Canad. Math. Bull. Vol. **22** (4), 1979

CENTRAL LIMIT THEOREM FOR ABSOLUTE DEVIATIONS FROM THE SAMPLE MEAN AND APPLICATIONS

BY
D. L. McLEISH

1. **Introduction.** The following type of argument is rendered almost believable by its frequent occurrence in elementary courses in statistics. Let ξ_i be a sequence of independent identically distributed random variables with means μ variances σ^2. Then $1/\sigma\sqrt{n}\sum_{i=1}^{n}(\xi_i-\mu)$ converges in law to the standard normal distribution. Since $\bar{\xi}_n = 1/n\sum_{i=1}^{n}\xi_i$ is a consistent (and fairly rapidly converging) estimator of μ, this result should remain essentially unchanged if we replace μ above by $\bar{\xi}_n$. What, in fact, occurs is that normality is preserved, but the variance is affected. In this case, the sum turns out to be identically 0, i.e. the normal distribution with mean and variance both 0. The defect in the above argument clearly is that $\bar{\xi}_n \to \mu$ only at the same rate that $1/\sqrt{n} \to 0$. Indeed, if we replace μ by $\bar{\xi}_{2n}$, convergence to the normal $(0, \frac{1}{2})$ variate obtains, and if it is replaced by $\bar{\xi}_{n^2}$, again we obtain convergence to the standard normal. These are all rather trivial consequences of an invariance principle (cf Billingsley [3, Theorem 24.2]).

In this note we wish to investigate another example of the same phenomenon and some of its applications. Consider the sums $\sum_{i=1}^{n}(\xi_i - d\bar{\xi}_n)I(\xi_i > d\bar{\xi}_n)$ and $\sum_{i=1}^{n} g_{d\bar{\xi}_n}(\xi_i)$ where $I(A)$ represents the indicator random variable, d is some positive constant and $g_y(x) = \min(x, y)$. We show that each of these sums is asymptotically normal; the asymptotic mean can be found simply by replacing $\bar{\xi}_n$ by μ, but the asymptotic variance can not. Applications of these results are then mentioned. The first sum above can be applied to demonstrating convergence of $\sum_{i=1}^{n}|\xi_i - \bar{\xi}_n|$, a statistic that may be used as a somewhat more robust estimate of scale than the sample variance. Let $Y_1, Y_2, \ldots, Y_n$ be non-negative independent identically distributed random variables from the one parameter family $P\{Y_i > x\} = F(x/\theta)$. F is some known distribution function, which may behave quite regularly (e.g. is exponential) for most values of x, except for some probability mass a long distance from the mean, usually as the result of the presence of outliers. One way of attempting to identify and reduce the noisy effect of these outliers is by Winsorizing all sample points more than a certain distance from the sample mean; i.e. by replacing each Y_i

Received by the editors September 19, 1977 and, in revised form, April 19, 1978.
(1) Work done at the Centre de Recherches Mathematiques, Montreal and supported by a grant from the Canadian Mathematical Congress.

by $g_{d\bar{Y}_n}(Y_i)$. This seems in some cases to be a clearer and more efficient way of identifying outliers than the usual practice of Winsorizing a fixed proportion of the observations.

The first sum is also applied to random sets generated in the following way. Let $x_1, x_2, \ldots, x_{n-1}$ be the order statistics from a uniform sample of size $n-1$ on $[0, 1]$. Suppose each X_i defines a ball of centre X_i and radius $d/2n$. Then the total coverage (i.e. the measure of the union of the balls) in the unit interval is $1-\sum_{i=0}^{n-1}(x_{i+1}-x_i-d/n)I(x_{i+1}-x_i>d/n)$, where, for convenience, we take $x_0=0$, $x_n=1$ and consider these points as covered by a ball as well.*

It is well-known that $(x_1, x_2, \ldots, x_{n-1})$ has the same joint distribution as $(s_1/s_n, s_2/s_n, \ldots, s_{n-1}/s_n)$ where $s_i=\sum_{i=1}^{i}\xi_i$ and the ξ_i are independent exponentially distributed random variables. Therefore the above coverage has the same distribution as the variable

$$1-\frac{1}{s_n}\sum_{i=1}^{n}(\xi_i-d\bar{\xi}_n)I(\xi_i>d\bar{\xi}_n). \tag{1.1}$$

