Progress in Probability, Vol. 43
© 1998 Birkhäuser Verlag Basel/Switzerland

A Central Limit Theorem for the Sock-Sorting Problem

WENBO V. LI AND GEOFFREY PRITCHARD

ABSTRACT. The problem of arranging $2n$ objects into n pairs in a prescribed way, when the objects are presented one at a time in random order, is considered. Using tools from the theory of empirical processes, we derive a functional central limit theorem, with a limiting Gaussian process closely related to the Brownian sheet.

1. Introduction and statement of results

This paper considers the following problem, which seems to have been first proposed in [5]; see also [8] and [10]. (For a related problem, see [6].) A collection of n different pairs of socks are scrambled in a laundry bag. Socks are drawn one at a time from the bag in random order, and laid on a table. When the mate of a sock on the table is drawn, the two are paired and put in a drawer. How much table space is required?

Our main result is a functional central limit theorem for the table-space usage as a function of the number of socks so far drawn. The Gaussian process occurring in the limit is an interesting one which can be described as the diagonal of a constrained Brownian sheet, among other representations. As a corollary, we obtain an asymptotic distribution for the maximum space required, as n becomes large. Simulations have shown that this asymptotic distribution closely approximates the true one as soon as n is greater than about 20.

We now state our main result. Let L_k be the number of socks on the table after k socks have been drawn.

THEOREM 1.1. *Let $X_n \in C[0,1]$ be the usual piecewise linear interpolation of the sequence $\{L_k\}_{k=0}^{2n}$, with $X_n(k/2n) = L_k$. Let $q(t) = 2t(1-t)$. Then*

$$\frac{X_n(t) - nq(t)}{2\sqrt{n}} \Longrightarrow \Delta(t).$$

Here $\Longrightarrow$ denotes weak convergence with respect to the uniform norm topology on $C[0,1]$, and $\Delta(t) = \sigma(t,t)$ with σ the Brownian sheet on the unit square constrained to be 0 on the whole boundary of the square. In other words,

$$\sigma(s,t) = V(s,t) - sV(1,t) - tV(s,1) + stV(1,1)$$

where V is the usual Brownian sheet with $Cov(V(s,t), V(s',t')) = (s \wedge s')(t \wedge t')$.

The proof we will present of Theorem 1.1 will proceed by embedding the process in the uniform distribution on $[0,1]^{2n}$, and using empirical-process theory. This is the "slickest" proof known to us, although it is not the only one. One alternative (our original proof of this result) proceeds by finding an integral operator T such that when T is applied to (a close approximation of) $X_n - nq$ the result is a martingale, making use of a central limit theorem for martingale triangular arrays, and finally transforming back to the original problem by applying the inverse of T. This proof still uses the embedding in the uniform distribution. After we had completed this first proof, it was pointed out to us by D. Mason and Z. Shi that the embedding idea is powerful enough to give a direct empirical-process proof. A third possible approach to this problem (suggested by T. Kurtz) is to apply a diffusion limit theorem for Markov chains (see, for example, Chapter 11 of [7]).

The process σ also appears in empirical-process theory as the limit of the Hoeffding, Blum, Kiefer, Rosenblatt empirical process; see [2].

In §2 we also consider other representations of the limit process Δ, and return to the original question of the maximum table-space usage.

2. Proof and remarks

Proof of Theorem 1.1. Let $\{X_i\}_{i=1}^n$,$\{Y_i\}_{i=1}^n$ be independent i.i.d. random samples from the uniform distribution on $[0,1]$. (X_i,Y_i may be thought of as the "times" at which the socks of the ith pair are drawn.) Define empirical distribution functions F_n, G_n by

$$G_n(s,t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \le s, Y_i \le t\}} \qquad \text{and} \qquad F_n(t) = \frac{1}{2n}\sum_{i=1}^{n}\left(1_{\{X_i \le t\}} + 1_{\{Y_i \le t\}}\right),$$

