Sankhyā : The Indian Journal of Statistics
2021, Volume 83-B, Supplement 1, pp. S26-S35

On the Exceedances of Exchangeable Random Variables

Satish Iyengar
University of Pittsburgh, Pittsburgh, USA

Abstract

Suppose that $\mathbf{X_n} = (X_1, \ldots, X_n)$ have mean 0, and a single-factor covariance $\Sigma = (\sigma_{ij})$ with $\sigma_{ii} = 1$ and $\sigma_{ij} = \rho \geq 0$ for $i \neq j$. For a threshold c, let S_n be the number of components of $\mathbf{X_n}$ that exceed c. We express the distribution of S_n in terms of a single integral, provide the limiting distribution as $n \to \infty$, and show that the limit resembles the Beta family. We then describe the shape of the exceedance distribution when the underlying distributions of the single-factor model have a certain likelihood ratio criterion with respect to its scale parameter, and we show that it obeys a majorization ordering.

AMS (2000) *subject classification.* Primary 62E10, 62H05.
Keywords and phrases. Exceedance, latent variables, likelihood ratio, Majorization.

1 Introduction

The problem of determining the probability of crossing a threshold by a random process has a wide range of applications, and a considerable history (Castillo et al., 2005; Leadbetter et al., 2011). The use of control charts or the study of k-of-n systems in reliability theory (Barlow and Proschan, 1965) are a classical example. Recent examples of considerable current interest include modeling flooding in hydrology (Huang et al., 2020) and the frequency of large forest fires (Alvarado et al., 1998) in climate science.

In many applications, (systems of) differential equations model the processes that underlie the exceedances. For example, stochastic differential equations with Brownian motion as the underlying driver in fields as diverse as mathematical finance (prices of equities) (Steele, 2000), hydrology (groundwater flow) (Cushman, 1987), and neuroscience (spike generation of a neuron) (Tuckwell, 1988). In other cases, the critical events are modeled as exceedances of rather simple probabilistic models, such as the multivariate normal.

Dedicated to Dr. C.R. Rao on the occasion of his 100th birthday.

The problem that we address here arose from a discussion with Professor Shun'ichi Amari about a paper by Shadlen and Newsome (1998) which dealt with the problem of modeling the natural variability of cortical neurons using a simple integrate-and-fire model. In this model, a neuron receives thousands of synaptic inputs. The magnitudes of the inputs could be viewed as i.i.d. Gaussian random variables, with positive (negative) values corresponding to excitatory (inhibitory). If the inputs were independent, the number of excitatory inputs would have a binomial distribution. However, the independence of the inputs is often not reasonable, so a natural generalization is to consider exchangeable random variables instead: the inputs may be dependent, but they are sense stationary in some sense. In short, we are interested in the structure of the probabilities of exceedances.

In Section 2 we define the exceedance statistic and its distribution. Although our primary interest is on an underlying Gaussian model, we state some of our results more generally. We show that the exceedance probabilities are expressible in terms of a single integral. We use it to describe the different shapes of the exceedance distribution and prove a majorization ordering for it.

2 Properties of the Exceedance Distribution

Consider i.i.d. symmetric random variables $Z_0, \ldots, Z_n$ with cdf F, pdf f, with $f(x) = f(-x) > 0$ or all x, and $E(Z_i^2) = 1$. Many of the results below are for the Gaussian, for which we denote the cdf and pdf by Φ and ϕ, respectively. For $\rho \geq 0$, let $X_i = \sqrt{1-\rho}Z_i + \sqrt{\rho}Z_0$, so that $\mathbf{X}_n = (X_1, \ldots, X_n)$ has mean 0 and covariance matrix $\Sigma = (\sigma_{ij})$ with $\sigma_{ii} = 1$ and $\sigma_{ij} = \rho$ for $i \neq j$. For a constant c the exceedance statistic is

$$S_n = \sum_{i=1}^{n} I(X_i \geq c).$$

By conditioning on Z_0, we get the exceedance distribution,

$$p_{n,k}(c|\rho) = P(S_n = k) = \binom{n}{k} \int_{-\infty}^{\infty} F\left(-\frac{c + t\sqrt{\rho}}{\sqrt{1-\rho}}\right)^{k} F\left(\frac{c + t\sqrt{\rho}}{\sqrt{1-\rho}}\right)^{n-k} f(t)\, dt. \tag{2.1}$$

