that $\mathrm{E}[nX_I f(W^*)] = 0$ by the assumptions on the dependence structure, and $\mathrm{E}[nX_I f(W)] = \mathrm{E}\{\mathrm{E}[nX_I f(W) \mid X]\} = \mathrm{E}Wf(W)$. Simple manipulations and a Taylor series expansion of $f(W) - f(W^*)$ (with integral remainder) yield

$$\mathrm{E}h(W) - \Phi h = \mathrm{E}\{f'(W)[1 - nX_I(W - W^*)]\} + \mathrm{E}\int_{W^*}^{W} nX_I(t - W^*)df'(t)\,. \tag{15}$$

With the bound (4b) and the observation that $\sum_{i=1}^{n}\sum_{j\in M_i}\mathrm{E}X_iX_j = \mathrm{Var}W = 1$, we obtain by the Cauchy–Schwarz inequality

$$|\mathrm{E}\{f'(W)[1-nX_I(W-W^*)]\}| \leq 2||h||\sqrt{\mathrm{E}\left\{\sum_{i=1}^{n}\sum_{j\in M_i}(X_iX_j-\mathrm{E}X_iX_j)\right\}^2}. \quad (16)$$

Note that another possible bound is

$$2||h||\mathrm{E}\left|\sum_{i=1}^{n}\sum_{j\in M_i}(X_iX_j-\mathrm{E}X_iX_j)\right|.$$

The latter bound does not require the existence of fourth order moments, but, as it is harder to compute, we shall not pursue it here.

Returning to the second term in (15), we use (4c) and integrate to obtain

$$\mathrm{E}\int_{W^*}^{W} nX_I(t-W^*)df'(t) \leq ||h'||\mathrm{E}\left\{\sum_{i=1}^{n}|X_i|(\sum_{j\in M_i}X_j)^2\right\}.$$

This completes the proof.

A discussion of non-smooth h. Note that in the last bound the term $||h'||$ may be large or even infinite for a non-smooth h such as an indicator of an interval. However, we used the relation $||f''|| \leq 2||h'||$ only in the interval (W^*, W). In the next theorem we apply Theorem 2.1, where h is the indicator function of a half interval, having a derivative which vanishes everywhere except for one point, and we can approximate it by a smooth function whose derivative vanishes in the random (but small) interval (W^*, W) with high probability. Thus we can refine the last bound. This requires further conditions. The following version from Dembo and Rinott (1996) is rather useful. We sketch its proof without going into technicalities. For further, more advanced treatment of the latter issue, see Rinott and Rotar (1996).

Theorem 2.2. *Let $Y_1, \ldots, Y_n$ be random variables satisfying $|Y_i - \mathrm{E}(Y_i)| \leq B$ a.s., $i = 1, \ldots, n$, $\mathrm{E}\sum_{i=1}^{n} Y_i = \lambda$, $\mathrm{Var}\sum_{i=1}^{n} Y_i = \sigma^2 > 0$ and $\frac{1}{n}\mathrm{E}\sum_{i=1}^{n}|Y_i - \mathrm{E}(Y_i)| = \mu$. Let $M_i \subset \{1, \ldots, n\}$ be such that $j \in M_i$ if and only if $i \in M_j$ and (Y_i, Y_j) is independent of $\{Y_k\}_{k\notin M_i\cup M_j}$ for $i, j = 1, \ldots, n$, and set $D = \max_{1\leq i\leq n}|M_i|$. Then*

$$\left|P\left(\frac{\sum_{i=1}^{n}Y_i-\lambda}{\sigma}\leq w\right)-\Phi(w)\right| \leq 7\frac{n\mu}{\sigma^3}(DB)^2. \quad (17)$$

Proof of Theorem 2.2. Without loss of generality assume that $\mathrm{E}(Y_i) = 0$. Hence $|Y_i| \le B$ a.s., $i = 1, \dots, n$ and $\mu = \mathrm{E}(|Y_I|)$, where the random index I is uniformly distributed over $\{1, \dots, n\}$ independently of the Y's. Set $X_i = Y_i/\sigma,\ i = 1, \dots, n$, and $W = \sum_{i=1}^{n} X_i$.

