A NOTE ON THE ANALYTIC APPROXIMATION OF EXCEEDANCE PROBABILITIES IN HETEROGENEOUS POPULATIONS

By

H. S. Battey

Department of Mathematics, Imperial College London, SW7 2AZ, UK

SUMMARY

This note derives analytic approximations to exceedance probabilities for order statistics from two heterogeneous populations. A limitation of this approach is that it entails a special condition that needs to be checked or justified on a case by case basis.

Some key words: Analytic approximations; Competing risks, Extremal-Types Theorem, Probabilistic screening.

1 Introduction

Suppose that $X_1, \dots, X_{N_X}$ are N_X independent copies of the random variable X, whose distribution function is F_X, and $Z_1, \dots, Z_{N_Z}$ are N_Z copies of Z, whose distribution function is F_Z. Both F_X and F_Z are assumed continuous. It is not known which of the $N_X + N_Z$ realizations of these random variables are from the X and Z populations. The following quantities arise in diverse contexts:

(i) the probability that the set of largest $s \geq N_X$ observations contains all N_X observations from the X population;

(ii) the probability that the minimum of the X population exceeds the rth largest observation from the Z population;

(iii) the probability that the maximum of the $N_X + N_Z$ observations belongs to the X population.

Exact solutions for such probabilities are rarely available, so that reliance is often on numerical approximations, making dependence on key aspects inexplicit.

The first two questions emerge naturally in screening-type problems. Suppose the X population represents individuals with a particular disease, which can be assessed accurately by a costly, or otherwise inconvenient, procedure. To better target this resource, an initial screening is performed. Thus in this context X and Z may represent, for instance, systolic blood pressure, concentrations of toxins or solutes in the blood, etc. for individuals in the diseased and healthy group. The first question above indicates the likely success of the screening at detecting all individuals in the diseased group. The second quantifies the number of false positives entailed in the identification of all N_X cases. Similar issues arise in statistical contexts, where it is often desirable to screen a large number of potentially explanatory variables, sometimes with a view to assessing causality more rigorously through a full factorial or other designed experiment.

A key unifying observation for addressing these questions is that there is no loss of generality by treating one of the two distributions as standard uniform. This is because the transformed random variable $U_i \triangleq 1 - F_Z(Z_i)$ is uniformly distributed on $[0, 1]$. Here and henceforth $\triangleq$ means equality by definition. On defining $V_i \triangleq 1 - F_Z(X_i)$ and modelling the density function of this random variable by $(1-\gamma)v^{-\gamma}$ for $\gamma < 1$, $0 \leq v \leq 1$, an approach that would require careful justification in any particular context, probability calculations of the nature outlined above can be expressed in terms of beta integrals, approximable in terms of elementary functions using Stirling's formula.

In the expression $(1-\gamma)v^{-\gamma}$, $\gamma = 0$ recovers the uniform density function and $\gamma \to 1$ and $\gamma \to -\infty$ represent strong departure from the uniform distribution in both directions.

Of course, $1 - F_Z(Z_i)$ is not the only transformation that delivers uniform random variables and there are settings in which one of the other three possibilities, $F_Z(Z_i)$, $1 - F_X(X_i)$, $F_X(X_i)$, may be more fruitful. We discuss this choice in the context of the examples given.

A referee has pointed out work by Bairamov and Parsi (2011) and Bayramoglu and Eryilmaz (2015), who study a very similar problem. Their results apply under weaker conditions than in the present paper, at the expense of more complicated analytic expressions. In particular, $X_1, \ldots, X_{N_X}$ and $Z_1, \ldots, Z_{N_Z}$ are treated as exchangeable random variables, independent of one another in the first paper and dependent in the second. While these two papers are the most closely related to the present work, there is an extensive literature studying the exact and asymptotic distributions of the number of exceedances based on order statistics. Notable early examples are Gumbel and von Schelling (1950) and Sarkadi (1957) who consider exceedence probabilities of order statistics from two samples of potentially different sizes drawn from the same population (i.e. $F_X = F_Z$). Certain special cases are recoverable both from their calculations and ours, as indicated below.