Several special cases of this expression are elementary (David, 1953): for example, for the Gaussian

$$p_{2,2}(0|\rho) = \frac{1}{2} - \frac{1}{2\pi}\cos^{-1}\rho \quad \text{and} \quad p_{3,3}(0|\rho) = \frac{1}{2} - \frac{3}{4\pi}\cos^{-1}\rho, \tag{2.2}$$

which also hold for $-1/(n-1) \leq \rho < 0$; these expressions also hold for elliptically contoured distributions (Iyengar and Tong, 1989). And for any such F,

$$p_{n,k}(0|1/2) = \frac{1}{n+1}, \tag{2.3}$$

is the probability that Z_0 is the $(n-k+1)^{\mathrm{st}}$ order statistic among $(Z_0, \ldots, Z_n)$. More generally, the connection to order statistics of equicorrelated random variables is clear: writing $F(-x) = 1 - F(x)$, we have

$$\begin{aligned} p_{n,k}(c|\rho) &= \binom{n}{k} \int_{-\infty}^{\infty} F\left(-\frac{c+t\sqrt{\rho}}{\sqrt{1-\rho}}\right)^{k} F\left(\frac{c+t\sqrt{\rho}}{\sqrt{1-\rho}}\right)^{n-k} f(t)\,dt \\ &= \binom{n}{k} \sum_{i=0}^{k} (-1)^{k-i} \binom{k}{i} \int_{-\infty}^{\infty} F\left(\frac{c+t\sqrt{\rho}}{\sqrt{1-\rho}}\right)^{n-i} f(t)\,dt \qquad (2.4) \\ &= \binom{n}{k} \sum_{i=0}^{k} (-1)^{k-i} \binom{k}{i} P(M_{n-i} \leq c), \end{aligned}$$

where M_n is the largest order statistic in an equicorrelated sample of size n.

For large n there is an easily derived approximation.

THEOREM 2.1. *For $0 \leq t \leq 1$, as $n \to \infty$*

$$P\left(\frac{S_n}{n} < t\right) \to F\left(\frac{c+\sqrt{1-\rho}F^{-1}(t)}{\sqrt{\rho}}\right), \tag{2.5}$$

or with a slight abuse of notation,

$$F^{-1}\left(\frac{S_n}{n}\right) \xrightarrow{d} \frac{c}{\sqrt{1-\rho}} + \sqrt{\frac{\rho}{1-\rho}} Z_0.$$

PROOF. By the strong law,

$$\frac{1}{n}\sum_{i=1}^{n} I\left(Z_i \geq \alpha\right) \to F(-\alpha) \ \text{ almost surely.}$$

Thus, by dominated convergence

$$\begin{aligned} P\left(\frac{S_n}{n}<t\right) &= \int_{\mathbb{R}}\left(\frac{S_n}{n}<t|Z_0=u\right) f(u)\,du \\ &= \int_{\mathbb{R}} P\left(\frac{1}{n}\sum_{i=1}^{n}\left[Z_i \geq \frac{c+u\sqrt{\rho}}{\sqrt{1-\rho}}\right]<t\right) f(u)\,du \\ &\to \int_{\mathbb{R}} I\left[F\left(-\frac{c+u\sqrt{\rho}}{\sqrt{1-\rho}}\right)<t\right] f(u)\,du \\ &= P\left[Z_0 > -\frac{c+\sqrt{1-\rho}F^{-1}(t)}{\sqrt{\rho}}\right] \\ &= F\left(\frac{c+\sqrt{1-\rho}F^{-1}(t)}{\sqrt{\rho}}\right) \end{aligned}$$

For the Gaussian, the following properties of $G_{a,b}(t) = \Phi(a + b\Phi^{-1}(t))$ for $a \in \mathbb{R}$ and $b > 0$, and its density

$$g_{a,b}(t) = \frac{b\phi(a+b\Phi^{-1}(t))}{\phi(\Phi^{-1}(t))} \tag{2.6}$$

are easy to verify.