Let $U_{i,j} = X_i X_j - \mathrm{E}X_i X_j$, so that

$$\mathrm{E}\left\{\sum_{i=1}^{n}\sum_{j\in M_i}(X_i X_j - \mathrm{E}X_i X_j)\right\}^2 = \sum_{i=1}^{n}\sum_{j\in M_i}\sum_{k=1}^{n}\sum_{\ell\in M_k}\mathrm{E}U_{i,j}U_{k,\ell}\,. \quad (18)$$

Observe that, for fixed i, there are at most $4D^3$ non-zero terms in the r.h.s. of (18), each of which is bounded by $2(B/\sigma)^3\mathrm{E}|X_i|$. Consequently, the r.h.s. of (18) is bounded by $8n\mu(DB/\sigma)^3\frac{1}{\sigma}$. As

$$\sigma^2 = \sum_{i=1}^{n}\sum_{j\in M_i}\mathrm{E}(Y_i Y_j) \le \sum_{i=1}^{n}\sum_{j\in M_i} B\mathrm{E}|Y_i| \le BDn\mu,$$

we conclude that the r.h.s. of (18) is bounded by $8(n\mu)^2(DB)^4/\sigma^6$. Taking the square root of the latter term we obtain a bound on the r.h.s. of (16) of the required order, $\frac{n\mu}{\sigma^3}(DB)^2$.

The second term in (15) was shown to be bounded by

$$||h'||\mathrm{E}\left\{\sum_{i=1}^{n}|X_i|(\sum_{j\in M_i}X_j)^2\right\}.$$

An easy calculation shows that it is bounded by $||h'||\frac{n\mu}{\sigma^3}(DB)^2$. We need to apply Theorem 2.1 where h is the indicator of half an interval. If h' were bounded the proof would be complete (apart from constants). However, as indicated above h' is not bounded. Technical arguments (to which we alluded above) are required now to show that h can be appropriately approximated to complete the proof.

In order to understand the above theorems in the simplest case, consider the classical CLT for iid variables. Let the summands Y_i be iid, with a finite fourth moment. Theorem 2.1 applies with $M_i = \{i\}$. The first term in (14) reduces to $\frac{2}{\sigma^2}||h||\sqrt{n\mathrm{Var}Y_1^2}$. Since $\sigma^2 = n\mathrm{Var}Y_1$, it follows that for a bounded h, this term has the order of $1/\sqrt{n}$. The second term in (14) reduces in the case of iid Y_i to $\frac{1}{\sigma^3}||h'||\mathrm{E}\{\sum_{i=1}^{n}|Y_i|^3\}$. If h' is bounded this term is also of the order of $1/\sqrt{n}$. For bounded iid variables Y_i, Theorem 2.2 yields the correct rate of convergence: $1/\sqrt{n}$. This follows immediately from (17) with $D = 1$. While the above discussion provides a rather elementary proof of some CLT's for iid variables, we emphasize that the main interest in

Stein's method lies in its applicability to a large variety of non-independent cases. □

2.2. Size bias coupling

We provide here one result from Goldstein and Rinott (1996). An indication of the proof was given in the introduction.

Theorem 2.3. *Let $W \geq 0$ be a random variable with distribution $dF(w)$ and let W^* be defined on the same probability space as W and having the marginal distribution $wdF(w)/\lambda$, where $\lambda = \mathrm{E}W$, and $\sigma^2 = \mathrm{Var}W$. Then, for any piecewise continuously differentiable $h : R \to R$,*

$$|\mathrm{E}h\left(\frac{W-\lambda}{\sigma}\right) - \Phi h| \leq 2\|h\|\frac{\lambda}{\sigma^2}\sqrt{\mathrm{Var}\mathrm{E}(W^* - W|W)} + \|h'\|\frac{\lambda}{\sigma^3}\mathrm{E}(W^* - W)^2.$$

2.3. The exchangeable pair coupling

We first quote a slightly modified but equivalent version of a theorem of Stein (1986, p. 35), and then some extensions. The proof was sketched in the introduction.