for $s,t \in [0,1]$. Let Q_n denote the quantile function corresponding to F_n, i.e. $Q_n(u) = \inf\{t : F_n(t) \ge u\}$. Then $nG_n(Q_n(k/2n), Q_n(k/2n))$ is the number of completed pairs after k socks have been drawn, i.e. is $(k - L_k)/2$. We thus see that $(X_n(u) - nq(u))/2\sqrt{n}$ is within $O(n^{-1})$ of

$$\frac{2nu - 2nG_n(Q_n(u), Q_n(u)) - nq(u)}{2\sqrt{n}} \tag{1}$$

uniformly in u. So it will suffice to show that quantity (1) converges weakly to $\Delta(u)$, using the Skorokhod topology on the space $D[0,1]$. To this end, write (1) as

$$\sqrt{n}\left(Q_n(u)^2 - G_n(Q_n(u), Q_n(u))\right) - \sqrt{n}\left(Q_n(u) - u\right)\left(Q_n(u) + u\right).$$

For the purpose of establishing the weak convergence, this may be replaced by

$$\sqrt{n}\left(u^2 - G_n(u,u)\right) - \sqrt{n}\left(u - F_n(u)\right)2u.$$

The replacement of $Q_n(u)$ by u in the first term is a "random time change" (see [1], p.144), justified by the Glivenko-Cantelli theorem ($Q_n(u) \to u$ a.s.; see [4], p.56). The replacement of $Q_n(u) - u$ by $u - F_n(u)$ is justified by the Bahadur-Kiefer theorem (see e.g. [3]). The expression we now have can be written

$$\sqrt{n}\left(u^2 - G_n(u,u)\right) - u\sqrt{n}(u - G_n(u,1)) - u\sqrt{n}(u - G_n(1,u)).$$

The following functional central limit theorem is well-known (see e.g. [9]):

$$\sqrt{n}(G_n(s,t) - st) \Longrightarrow B(s,t)$$

where $B(s,t) = V(s,t) - stV(1,1)$, with V the usual Brownian sheet. Since we have reduced our process to a linear transformation of this one, we conclude that it converges weakly to

$$-B(u,u) + uB(1,u) + uB(u,1),$$

which is $-\Delta(u)$. This has the same law as $\Delta(u)$. □

The limiting process Δ can be described in several ways, one of which was given in the statement of Theorem 1.1. Another representation (discovered as a consequence of the alternative proof of Theorem 1.1 mentioned in the introduction) is the stochastic integral:

$$\Delta(t) = (1-t)^2 \int_0^t (2s)^{1/2}(1-s)^{-3/2} dW(s),$$

where W is the usual Wiener process. To check this, it suffices to note that both processes are Gaussian, and observe that their covariance structures are the same, namely

$$\operatorname{Cov}(\Delta(s), \Delta(t)) = (s \wedge t - st)^2.$$

Another way to view Δ is as a diffusion process, described by the associated stochastic differential equation. This may be easily found from the stochastic integral representation to be

$$d\Delta(t) = -2\left(\frac{\Delta(t)}{1-t}\right) dt + \sqrt{2t(1-t)} dW(t), \qquad \Delta(0) = 0.$$

Compare this with the equation for the more common Brownian bridge process B:

$$dB(t) = -\left(\frac{B(t)}{1-t}\right) dt + dW(t), \qquad B(0) = 0.$$

Though Δ is similar to B in vanishing at the endpoints 0 and 1, its equation has a stronger "drift" coefficient and a variable "speed" coefficient.

Finally, we return to the original problem – that of the maximum table-space usage. An asymptotic law for this quantity is now an easy corollary of our main result.

PROPOSITION 2.1. *Let $K_n = \max_{k=0}^{2n} L_k$. Then*

$$\frac{K_n - n/2}{\sqrt{n/4}} \Longrightarrow Y,$$

where Y has a standard normal distribution.

Remark. This says that K_n has the same asymptotic behaviour as L_n; essentially because for large n the maximum L_k will occur for $k \approx n$.