2 Exceedance probability formulae: examples

2.1 Competing maxima

We first address the most challenging of the questions specified in §1. This serves as an exemplar for the other cases. The probability that the maximum of $N_X + N_Z$ observations belongs to the X population is the probability that $X_{\max} \triangleq \max\{X_1, \ldots, X_{N_X}\}$ exceeds $Z_{\max} \triangleq \max\{Z_1, \ldots, Z_{N_Z}\}$. Since distribution functions are monotonically increasing,

$$p = \mathrm{pr}(X_{\max} > Z_{\max}) = \mathrm{pr}(V_{\min} < U_{\min}),$$

where $V_{\min} \triangleq \min\{V_1, \ldots, V_{N_X}\}$ and $U_{\min} \triangleq \min\{U_1, \ldots, U_{N_Z}\}$. Note that the density function of $U_{\min}$ at u is $N_Z(1-u)^{N_Z-1}$. Thus consider initially the event $A_i \triangleq \{V_i < U_{\min}\}$, with associated probability

$$\mathrm{pr}(A_i) = N_Z \int_0^1 v^{1-\gamma}(1-v)^{N_Z-1} dv.$$

This is of the form of a beta integral of indices $2-\gamma$ and N_Z. Using $x\Gamma(x) = \Gamma(x+1)$ and Stirling's formula in the form $\Gamma(x)/\Gamma(x+a) \simeq x^{-a}$ for large x and fixed a, we obtain

$$\mathrm{pr}(V_i < U_{\min}) \simeq \frac{\Gamma(2-\gamma)\Gamma(N_Z+1)}{\Gamma\{N_Z+1+(1-\gamma)\}} \simeq \frac{\Gamma(2-\gamma)}{(N_Z+1)^{(1-\gamma)}}, \quad (N_Z \to \infty).$$

Since $V_1, \ldots, V_{N_X}$ are independent and identically distributed, for any fixed v,

$$\mathrm{pr}(V_{i_1} < v, \ldots, V_{i_k} < v) = v^{k(1-\gamma)}$$

for an arbitrary set of k indices $i_1, \ldots, i_k$. It follows that there are similar approximations to the probabilities of all combinations of joint events, specifically

$$\mathrm{pr}(A_{i_1}, \ldots, A_{i_k}) \simeq \frac{\Gamma\{k(1-\gamma)+1\}}{(N_Z+1)^{k(1-\gamma)}}, \quad (N_Z \to \infty).$$

We conclude that

$$p = \mathrm{pr}\Big(\bigcup_{i=1}^{N_X} A_i\Big) \simeq \sum_{k=1}^{N_X} (-1)^{k-1} \binom{N_X}{k} \frac{\Gamma\{k(1-\gamma)+1\}}{(N_Z+1)^{k(1-\gamma)}}, \quad (N_Z \to \infty). \tag{1}$$

An analysis of convergence is given in the supplementary material. The approximation is essentially exact, the only error coming from the use of Stirling's formula. Indeed, Stirling's approximation to $\Gamma(x)$, while derived under the notional limiting operation $x \to \infty$ provides a remarkably accurate approximation even for small values of x. It is therefore reasonable to apply Stirling's formula to the numerators, giving the following approximation in terms of elementary functions:

$$p \approx \frac{\sqrt{2\pi}}{e} \sum_{k=1}^{N_X} (-1)^{k-1} \binom{N_X}{k} \frac{\{k(1-\gamma)+1\}^{k(1-\gamma)+\frac{1}{2}}}{e^{k(1-\gamma)}(N_Z+1)^{k(1-\gamma)}}. \tag{2}$$

Note that while the random variables $V_1, \ldots, V_{N_X}$ are independent, $\mathbb{1}\{V_1 < U_{\min}\}, \ldots, \mathbb{1}\{V_{N_X} < U_{\min}\}$ are not, due to their mutual dependence on $U_{\min}$. It is for this reason that the full inclusion-exclusion formula is needed in (1) rather than the simplification

$$\mathrm{pr}\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} (-1)^{i-1} \binom{n}{i} \eta^i = 1 - (1-\eta)^n,$$

applying to independent events E_i with probabilities $\eta = \mathrm{pr}(E_i)$ for all i.