Theorem 2.4. *Let (W, W') be a pair of exchangeable random variables (i.e., their joint distribution is symmetric), and suppose $\mathrm{E}W = 0$, $\mathrm{Var}W = 1$ and*

$$\mathrm{E}(W'|W) = (1-\lambda)W \tag{19}$$

for some positive $\lambda < 1$. Then for any continuously differentiable bounded function h,

$$|\mathrm{E}h(W) - \Phi h| \leq \frac{1}{\lambda}||h||\sqrt{\mathrm{Var}\{\mathrm{E}[(W'-W)^2|W]\}} + \frac{1}{4\lambda}(||h'||)\,\mathrm{E}\{|W'-W|^3\}.$$

When h' is not bounded, the above bound blows up, and as before a more careful (and rather technical) analysis is needed. For indicators of half intervals (for further details and other classes of functions, see Rinott and Rotar (1997)) we have Theorem 2.5 below. It also extends the range of applications of this approach by replacing (19) by a weaker condition, allowing (19) to hold only approximately, with a remainder denoted by R. The discussion of weighted nondegenerate U-statistics demonstrates the utility of this extension.

Theorem 2.5. *Let* (W, W') *be exchangeable,* $\mathrm{E}W = 0$, $\mathrm{Var}W = 1$. *Define the r.v.* $R = R(W)$ *by*

$$\mathrm{E}(W'|W) = (1-\lambda)W + R, \tag{20}$$

where λ *is a number satisfying* $0 < \lambda < 1$. *Then, for all real* w,

$$|P(W \le w) - \Phi(w)| \le \frac{6}{\lambda}\sqrt{\mathrm{Var}\{\mathrm{E}[(W'-W)^2|W]\}} + 19\frac{\sqrt{\mathrm{E}R^2}}{\lambda} + 6\sqrt{\frac{1}{\lambda}\mathrm{E}\{|W'-W|^3\}}. \tag{21}$$

Also, if

$$|W'-W| \le A$$

for some constant A, *then*

$$|P(W \le w) - \Phi(w)| \le \frac{12}{\lambda}\sqrt{\mathrm{Var}\{\mathrm{E}[(W'-W)^2|W]\}} + 37\frac{\sqrt{\mathrm{E}R^2}}{\lambda} + 48\frac{A^3}{\lambda} + 8\frac{A^2}{\sqrt{\lambda}}. \tag{22}$$

3. Some examples and applications

We describe some examples without proofs and details.

3.1. Local maxima and Nash equilibria

Consider the vectors $\mathbf{a} = (a_1, \ldots, a_p)$, where each of the p coordinates a_i takes values in $\{1, 2, ..., s\}$. Let $\mathbf{a}|b_i$ denote a vector which differs from $\mathbf{a}$ in a single coordinate, more precisely, $\mathbf{a}|b_i = (a_1, \ldots, a_{i-1}, b_i, a_{i+1}, \ldots, a_p)$, for $i = 1, ..., p$ and $b_i \in \{1, ..., s\}$. Let $V_\mathbf{a}$, defined for all $\mathbf{a}$, be iid random variables. We say that there is a "local maximum" at $\mathbf{a}$ if $V_\mathbf{a} \ge V_{\mathbf{a}|b_i}$ for all possible i and b_i. Let M denote the number of local maxima. We are interested in the asymptotic distribution of M for large s and p. More generally, for each $\mathbf{a}$ define a vector $\mathbf{V}_\mathbf{a} = (V_\mathbf{a}^{(1)}, \ldots, V_\mathbf{a}^{(p)})$, where again the vectors defined are iid. We shall discuss a variety of dependency conditions on the coordinates $V_\mathbf{a}^{(1)}, \ldots, V_\mathbf{a}^{(p)}$ within each vector.