(a) $G_{0,1}$ is the uniform.

(b) If $b = 1$ and $a > 0$ ($a < 0$), the density decreases (increases) from ∞ to 0 (0 to ∞).

(c) If $b > 1$ the density is bounded with $g_{a,b}(0) = g_{a,b}(1) = 0$ and it is unimodal with mode at $t = \Phi(ab/(1-b^2))$.

(d) If $b < 1$ the density is bounded with $g_{a,b}(0) = g_{a,b}(1) = \infty$ and it is U-shaped with minimum at $t = \Phi(ab/(1-b^2))$.

(e) The raw moments of this distribution are

$$\int_0^1 t^k g_{a,b}(t)\,dt = \int_{\mathbb{R}} \Phi\left(\frac{x-a}{b}\right)^k \phi(x)\,dx;$$

The first moment is $1-\Phi(a/\sqrt{1+b^2})$; the rest are easily computed, but not expressible in elementary terms.

Thus, the family (2.6) of limiting distributions resembles the Beta(α,β) family. For the special case $c = 0$ that resemblance holds for finite n for not only the Gaussian, but other latent distributions that satisfy a certain likelihood ratio ordering with respect to the scale parameter.

Definition 2.2. We say that the cdf F and its pdf f satisfy the LR condition if $f(x) = f(-x)$ and that the ratio

$$L(x|\sigma) = \frac{1}{\sigma}\frac{f(x/\sigma)}{f(x)}$$

is decreasing (increasing) in $|x|$ for $0 < \sigma \leq 1$ $(1 \leq \sigma < \infty)$. The Gaussian, Laplace, and t-distributions all satisfy this LR condition.

We next prove the intuitively clear result that for $c = 0$ the exceedance distribution is either U-shaped or unimodal with mode or minimum at the middle. The proof is rather involved, requiring a detailed study of the integrands and the use of the likelihood ratio method which transfers attention from any ρ to $\rho = 1/2$, for which the exceedance distribution is uniform from (2.3).

Theorem 2.3. *Suppose that F and f satisfy the LR condition in (2.2), and that $c = 0$, so that the exceedance distribution is symmetric. Then for $0 \leq \rho \leq 1/2$ the exceedance distribution is unimodal with probabilities decreasing away from the mode. And for $1/2 \leq \rho \leq 1$ it is U-shaped with probabilities increasing away from the minimum. The mode or minimum is at $n/2$ for n even and at $(n \pm 1)/2$ for n odd.*

Proof. We prove this result for $0 \leq \rho \leq 1/2$; the the proof for $1/2 \leq \rho \leq 1$ is similar, so we omit it. We first show that if $0 \leq \rho \leq 1/2$, then $p_1^{(n)} - p_0^{(n)} \geq 0$, and that for $2 \leq k \leq (n+1)/2$ we have $p_k^{(n)} - p_{k-1}^{(n)} \geq 0$. Let $\alpha = \sqrt{\rho/(1-\rho)}$. Then

$$\begin{aligned} p_1^{(n)} - p_0^{(n)} &= n\int_{-\infty}^{\infty} F(-\alpha t)F(\alpha t)^{n-1}f(t)\,dt - \int_{-\infty}^{\infty} F(-\alpha t)^n f(t)\,dt \\ &= \int_{-\infty}^{\infty} [nF(-\alpha t)^{n-1} - (n+1)F(t)^n]L(t|\alpha)f(t)\,dt \qquad (2.7) \\ &= \int_0^{\infty} [g_n(F(t)) + g_n(F(-t))]L(t|\alpha)f(t)\,dt \\ &= \int_0^{\infty} h_n(F(t))L(t|\alpha)f(t)dt, \end{aligned}$$