Reversing the roles of X_i and Z_i in the construction of the uniform random variables leads to a simpler formula. Let $\widetilde{U}_i \triangleq F_X(X_i)$ and $W_i \triangleq F_X(Z_i)$ so that $\widetilde{U}_1, \ldots, \widetilde{U}_{N_X}$ are independent uniformly distributed random variables. On modelling the density function of the random variables $W_1, \ldots, W_{N_Z}$ as $(1-\gamma)w^{-\gamma}$, the probability p is now $\mathrm{pr}(\widetilde{U}_{\max} > W_{\max})$, where $\widetilde{U}_{\max} \triangleq \max\{\widetilde{U}_1, \ldots, \widetilde{U}_{N_X}\}$ with density function $N_X u^{N_X-1}$ at u, and $W_{\max} \triangleq$

$\max\{W_1, \ldots, W_{N_Z}\}$. The required probability is

$$p = \mathrm{pr}\Big(\bigcap_{i=1}^{n}\{\ W_i < \tilde{U}_{\max}\}\Big) = N_X \int_0^1 w^{N_Z(1-\gamma)+N_X-1}dw = \frac{N_X}{N_X + N_Z(1-\gamma)}. \quad (3)$$

The choice between which transformation to use, leading to either (1) or (3) depends on which of F_Z and F_X is known, or for which of the populations $V_1, \ldots, V_{N_X}$ or $W_1, \ldots, W_{N_Z}$ the density parameterization $(1-\gamma)v^{-\gamma}$ is most reasonable.

If γ is set to zero in equation (1) and the binomial coefficient is written in terms of gamma functions and simplified using Stirling's approximation, the right hand side of equation (1) is, for $N_Z > N_X$,

$$\sum_{k=1}^{N_X}(-1)^{k-1}\left(\frac{N_X+1}{N_Z+1}\right)^k = \frac{N_X+1}{(N_X+1)+(N_Z+1)}\left\{1+(-1)^{N_X+1}\left(\frac{N_X+1}{N_Z+1}\right)^{N_X+1}\right\}.$$

For large N_Z this is approximately $N_X/(N_X+N_Z)$ which is what one would obtain from direct calculation, noting that V_i and U_i are both uniformly distributed when $\gamma = 0$. This is equation (1.6) of Gumbel and von Schilling (1950). The formula (3) similarly becomes $N_X/(N_X+N_Z)$ when $\gamma = 0$. Also intuitively, for $\gamma = 1$, formula (1) reduces to

$$\sum_{k=1}^{N_X}(-1)^{k-1}\binom{N_Z}{k} = -\left\{\sum_{k=0}^{N_X}\binom{N_Z}{k} - 1\right\} = 1.$$

while for $\gamma \to -\infty$, it tends to zero. Similarly for formula (3).

2.2 Probabilities quantifying screening properties

As noted in §1, probabilities (i) and (ii) are relevant for assessing the properties of physical or abstract screening procedures. These questions lead to simpler calculations than that of §2.1 once the transformation to $U_i = 1 - F_Z(Z_i)$ and $V_i = 1 - F_Z(X_i)$ has been made.