If we consider a game with p players where $V_\mathbf{a}^{(i)}$ represents the payoff to player i when the p players choose the pure strategies $a_1, \ldots, a_p$ respectively, then $\mathbf{a} = (a_1, \ldots, a_p)$ is a Nash equilibrium point (in pure strategies) if the condition $V_\mathbf{a}^{(i)} \ge V_{\mathbf{a}|b_i}^{(i)}$ holds for all i and $b_i \in \{1, ..., s\}$. We shall study

the number N of Nash equilibria. In the case that $V_{\mathbf{a}}^{(1)} = \ldots = V_{\mathbf{a}}^{(p)}$ (with probability 1), the notions of Nash equilibria and of local maxima described above coincide. In this case all players share the same payoff.

Baldi, Rinott and Stein (1989) computed $\mathrm{E}M = s^p/[(s-1)p+1]$ (easy), and $\mathrm{Var}M = s^p(p-1)(s-1)/2[(s-1)p+1]^2$, and proved that M is asymptotically normal when either $p \to \infty$ or $s \to \infty$ (or both). They used a version of Theorem 2.3 with $W = M$, and W^* is constructed as follows: choose a vertex $\mathbf{a}$ at random. If $V_{\mathbf{a}}$ is a local maximum, set $W^* = W$. Otherwise, interchange the values of V at $\mathbf{a}$ and the largest of its values at neighboring vertices $\mathbf{a}|b_i$, that is, the largest among all $V_{\mathbf{a}|b_i}$, leaving all other V's unchanged. It is easy to see that after this interchange there will be a local maximum at $\mathbf{a}$. It can be shown that the resulting number of local maxima has the right distribution for W^*. Also, from the above one sees that W and W^* are close, as required.

Another possible approach to this problem is to use local dependence. Define a distance between vectors (of strategies), $d(\mathbf{a}, \mathbf{b})$= the number of coordinates in which $\mathbf{a}$ and $\mathbf{b}$ differ, and let $Y_{\mathbf{a}}$ take the value 1 if there is a local maximum at $\mathbf{a}$, and 0 otherwise. Then $M = \sum_{\mathbf{a}} Y_{\mathbf{a}}$. It is not hard to show that if $d(\mathbf{a}, \mathbf{b}) > 2$, then $Y_{\mathbf{a}}$ and $Y_{\mathbf{b}}$ are independent. Moreover the conditions of Theorem 2.2 can be easily verified, with the appropriate constants, and with the dependency neighborhoods defined by $M_{\mathbf{a}} = \{\mathbf{b} : d(\mathbf{a}, \mathbf{b}) \leq 2\}$, leading to results on asymptotic normality of the number of local maxima M.

With some technical complications, the same holds for the number of Nash equilibria N, leading to the following result from Rinott and Scarsini (1999).

Theorem 3.1. *If the components of* $\mathbf{V}_{\mathbf{a}} = (V_{\mathbf{a}}^{(1)}, \ldots, V_{\mathbf{a}}^{(p)})$ *are positively quadrant dependent, then there exists a constant* c*, depending on the distribution of* $\mathbf{V}_{\mathbf{a}}$*, such that, for all* t*,*

$$\left| P\left(\frac{N-\lambda}{\sigma} \leq t \right) - \Phi(t) \right| \leq c \frac{s^4 p^4}{s^{p/2} Q^{1/2}}, \tag{23}$$

where $Q =$ *the probability of a Nash point at a given* $\mathbf{a}$, $\lambda = \mathrm{E}N = s^p Q$, *and* $\sigma^2 = \mathrm{Var}N$.

This theorem is stated in Rinott and Scarsini (1999) for the special case in which the components of $\mathbf{V}_{\mathbf{a}}$ are normal with positive correlations, along with further results such as bounds on σ^2 and on Q, and conditions under which the r.h.s. of (23) converges to 0, with some discussion comparing the applicability of Theorems 2.2 and 2.3. The condition of positive quadrant dependence is, of course, a general condition of positive dependence, for which positively correlated normals provide an example. It can be shown that, in the case in which the players' payoffs are independent, N converges

to a Poisson ($\lambda = 1$) variable whereas, in the case of suitably defined negative dependence (e.g., in a zero-sum game), N converges to 0.