where $g_n(y) = ny^{n-1} - (n+1)y^n$ and $h_n(y) = g_n(y) + g_n(1-y)$. We need the following facts, all of which are derived by a close examination of these polynomials. First, g_n has roots at 0 and $n/(n+1)$; its boundary values are $g_n(0) = 0$, $g_n(1) = -1$, $g_n'(0) = 0$, and $g_n'(1) = -n$; g_n has its maximum value at $y = (n-1)/(n+1)$. Next, h_n is strictly positive for $1/2 \leq y \leq n/n+1$, and is strictly decreasing for $n/(n+1) \leq y \leq 1$; thus, there is a unique y^* between $n/(n+1)$ and 1 such that $h_n(y^*) = 0$. Now let $F(t^*) = y^*$. Then

$$\begin{aligned} p_1^{(n)} - p_0^{(n)} &= \int_0^\infty h_n(F(t))L(t|\alpha)f(t)\,dt \\ &= \int_0^{t^*} h_n(F(t))L(t|\alpha)f(t)\,dt + \int_{t^*}^\infty h_n(F(t))L(t|\alpha)f(t)\,dt \qquad (2.8) \\ &\geq L(t^*|\alpha)\left[\int_0^{t^*} h_n(F(t))f(t)\,dt + \int_{t^*}^\infty h_n(F(t))f(t)\,dt\right] \\ &= L(t^*|\alpha)\int_0^\infty h_n(F(t))f(t)\,dt = 0. \end{aligned}$$

The proof of $p_k^{(n)} - p_{k-1}^{(n)} \geq 0$ for $2 \leq k \leq (n+1)/2$ and $0 \leq \rho \leq 1/2$ is similar, but the functions corresponding to g_n and h_n are more involved. Start with

$$\begin{aligned} p_k^{(n)} - p_{k-1}^{(n)} &= \binom{n}{k}\int_{-\infty}^\infty F(-\alpha t)^k F(\alpha t)^{n-k} f(t)dt \\ &\quad -\binom{n}{k-1}F(-\alpha t)^{k-1}F(\alpha t)^{n-k+1}f(t)\,dt \\ &= \binom{n}{k}\int_0^\infty h_{n,k}(F(t))L(t|\alpha)f(t)\,dt, \qquad (2.9) \end{aligned}$$

where

$$g_{n,k}(y) = (1-y)^{k-1}y^{n-k}\left(1 - \frac{n+1}{n-k+1}y\right),$$

and

$$h_{n,k}(y) = g_{n,k}(y) + g_{n,k}(1-y) \quad \text{for } \frac{1}{2} \leq y \leq 1.$$

Note that $g_{n,k}$ has roots at 0, 1, and $(n-k+1)/(n+1)$, and that $g'_{n,k}$ has roots at

$$\frac{n-k+1}{n+1} \pm \frac{\sqrt{k[1-(k-1)/n]}}{n+1}.$$

Next, we show that $h_{n,k}$ has a unique root y^* between 1/2 and 1, with $h_{n,k}$ positive (negative) to the left (right) of y^*, so we can then use the same proof as before. Using the properties of $g_{n,k}$, we see that $h_{n,k}$ is strictly positive for $1/2 \leq y \leq (n-k+1)/(n+1)$, and strictly decreasing in the interval

$$\left[\frac{n-k+1}{n+1}, \frac{n-k+1}{n+1} + \frac{1}{n+1}\sqrt{k\left(1-\frac{k-1}{n}\right)}\right] = [L_{n,k}, U_{n,k}].$$

We must therefore show that $h_{n,k} \leq 0$ in $[U_{n,k}, 1]$. This is clearly true for $y = 1$; for $y < 1$ we have

$$h_{n,k}(y) \leq 0 \iff \left(\frac{y}{1-y}\right)^{n-2k+1} \geq 1 + \frac{n-2k+1}{(n+1)y-(n-k+1)}.$$