We start by considering the probability that all members of the X population are discovered before the first falsely detected individual from the Z population. This is

$$\begin{aligned} \mathrm{pr}(U_{\min} > V_{\max}) &= \mathrm{pr}\Big(\bigcap_{i=1}^{N_X}\{V_i < U_{\min}\}\Big) \\ &= N_Z\int_0^1 v^{N_X(1-\gamma)}(1-v)^{N_Z-1}dv \simeq \frac{\Gamma\{N_X(1-\gamma)+1\}}{(N_Z+1)^{N_X(1-\gamma)}}, \quad (N_X \to \infty). \end{aligned}$$

The rth smallest order statistic $U_{(r)}$ constructed from $U_1, \ldots, U_{N_Z}$ is distributed as a beta random variable with indices r and $N_Z - r + 1$ so that its density function is given by

$$\frac{N_Z!}{(r-1)!(N_Z-r)!} v^{r-1}(1-v)^{N_Z-r}.$$

Probability (ii) is therefore addressed by

$$\begin{aligned} \mathrm{pr}(U_{(r)} > V_{\max}) &= \frac{N_Z!}{(r-1)!(N_Z-r)!} \int_0^1 v^{N_X(1-\gamma)+r-1}(1-v)^{N_Z-r} dv \\ &= \frac{N_Z!}{(r-1)!} \frac{\Gamma\{N_X(1-\gamma)+r\}}{\Gamma\{N_X(1-\gamma)+N_Z+1\}}, \end{aligned}$$

which can be further simplified using Stirling's formula. There are no technical difficulties in considering the generalization $\mathrm{pr}(U_{(r)} > V_{(s)})$ for $1 \leq r \leq N_Z$, $1 \leq s \leq N_X$.

3 Comparison to approximations based on limit theorems

In highly influential work, Fisher and Tippett (1928) characterized the set of probability laws $\mathcal{L}$ such that, for independently distributed random variables $X_1, \ldots, X_n$ with a common but arbitrary distribution function F, there exist sequences $a_X(n)$ and $b_X(n)$ such that

$$\lim_{n\to\infty} \mathrm{pr}\left(\frac{\max\{X_1,\ldots,X_n\} - b_X(n)}{a_X(n)}\right) = \lim_{n\to\infty} F^n\{a_X(n)x + b_X(n)\} = L(x), \quad L \in \mathcal{L}. \quad (4)$$

When equation (4) holds, the random variable X is said to be in the domain of attraction of L. The set $\mathcal{L}$ has just three elements. These are, in the notation of Fisher and Tippett (1928), for $\alpha > 0$:

$$\begin{aligned} \text{Type I} &: L(x) = \exp(-e^{-x}), \quad x \in \mathbb{R}; \\ \text{Type II} &: L(x) = L_\alpha(x) = \begin{cases} 0, & x \leq 0, \\ \exp(-x^{-\alpha}), & x > 0; \end{cases} \\ \text{Type III} &: L(x) = L_\alpha(x) = \begin{cases} \exp\{-(-x)^\alpha\}, & x < 0 \\ 1, & x \geq 0. \end{cases} \end{aligned}$$

Gnedenko (1943) proved that these are the only three types that can arise as limit laws. The scaling sequence $a_X(n)$ can be taken as 1 when $X_1, \dots, X_n$ are in the domain of attraction of the Type I limit law.

These limit laws are plausible approximations for the context of §2.1 provided that both N_X and N_Z are large. In the context of §2.1, suppose that both types of random variables $X_1, \dots, X_{N_X}$ and $Z_1, \dots, Z_{N_Z}$, are in the domain of attraction of the Type I limit law so that $a_Z(N_Z) = a_X(N_X) = 1$. The following calculation can be straightforwardly adapted for the eight other possible combinations. Let $G_Z(N_Z) \triangleq Z_{\max} - b_Z(N_Z)$ and $G_X(N_X) \triangleq X_{\max} - b_X(N_X)$ with density and distribution functions f_{G_Z}, f_{G_X}, F_{G_Z} and F_{G_X}. We will, for notational convenience, drop the arguments N_Z and N_Z. On writing G for a random variable with the standard Type I distribution, having density and distribution functions f_G and F_G, we have