The density of this variable was found by Votaw (1946) for finite n and Londhe (1976) found moments and attempted to prove asymptotic normality. Using different methods than his, we show in this paper that convergence to normality holds. This variable may arise in practice in military applications (e.g. bombs are dropped at random points along a railway line) or in biological ones (e.g. seeds are distributed at random points along a trench of length l). If each plant eventually covers an interval of length dl, then $1/s_n\sum_{i=1}^{n}(\xi_i-d\bar{\xi}_n)I(\xi_i>d\bar{\xi}_n)$ represents the proportion of the total trench that remains uncovered). For this latter example, it may be useful to permit the length of the covered segment d to depend on its position ($x_i\approx i/n$) because of variation in fertility. This justifies our subscripting d as d_{ni} in the results.

2. **Results.** Our main theorem is the following:

2.1. THEOREM. *Let $\xi_1, \xi_2, \ldots \xi_n \cdots$ be* i.i.d. *positive random variables with mean* 1, $F(x)=P(\xi_1>x)$ *and* $\int_0^\infty xF(x)\,dx<\infty$. *For any* $0\le u\le v<\infty$ *define* $\mu(u)=\int_u^\infty F(x)\,dx$ *and*

$$b(u, v)=2\int_v^\infty tF(t)\,dt-\mu(v)\{u+v+\mu(u)\}.$$

Let $\{d_{ni}; i=1, 2, \ldots, n, n=1, 2, \ldots\}$ be a triangular array of positive constants for which the following conditions hold;

(a) There exists a $k<\infty$ and a neighbourhood of $\{d_{n,i}; i\le n, n\ge k\}$ on which $F(x)$ is continuous.

* The results will hold with or without including these endpoints.

https://doi.org/10.4153/CMB-1979-051-x Published online by Cambridge University Press

(b) With $t_n = 1/n \sum_{i=1}^{n} d_{ni}F(d_{ni})$ we have that the sequence

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}\{b(d_{ni}, d_{ni}) - 2t_n b(0, d_{ni})\} + t_n^2 b(0,0) \to \sigma^2 \quad \text{as} \quad n \to \infty.$$

(c) $\max_{i \le n} d_{ni} = o(n)$.

(d) $1/n \sum_{i=1}^{n} r_{ni}$ is uniformly bounded.

Then the sum

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{(\xi_i - d_{ni}\bar{\xi}_n)I(\xi_i > d_{ni}\bar{\xi}_n) - \mu(d_{ni})\} \tag{2.1.1}$$

is asymptotically distributed as a normal $(0, \sigma^2)$ random variable.

Proof. We rewrite (2.1.1) as the following sum:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\xi_i - d_{ni}\bar{\xi}_n)I(\xi_i > d_{ni}\bar{\xi}_n) = \frac{1}{\sqrt{n}}\sum(\xi_i - d_{ni})I(\xi_i > d_{ni}) + \mathrm{M}_n t_n + \mathrm{M}_n U_n + V_n$$

where

$$\mathrm{M}_n = \sqrt{n}(1 - \bar{\xi}_n)$$

$$t_n = \frac{1}{n}\sum_{i=1}^{n} d_{ni}F(d_{ni})$$

$$U_n = \frac{1}{n}\sum_{i=1}^{n} d_{ni}\{I(\xi_i > d_{ni}) - F(d_{ni})\}$$

$$V_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\xi_i - d_{ni}\bar{\xi}_n)\{I(\xi_i > d_{ni}\bar{\xi}_n) - I(\xi_i > d_{ni})\}.$$

For some $\varepsilon > 0$ to be specified later, set $A_n = \{\omega;\ |\bar{\xi}_n - 1| < \varepsilon\}$; $B_{ni} = \{\omega;\ \xi_i$ is between d_{ni} and $d_{ni}\bar{\xi}_n\}$.