Since $n \geq 2k+1$, it suffices to show that $h_{n,k}(y) \leq 0$ in the interval

$$\left[\frac{n-k+1}{n+1} + \frac{\sqrt{k/2}}{n+1}, 1\right].$$

In this interval we have

$$\left(\frac{y}{1-y}\right)^{n-2k+1} \geq \left(\frac{n-k+1+\sqrt{k/2}}{k-\sqrt{k/2}}\right)^{n-2k+1}$$

because the function on the left is strictly increasing in y. Thus, it is now enough to show that

$$1 + \frac{n-2k+1}{\sqrt{k/2}} \leq \left(\frac{n-k+1+\sqrt{k/2}}{k-\sqrt{k/2}}\right)^{n-2k+1}$$

for $k \geq 2$ and $n \geq 2k-1$. This inequality is trivial for $n = 2k-1, 2k, 2k+1$. Finally, writing $u = n-2k+1$, we must verify that for all $u \geq 0$,

$$1 + \frac{u}{\sqrt{k/2}} \leq \left(\frac{u+k+\sqrt{k/2}}{k-\sqrt{k/2}}\right)^{u},$$

which is an easy (if tedious) verification, and our proof is complete. A small note: the details of this proof requires $n \geq 3$; the result also holds for $n = 2$ using the expressions Eq. 2.1.

Our next result concerns majorization properties that the exceedance distribution: see Marshall and Olkin (1979). Let $x, y \in \mathbb{R}^n$ be nonincreasing sequences of numbers; that is, $x_1 \geq x_2 \geq \cdots \geq x_n$, and similarly for y. Then x majorizes y, written $x \succ y$, if for $k = 1, \ldots, n$

$$\sum_{i=1}^{k} x_k \geq \sum_{i=1}^{k} y_k \quad \text{and} \quad \sum_{i=1}^{n} x_k = \sum_{i=1}^{n} y_k.$$

Theorem 2.4. *As $n \to \infty$, Suppose that the cdf F and its density f satisfy the LR condition, that $c = 0$, and $p(\rho) = (p_0(\rho), \dots, p_n(\rho))$ be the exceedance distribution. If $0 \leq \rho_1 < \rho_2 \leq 1/2$, then $p(\rho_1) \succ p(\rho_2)$; and if $1/2 \leq \rho_1 < \rho_2 \leq 1$, then $p(\rho_2) \succ p(\rho_1)$;*

Proof. We prove this result for $n = 2m+1$ is odd and $0 \leq \rho \leq 1/2$; the proof for even n and $1/2 \leq \rho \leq 1$ is similar. Because $c = 0$, $p_{n,i}(\rho) = p_{n,2m+1-i}(\rho)$, and

$$p_{n,0}(\rho) \leq \cdots \leq p_{n,m}(\rho) = p_{n,m+1}(\rho) \geq \cdots \geq p_{n,2m+1}(\rho) \quad \text{for } 0 \leq \rho \leq 1/2.$$

Thus, it suffices to show that the functions $p_{n,m}$, $2p_{n,m}$, $2p_{n,m} + p_{n,m-1}$, $\dots$, $2p_{n,m} + 2p_{n,m-1}, \dots$, all decrease with ρ. To do this, we will show that all the derivatives are negative. As before, let $\alpha = \sqrt{\rho/(1-\rho)}$, and note that α is a strictly increasing function of ρ. Thus, for $j < m$, we have

$$\begin{aligned} H(\alpha) &= \sum_{i=0}^{j} p_{n,m-i}(\rho) = \sum_{i=0}^{j} \binom{2m+1}{m-i} \int_{\mathbb{R}} F(-\alpha t)^{m-i} F(\alpha t)^{m+i+1} f(t)\, dt \\ &= \binom{2m+1}{m} \sum_{i=0}^{j} \frac{m!(m+1)!}{(m-i)!(m+i+1)!} \int_{\mathbb{R}} F(-\alpha t)^{m-i} F(\alpha t)^{m+i+1} f(t)\, dt. \end{aligned}$$

Next, writing $h(t) = t\phi(t)\phi(\alpha t)$, and dropping the constant combinatorial coefficient

$$\binom{2m+1}{m} m!(m+1)!,$$

the derivative of H is

$$\begin{aligned} H'(\alpha) &\propto \int_{\mathbb{R}} \left[\sum_{i=0}^{j} \frac{F(-\alpha t)^{m-i} F(\alpha t)^{m+i}}{(m-i)!(m+i+1)!} - \sum_{i=0}^{j} \frac{F(-\alpha t)^{m-i-1} F(\alpha t)^{m+i+1}}{(m-i)!(m+i+1)!} \right] h(t)\, dt \\ &= \int_{\mathbb{R}} k(t)h(t)dt. \end{aligned}$$