$$\begin{aligned} p = \mathrm{pr}(Z_{\max} \leq X_{\max}) &= \int_{-\infty}^{\infty} \mathrm{pr}(G_Z \leq v + b_X - b_Z) f_{G_X}(v) dv \\ &= \int_{-\infty}^{\infty} \{\mathrm{pr}(G_Z \leq v + b_X - b_Z) - \mathrm{pr}(G \leq v + b_X - b_Z)\} f_{G_X}(v) dv \\ &+ \int_{-\infty}^{\infty} \mathrm{pr}(G \leq v + b_X - b_Z) \{f_{G_X}(v) - f_G(v)\} dv \\ &+ \int_{-\infty}^{\infty} \mathrm{pr}(G \leq v + b_X - b_Z) f_G(v) dv \triangleq \mathrm{I}_1 + \mathrm{I}_2 + \mathrm{I}_3. \end{aligned} \tag{5}$$

This illustrates that if the extreme value limit laws are used to approximate the probability $p = \mathrm{pr}(Z_{\max} \leq X_{\max})$, then the error incurred is given by the sum of the integrals $\mathrm{I}_1 + \mathrm{I}_2$. The relevant form of convergence for this type of problem is therefore an appropriately weighted $\mathbb{L}_1(\mathrm{Leb})$ norm of the density and distribution functions. This is stronger than uniform convergence of distribution functions, which has been studied for several starting distributions F_X. Notably, Hall (1979) showed that for the maxima of standard normally distributed random variables, the uniform convergence rate in (4) to the Type I limit is no better than $(\log \log n)^2 / \log n$.

The conclusion is that, while these limiting approximations are appealing in that they deliver simple easily interpretable solutions, their adequacy in the present context, particularly for small N_X or N_Z is not guaranteed and depends heavily on the the distributions of X and Z. The proposal discussed in §2 provides a compromise between simplicity and adequacy of the resulting analytic approximation.

4 An idealized case

For an example in which the density function of $V_i = 1 - F_Z(X_i)$ is exactly of the form $(1-\gamma)v^{-\gamma}$ used above, let X_i be exponentially distributed of rate ξ and Z_i be exponentially distributed of rate λ. Then $F_Z^{-1}(z) = -\log(1-z)/\lambda$ so that

$$\text{pr}(V_i \leq v) = 1 - F_X\{F_Z^{-1}(1-v)\} = v^{\xi/\lambda}, \quad 0 < v < 1.$$

It follows that the density function of each V_i is given by $(1-\gamma)v^{-\gamma}$, where $\gamma = 1 - \xi/\lambda$. Thus the probability p is approximated by formulae 1 or 2, or a truncated version thereof, with $\gamma = 1 - \xi/\lambda$.

A direct calculation for the exponentials would entail solving the integral

$$p = N_X \xi \int_0^\infty \{1 - \exp(-\lambda m)\}^{N_Z} \{1 - \exp(-\xi m)\}^{N_X - 1} dm,$$

which does not appear to have an exact analytic solution. An alternative is to approximate p using the Fisher and Tippett Type I limiting form of the rescaled exponential maxima. This is term I_3 in equation (5) and thus incurs the error $\mathrm{I}_1 + \mathrm{I}_2$. The scaling constants are

$$b_Z = b_Z(N_Z) = F_Z^{-1}(1 - N_Z^{-1}) = \frac{-\log\{1 - (1 - N_Z^{-1})\}}{\lambda} = \frac{\log N_Z}{\lambda}$$

and similarly for b_X. Thus, on letting $B = \xi^{-1}\log(N_X) - \lambda^{-1}\log(N_Z)$,

$$\mathrm{I}_3 = \int_{-\infty}^{\infty} \exp\{-e^{-(v+B)}\}\exp\{-(v + e^{-v})\}dv = \frac{e^B}{1 + e^B} = \frac{N_X^{1/\xi}}{N_Z^{1/\lambda} + N_X^{1/\xi}}. \tag{6}$$