3.2. Functions of stationary processes and the exchangeable pair coupling

The exchangeable pair coupling will now be briefly explained. The reader might wonder why and when, for a random variable W, it may be a natural to construct the required W' with the conditions of Theorems 2.4 or 2.5.

Following Diaconis (1989), Diaconis and Sturmfels (1998), Barbour (1990, 1997) and Rinott and Rotar (1997), consider a stationary process $\{\mathbf{X}^{(t)}\}$, where $t = 1, 2, ...$ is a time parameter, and suppose we want to study the proximity to normality of some function of the process $\Gamma(\mathbf{X}^{(t)})$. It is then natural to choose $(W, W') = (\Gamma(\mathbf{X}^{(t)}), \Gamma(\mathbf{X}^{(t+1)}))$. As of $\{\mathbf{X}^{(t)}\}$ is stationary, W and W' have the same marginal distributions. Exchangeability of this pair clearly holds if the process $\{\mathbf{X}^{(t)}\}$ is reversible. A typical case is that of $\mathbf{X}^{(t)} = (X_1^{(t)}, ..., X_n^{(t)})$, where n denotes the size, or dimension, of the process, and $\Gamma(\mathbf{X}^{(t)}) = X_1^{(t)} + ... + X_n^{(t)}$.

If W is close to normal, one may hope that the pair (W, W') is close to bivariate normal, and then the linearity of the conditional expectation of W' as a function of W should hold approximately; this indicates that (20) is a natural condition in the present setup, and one may expect the remainder term R to be small. In fact, R can be viewed as a remainder term in the expansion of the conditional expectation of $W' - W$, centered at W.

The reader may now try to prove the classical CLT for the sum of iid variables $X_1, ..., X_n$ having a common cdf F, by setting $\mathbf{X}^{(1)} = (X_1, ..., X_n)$, and then obtaining $\mathbf{X}^{(t+1)}$ from $\mathbf{X}^{(t)}$ by choosing an index i at random and replacing $X_i^{(t)}$ by an independent copy with the distribution F, leaving the other coordinates unchanged. This defines a stationary, reversible Markov chain and, since only one coordinate changes at a time, one can hope that W and W' will be close. Making use of (21) leads to a rate of $1/n^{1/4}$, whereas for bounded variables, (22) yields the rate of $1/\sqrt{n}$. While the independent case provides a simple test case, the value of the method lies in its ability to yield results under interesting dependence structures.

Asymptotic normality of more complex functions, such as U-statistics of the form

$$U = \Gamma(\mathbf{X}) = \sum \gamma(X_{i_1}, ...X_{i_k}),$$

where the function γ may depend on n, can be studied by the same method, using the same process. Clearly the summands are no longer independent. For degenerate U-statistics, that is, when $P(\mathrm{E}\{\gamma(X_1, ..., X_k) \mid X_1\} = 0\ ) = 1$, one obtains $R = 0$ in (20), otherwise $R \neq 0$. General results giving conditions for normality for the different cases, and for more general types of

U-statistics by this method, are given in Rinott and Rotar (1997). Finally we mention that the latter paper provides a detailed study of a case where $X_1^{(t)}, ..., X_n^{(t)}$ are dependent, denoting the states of a certain stationary particle system (the anti-voter model) at n different sites at time t. Although the anti-voter chain is not reversible, further arguments lead to conditions for asymptotic normality of $X_1^{(t)} + ... + X_n^{(t)}$ for large n.

Remark 3.2. The reference list below is very incomplete and contains only papers to which we referred above and a few others with further applications.