We have

$$\int_{A_n}|V_n|\,dP \le \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_{A_nB_{ni}}|\xi_i - d_{ni}\bar{\xi}_n|\,dP$$

$$\le \frac{1}{n}\sum_{i=1}^{n} d_{ni}\int_{A_nB_{ni}}\sqrt{n}\,|\bar{\xi}_n - 1|\,dP$$

It is easy to show that M_n is uniformly integrable for indeed $E\mathrm{M}_n^2$ is bounded. Moreover, $P(A_nB_{ni}) \le F(d_{ni}(1-\varepsilon) - F(d_{ni}(1+\varepsilon))$. Since F is continuous in an open interval containing all the d_{ni}, there is a closed subset S of this region containing the d_{ni} as interior points. In addition for arbitrary $\eta > 0$, we can find a finite k such that $F(k) < \eta$. Then F is uniformly continuous on the compact region $S \cap [0, k]$ so that for a sufficiently small ε,

$$\sup_i[F(d_{ni}(1-\varepsilon)) - F(d_{ni}(1+\varepsilon))] < \eta$$

https://doi.org/10.4153/CMB-1979-051-x Published online by Cambridge University Press

This, the uniform integrability of M_n and the uniform boundedness of $1/n \sum d_{ni}$ insures that $\int_{A_n} |V_n|\, dP$ can be made arbitrarily small by choosing ε sufficiently small. Moreover, since $P(A_n) \to 1$ for any ε, we have shown that $V_n \to^P 0$.

Note also that

$$EU_n^2 = \frac{1}{n^2} \sum_{i=1}^{n} d_{ni}^2 F(d_{ni})[1 - F(d_{ni})]$$

$$\leq \frac{\max_i |d_{ni}|}{n} \cdot t_n$$

which converges to 0. Therefore $M_n U_n \to^P 0$.

Finally, since t_n is uniformly bounded and for any positive d,

$$\int_{(\xi_i - d)^2 > c} (\xi_i - d)^2\, dP \leq \int_{\xi_i^2 > c} \xi_i^2\, dP \to 0 \quad \text{as} \quad c \to \infty,$$

the Lindeberg condition must hold for the array

$$\frac{1}{\sqrt{n}} (\xi_i - d_{ni}) I(\xi_i > d_{ni}) + \frac{t_n}{\sqrt{n}} (1 - \xi_i)$$

and therefore convergence to the $N(0, \sigma^2)$ distribution follows, from the Lindeberg–Feller central limit theorem. Q.E.D.

2.2 COROLLARY. *Let* $g_d(x) = \min(x, d)$. *Then under the conditions of Theorem* 2.1 *with* σ^2 *replaced by the assumed existent limit*

$$\sigma_2^2 = \lim_{n \to \infty} (1 + t_n)^2 b(0, 0) - \frac{2(1 + t_n)}{n} \sum_{i=1}^{n} b(0, d_{ni}) + \frac{1}{n} \sum_{i=1}^{n} b(d_{ni}, d_{ni}),$$

we have that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{g_{d_{ni}\bar{\xi}_n}(\xi_i) + \mu(d_{ni})\} - \sqrt{n}$$

converges to a normal, $N(0, \sigma_2^2)$ *variate.*

Proof.

$$\sum_{i=1}^{n} \xi_i = \sum_{i=1}^{n} (\xi_i - d_{ni}\bar{\xi}_n) I(\xi_i > d_{ni}\bar{\xi}_n) + \sum_{i=1}^{n} g_{d_{ni}\bar{\xi}_n}(\xi_i).$$

Therefore, as in Theorem 2.1

$$\frac{1}{\sqrt{n}} \sum g_{d_{ni}\bar{\xi}_n}(\xi_i) = \sqrt{n}\bar{\xi}_n - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\xi_i - d_{ni}) I(\xi_i > d_{ni}) - t_n M_n + o_p(1)$$

where by $o_p(1)$ we mean a sequence of random variables converging in

probability to zero. Adding $1/\sqrt{n}\sum_{i=1}^{n}\mu(d_{ni})-\sqrt{n}$ to both sides gives

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{(1+t_n)(\xi_i-1)-(\xi_i-d_{ni})I(\xi_i>d_{ni})+\mu(d_{ni})\}+o_p(1)$$

which clearly converges to the $N(0,\sigma_2^2)$ distribution.