The simulations in §5.1 show that this approximation is considerably less accurate than formula (1) even for very large values of N_X and N_Z, suggesting that at least one of the two error terms I_1 and I_2 from equation (5) decays slowly for exponentially distributed random variables. The simpler formula (3) fails because the representation $(1-\gamma)w^{-\gamma}$ does not hold even as an approximation for the density function of $W_i \triangleq F_X(Z_i)$ in this example.

5 Numerical assessment

5.1 Empirical analysis of an idealized case

In each of 10000 Monte Carlo replications we generated $N_Z = 100$ random variables $Z_1, \ldots, Z_{N_Z}$ from an exponential distribution of rate $\lambda \in \{1.1, 1.2, \ldots, 2\}$ and $N_X = 20$ random variables $X_1, \ldots, X_{N_X}$ from an exponential distribution of rate $\xi = 1$. The maxima of these two sets of random variables, $Z_{\max}$ and $X_{\max}$ were recorded. The simulated probability that $X_{\max}$ exceeds $Z_{\max}$ was obtained by averaging the indicator random variables $\mathbb{1}(X_{\max} > Z_{\max})$ over the Monte Carlo replications. The maximum likelihood estimate $\hat{\gamma}$ was also obtained in each simulation by fitting a density of the form $(1-\gamma)v^{-\gamma}$ to $V_i = 1-F_Z(X_i)$ for $i = 1, \ldots, N_X$. These maximum likelihood estimates were then averaged over Monte Carlo replications and used in the formula (1). These are plotted against λ in the left panel of Figure 1 along with the version that uses the exact value $\gamma = 1 - \xi/\lambda$. The latter approximation is essentially exact.

The experiment was repeated for $N_X = 50$ and the results are shown in the right panel of Figure 1. As expected, the quality of the analytic approximation is unaffected but the quality of the approximation based on the maximum likelihood estimate of γ is improved because the bias in the maximum likelihood estimates decreases with increasing N_X.

We also report the approximation (6) based on on Fisher's and Tippett's (1928) limit laws. These are evidently inaccurate for the sample sizes under consideration, even though the true values of ξ and λ are used. Further simulations (not reported) indicate that the approximation (6) becomes more accurate as N_X and N_Z increase, but remains rather poor even when $N_X = N_Z = 10^5$.

Figure 2 illustrates the sensitivity of approximation (1) to N_Z for two different values of N_X. Thus, although the formula uses the notional limiting operation $N_Z \to \infty$ in Stirling's formula, the approximation (1) is accurate even for small N_Z. This is expected in view of the remarkable accuracy of Stirling's approximation to $\Gamma(x)$ even for small x.

5.2 Violation of the main assumption

If the model $(1-\gamma)v^{-\gamma}$ for the density function of the random variables $V_i = 1 - F_Z(X_i)$ is not satisfied to an adequate order of approximation, formula (1) is likely to give inaccurate