Proof. Let $\Phi(\alpha) = P(Y \leq \alpha)$. We need to show $P\left((K_n - \frac{1}{2}n)/\frac{1}{2}\sqrt{n} > \alpha\right) \to \Phi(\alpha)$ for all $\alpha \in \mathbb{R}$. Noting $K_n \geq L_n$ gives $\underline{\lim}_n P\left((K_n - \frac{1}{2}n)/\frac{1}{2}\sqrt{n} > \alpha\right) \geq \Phi(\alpha)$. Also, with X_n as in Theorem 1.1 and any $c > 0$,

$$\begin{aligned}
&\overline{\lim_n} P\left(K_n > \frac{1}{2}n + \frac{1}{2}\sqrt{n}\alpha\right) \\
&\leq \overline{\lim_n} P\left(\sup_{|t-\frac{1}{2}|\leq c} X_n(t) > \frac{1}{2}n + \frac{1}{2}\sqrt{n}\alpha\right) + \overline{\lim_n} P\left(\sup_{|t-\frac{1}{2}|\geq c} X_n(t) > \frac{1}{2}n + \frac{1}{2}\sqrt{n}\alpha\right) \\
&\leq \overline{\lim_n} P\left(\sup_{|t-\frac{1}{2}|\leq c} \frac{X_n(t) - nq(t)}{2\sqrt{n}} > \frac{\alpha}{4}\right) \\
&\quad + \overline{\lim_n} P\left(\sup_{|t-\frac{1}{2}|\geq c} \frac{X_n(t) - nq(t)}{2\sqrt{n}} > \frac{\frac{1}{2}n + \frac{1}{2}\sqrt{n}\alpha - nq(\frac{1}{2}+c)}{2\sqrt{n}}\right) \\
&\leq P\left(\sup_{|t-\frac{1}{2}|\leq c} \Delta(t) > \frac{\alpha}{4}\right) + \overline{\lim_n} P\left(\sup_{0\leq t\leq 1} \Delta(t) > c^2\sqrt{n} + \frac{\alpha}{4}\right).
\end{aligned}$$

The second term is 0 for any $c > 0$, since $\sup_{0\leq t\leq 1} \Delta(t) < \infty$ a.s. The first term goes to $\Phi(\alpha)$ as $c \to 0$ by the Dominated Convergence Theorem, noting that $\lim_{c\to 0} \sup_{|t-\frac{1}{2}|\leq c} \Delta(t) = \Delta(1/2)$ a.s. □

Acknowledgement. The authors wish to thank D. Mason, Z. Shi, and T. Kurtz for useful discussions.

References

[1] P. Billingsley, *Convergence of probability measures* Wiley, New York, 1968.

[2] Csörgo, *Strong approximations of the Hoeffding, Blum, Kiefer, Rosenblatt multivariate empirical process*, J. Multivariate Analysis **9** (1979), 84–100.

[3] P. Deheuvels and D. Mason, *Bahadur-Kiefer-type processes*, Ann. Prob. **18** (1990), 669–697.

[4] R. Durrett, *Probability: Theory and examples* 1st ed., Wadsworth, Pacific Grove, California, 1991.

[5] M. P. Eisner, Problem 216, College Mathematics Journal **13** (1982), 206.

[6] D. M. Friedlen, Problem E3265, Amer. Math. Monthly **97** (1990), 242–244.

[7] S. Ethier and T. Kurtz, *Markov processes: characterization and convergence* Wiley, New York, 1986.

[8] R. Luttmann, Problem E3148, Amer. Math. Monthly **95** (1988), 357–358.

[9] E. Giné and J. Zinn, *Some limit theorems for empirical processes*, Ann. Prob. **12** (1984), 929–989.

[10] S. Rabinowitz *Index to Mathematical Problems 1980–1984*, MathPro Press, Westford, MA, 1992.

Wenbo V. Li
Department of Mathematical Sciences
University of Delaware
Newark, DE 19716, USA
wli@math.udel.edu

Geoffrey Pritchard
Department of Mathematics
Texas A&M University,
College Station, TX 77843, USA
Geoffrey.Pritchard@math.tamu.edu