Note that because $h(t)$ is an odd function, we have

$$\int_{\mathbb{R}} k(t)h(t)dt = \int_0^{\infty} [k(t) - k(-t)]h(t)dt.$$

Applying that to the expression for $H'(\alpha)$ we get a collapsing sum that leads to

$$H'(\alpha) = -C(m,j)\int_0^\infty [F(-\alpha t)F(\alpha t)]^{m-j-1}[F(\alpha t)^{2j+2} - F(-\alpha t)^{2j+2}]h(t)dt < 0,$$

where

$$C(m,j) = (m+1)\binom{2m+1}{m+j+1}\binom{2m}{m}^{-1}.$$

We now see that $\sum_{i=1}^{j} p_{n,m-i}(\rho)$ is decreasing in ρ for $j < m$; the case of $j = m$ is trivial because $\sum_{i=1}^{m} p_{n,m-i}(\rho) = 1/2$. Finally, the same calculations show that the same result holds for

$$2\sum_{i=0}^{j} p_{n,m-i}(\rho) + p_{n,m-j-1}(\rho),$$

and our proof is complete.

3 Discussion

Numerical examples indicate that the approximation in Eq. 2.5 is good for $n \geq 20$ near the mode, but that $n \geq 50$ gives better results in the tails. Of course, for the neuroscience applications that motivated this work the approximation is quite good because n is in the thousands. Our main results – the shape of the exceedance distribution and the the majorization result – are limited in scope because $c = 0$. Extending these results to $c \neq 0$ requires knowledge of the location of the mode, complicating the computations considerably. Our numerical work indicate that the beta-distribution-like shapes may well hold for any c and $\rho \geq 0$. However, the majorization result does not generalize to $c \neq 0$.

References

ALVARADO, E, SANDBERG, DV and PICKFORD, SG (1998). Modeling large forest fires as extreme events. *Northwest Sci.* **72**, 66–75.

BARLOW, RE and PROSCHAN, F (1965). *Mathematical Theory of Reliability.* Wiley, Hoboken.

CASTILLO, E, HADI, AS, BALAKRISHNAN, N and SARABIA, JM (2005). *Extreme Value and Related Models with Applications in Engineering and Science.* Wiley, Hoboken.

CUSHMAN, JH (1987). Development of stochastic partial differential equations for subsurface hydrology. *Stochastic Hydrol. Hydraulics* **1**, 241–262.

DAVID, FN (1953). A note on the evaluation of the multivariate normal integral. *Biometrika* **40**, 458–459.

HUANG, Y, LIANG, Z, HU, Y, LI, B and WANG, J (2020). Theoretical derivation for the exceedance probability of corresponding flood volume of the equivalent frequency regional composition method in hydrology. *Hydrol. Res.* **51**, 1274–1292.

IYENGAR, S and TONG, YL (1989). Convexity of elliptically contoured distributions with applications. *Sankhya A* **51**, 13–29.

LEADBETTER, ML, LINDGREN, G and ROOTZÉN, H (2011). *Extremes and Related Properties of Random Sequences and Processes.* Springer, Berlin.

MARSHALL, A and OLKIN, I (1979). *Inequalities: Theory of Majorization and its Applications.* Academic Press, Cambridge.

SHADLEN, MN and NEWSOME, WT (1998). The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. *J Neuroscience* **18**, 3870–3896.

STEELE, JM (2000). *Stochastic Calculus and Financial Applications.* Springer, Berlin.

TUCKWELL, HC (1988). *Introduction to Theoretical Neurobiology.* Cambridge University Press, Cambridge.

Publisher's Note. Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Satish Iyengar
Department of Statistics,
University of Pittsburgh,
Pittsburgh, PA, USA
E-mail: ssi@pitt.edu

Paper received: 30 December 2020; accepted 6 January 2021.