3. **Applications.** Let ξ_i, $i=1,2,\ldots$ be i.i.d. random variables (not necessarily positive) with mean μ, variance σ^2. Let $\sigma_3^2=\mathrm{Var}\{(\xi-\mu)[I(\xi>\mu)-P(\xi>\mu)]\}$. Then as in Theorem 2.1, if $F(x)$ is continuous at μ (which we assume w.l.o.g. to be 1),

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}&(|\xi_i-\bar{\xi}_n|-E|\xi_i-1|)\\&=\frac{2}{\sqrt{n}}\sum_{i=1}^{n}(\xi_i-\bar{\xi}_n)I(\xi_i>\xi_n)-\mu(1)\\&=\frac{2}{\sqrt{n}}\sum_{i=1}^{n}\{(\xi_i-1)I(\xi_i>1)-\mu(1)-F(1)(\xi_i-1)\}+o_p(1)\end{aligned}$$

Therefore, this converges to the normal $(0,4\sigma_3^2)$ distribution.

Random sets. Let $D(x)$ be a non-negative Riemann integrable function on $[0,1]$ and $F(x)=e^{-x}$. In this case $\mu(u)=e^{-u}$ and $b(u,v)=e^{-v}(2-e^{-u})$ for $0\le u\le v$. Therefore if $d_{ni}=D(i/n)$ for $i=1,2,3,\ldots,$

$$t_n=\frac{1}{n}\sum_{i=1}^{n}D\left(\frac{i}{n}\right)e^{-D(i/n)}\to t=\int_0^1 D(x)e^{-D(x)}\,dx.$$

Therefore,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\xi_i-D\left(\frac{i}{n}\right)\bar{\xi}_n\right)I\left(\xi_i>D\left(\frac{i}{n}\right)\bar{\xi}_n\right)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}e^{-D(i/n)}$$

converges to $N(0,\sigma^2)$ distribution with

$$\sigma^2=(2-2t)\int_0^1 e^{-D(x)}\,dx-\int_0^1 e^{-2D(x)}\,dx-t^2.$$

Convergence of the total coverage,

$$\sqrt{n}\sum_{i=1}^{n}\;x_{i+1}-x_i-\frac{D\left(\frac{i}{n}\right)}{n}\,I\left(x_{i+1}-x_i>\frac{D\left(\frac{i}{n}\right)}{n}\right)$$

follows from this and (1.1).

Optimal Winsorizing. Suppose F is known and i.i.d. random variables $x_1,x_2,\ldots,x_n$ are observed with the same distribution as $\theta\xi_1$ where θ is some unknown parameter. Our purpose is to estimate the mean $Ex_1=\theta$ when F is

https://doi.org/10.4153/CMB-1979-051-x Published online by Cambridge University Press

such that Winsorizing may improve efficiency (consider for example the distribution $F(x)=\int_x^\infty (0.9)f(y)+(0.1)g(y)\,dy$ where $f(y), g(y)$ are the Gamma densities with $\alpha=1$, $\beta=\frac{1}{2}$, and $\alpha=55$, $\beta=0.1$ respectively. We may wish to use the consistent estimate

$$\frac{1}{n(1-\mu(d))}\sum_{i=1}^{n} g_{d\bar{x}_n}(x_i)$$

where d is a constant chosen to minimize the asymptotic variance

$$\operatorname{Var}\left\{\frac{(1+dF(d))\xi-(\xi-d)I(\xi>d)}{1-\mu(d)}\right\}$$
$$=\frac{(1+dF(d))^2 b(0,0)-2(1+dF(d))b(0,d)+b(d,d)}{(1-\mu(d))^2}.$$

As an example, we choose the mixture of exponential distributions defined by

$$F(x)=0.9e^{-2x}+0.1\min(e^{-(x-4.5)},1) \quad \text{for} \quad x>0,$$

$F(x)=1$ for $x<0$. In this case

$$\mu(d)=0.45e^{-2d}+0.1(5.5-d).$$

The asymptotic variance of the estimator above was computed for various values of d using a hand calculator ($Hp-67$) with the minimum at $d=1.63$, the corresponding variance 1.6079. The resulting asymptotic efficiency as compared with the sample mean (variance $=2.575$) is 160%. Moreover, the correct choice of d is not critical; values of d between 1 and 2.5 all yield asymptotic efficiencies of at least 145%.

References

1. A. R. Londhe (1976), *The Limiting Distribution of the Measure of A Random Set.* Thesis: U. of Alberta.

2. D. F. Votaw (1946), *The Probability Distribution of the Measure of A Random Linear Set.* Annals of Math. Statist. **17,** 240–244.

Department of Mathematics
The University of Alberta
Edmonton, Alberta T6G 2G1