References

1. Avram, F., Bertsimas, D. (1993): On central limit theorems in geometrical probability. *The Annals of Applied Probability* **3**, 1033–1046
2. Baldi, P., Rinott, Y. (1989): On normal approximations of distributions in terms of dependency graphs. *The Annals of Probability* **17**, 1646–1650
3. Baldi, P., Rinott Y., Stein, C. (1989): A normal approximation for the number of local maxima of a random function on a graph. In: Anderson, T. W., Athreya, K. B., Iglehart, D. L. (eds): Probability, statistics, and mathematics. Papers in honor of Samuel Karlin. Academic Press, Boston, pp. 59–81
4. Barbour, A.D. (1990): Stein's method for diffusion approximations. *Probability Theory and Related Fields* **84**, 297–332
5. Barbour, A.D. (1997): Stein's method. In: Kotz, S., Read, C. B., Bakks, D.L. (eds): Encyclopedia of statistical sciences. Update Vol. Wiley, New York
6. Barbour, A.D., Hall, P. (1984): Stein's method and the Berry-Esseen theorem. *The Australian Journal of Statistics* **26**, 8–15
7. Barbour, A.D., Holst, L., Janson, S. (1992): Poisson approximation. Oxford Studies in Probability **2**. Clarendon Press, Oxford
8. Barbour, A.D., Karoński, M., Ruciński, A. (1989): A central limit theorem for decomposable random variables with applications to random graphs. *Journal of Combinatorial Theory. Series B* **47**, 125–145
9. Bolthausen, E. (1982): Exact convergence rates in some martingale central limit theorems. *The Annals of Probability* **10**, 672–688
10. Bolthausen, E. (1984): An estimate of the remainder in a combinatorial central limit theorem. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **66**, 379–386
11. Bolthausen, E., Götze, F. (1993): The rate of convergence for multivariate sampling statistics. *The Annals of Statistics* **21**, 1692–1710
12. Chen, L.H.Y. (1998): Stein's method: some perspectives with applications. Probability towards 2000. (Lecture Notes in Statistics) **128**. Springer, New York, pp. 97–122
13. Dembo, A, Rinott, Y. (1996): Some examples of normal approximations by Stein's method. In: Aldous, D., Pemantle, R. (eds): Random Discrete Structures. (The IMA Volumes in Mathematics and its Applications, **76**). Springer, New York, pp. 25–44
14. Diaconis, P. (1989): An example of Stein's method. Stanford Statistics Department Technical Report
15. Diaconis, P., Sturmfels, B. (1998): Algebraic algorithm for sampling from conditional distributions. *Annals of Statistics* **26**, 363–397
16. Goldstein, L., Reinert, G. (1997): Stein's method and the zero bias transformation with applications to simple random sampling. *The Annals of Applied Probability* **7**, 935–952

17. Goldstein, L., Rinott, Y. (1996): Multivariate normal approximations by Stein's method and size bias couplings. *Journal of Applied Probability* **33**, 1–17
18. Götze, F. (1991): On the rate of convergence in the multivariate CLT. *The Annals of Probability* **19**, 724–739
19. Lindvall, T. (1992): Lectures on the coupling method. (Wiley Series in Probability and Mathematical Statistics.) Wiley, New York
20. Rinott, Y., Rotar, V. (1996): A multivariate CLT for local dependence with $n^{-1/2}logn$ rate and applications to multivariate graph related statistics. *Journal of Multivariate Analysis* **56**, 333–350
21. Rinott, Y., Rotar, V. (1997): On coupling constructions and rates in the CLT for dependent summands with applications to the anti voter model and weighted U-statistics. *The Annals of Applied Probability* **7**, 1080–1105
22. Rinott, Y., Scarsini, M. (2000): On the number of pure strategy nash equilibria in random games. To appear in Games and Economic Behavior
23. Ruciński, A. (1992): Proving normality in combinatorics. In: Frieze, A., Łuczak, T. (eds.): Random graphs. Volume 2. Wiley, New York, pp. 215–231
24. Stein, C. (1972): A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics Probability **2**. Univ. California Press, Berkeley, pp. 583–602
25. Stein, C. (1986): Approximate computation of expectations. Institute of Mathematical Statistics, Hayward, CA
26. Stein, C. (1992): A way of using auxiliary randomization. In: Chen, L.H.Y., Choi, K.P., Hu, K., Lou, J.H. (eds): Probability theory. De Gruyter, Berlin, pp. 159–180