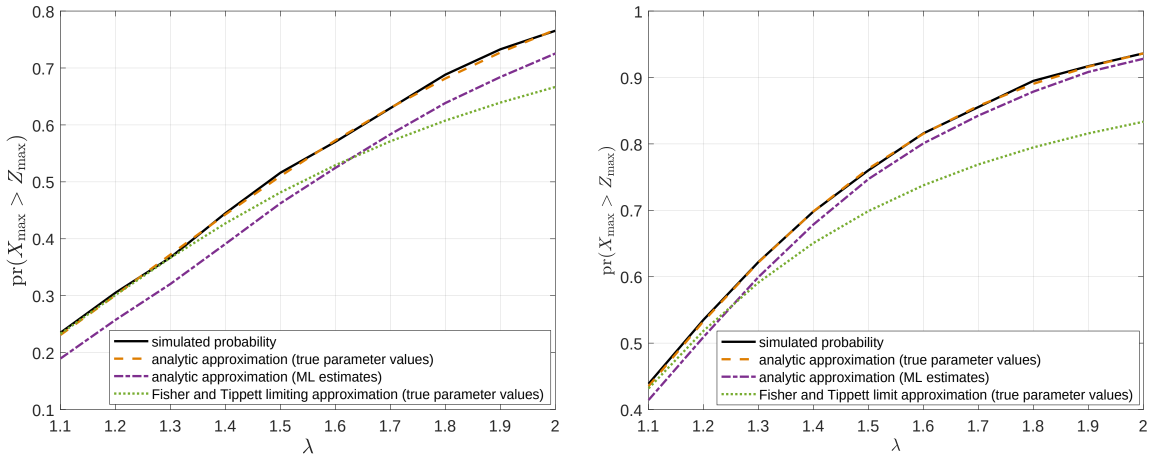

Figure 1: Simulated exceedance probabilities and the near-exact formula based on equation (1) for $N_Z = 100$, $\xi = 1$ and different values of λ, and for $N_X = 20$ (left) and $N_X = 50$ (right). Also depicted is formula (1) with ξ and λ replaced by the average of their maximum likelihood estimates over the 10^5 Monte Carlo replications and the limiting approximation (6) using the true values of ξ and λ.

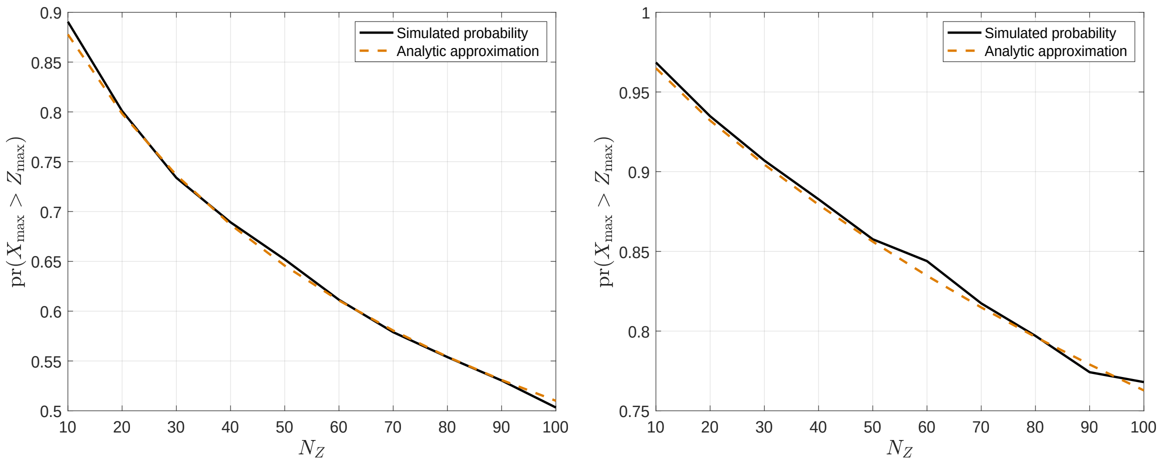

Figure 2: Simulated exceedance probabilities and the exact formula based on equation (1) for $\xi = 1$, $\lambda = 1.5$ and different values of N_Z, and for $N_X = 20$ (left) and $N_X = 50$ (right).

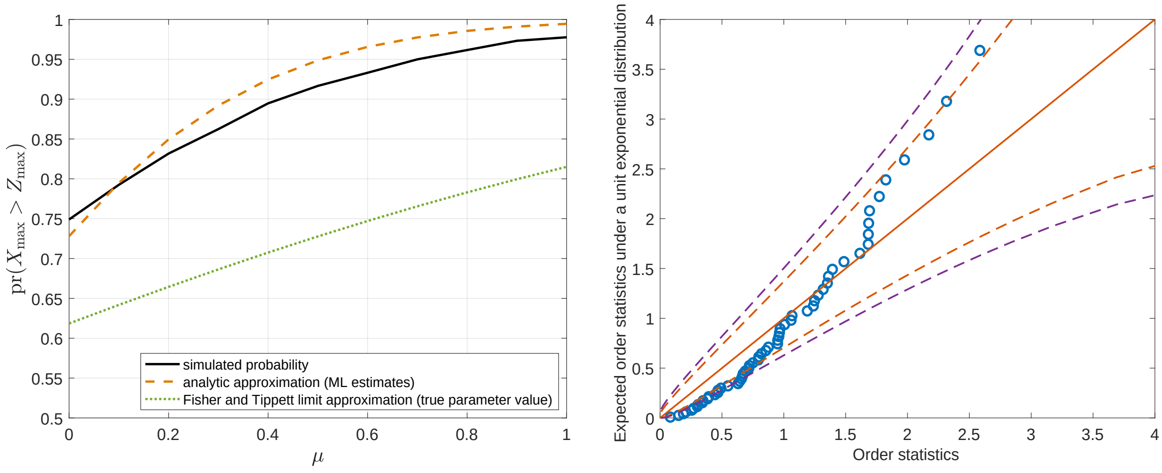

Figure 3: Left: simulated exceedance probabilities and the formula based on equation (1) for an example in which the key assumption is violated, for $N_Z = 20$ and $N_X = 60$. Also displayed is the approximation based on Fisher's and Tippett's limit laws. Right: probability plot of $-\log V_i^{1-\hat{\gamma}}$ against the unit exponential order statistics.

conclusions. The quality of the approximation can be assessed statistically for large N_X at a suitable confidence level α by fitting the density $(1-\gamma)v^{-\gamma}$ by maximum likelihood and comparing the realization of $2(1-\hat{\gamma})\sum_{i=1}^{n}\log V_i$ to the α upper quantile of a χ^2 distribution with $2N_X$ degrees of freedom. For an informal indication the order statistics of $-(1-\hat{\gamma})\log V_i$ can be plotted against the unit exponential order statistics. This is illustrated in Figure 3.

The experiment is as described in the previous section except that $Z_1, \ldots, Z_{N_Z}$ are standard normally distributed and $X_1, \ldots, X_{N_X}$ are normally distributed of unit variance and means μ as displayed on the axis of Figure 3 (left). The values of N_Z and N_X are 20 and 60. The parameter γ of the representation $(1-\gamma)v^{-\gamma}$ is estimated by maximum likelihood and used in the formula (1). Also depicted is the limiting approximation I_3 from equation (5). As in equation (6) this approximation is $e^B(1+e^B)^{-1}$ but with $B = \mu + \Phi^{-1}(1-N_X^{-1}) - \Phi^{-1}(1-N_Z^{-1})$, where Φ is the standard normal distribution function. The true value of μ is used in this latter approximation, yet the approximation I_3 is poor. This is unsurprising because, even for the weaker uniform convergence of distribution functions, convergence rates are extremely slow for normal distributions, as discussed in §3. The approximation (1) is less inaccurate and qualitatively successful, in spite of considerable violation of the assumption used in the calculation, as indicated by the right hand panel of Figure 3. This is a plot of the order statistics of $-(1-\hat{\gamma})\log V_i$ against the unit exponential order statistics.

6 Conclusion

We have illustrated how the transformation to $U_i = 1 - F_Z(Z_i)$ and $V_i = 1 - F_Z(X_i)$ facilitates probability calculations involving order statistics for the two populations. The accuracy of the ensuing approximations hinges of the plausibility of the $(1-\gamma)v^{-\gamma}$ assumption for the probability density function of each V_i. This can be assessed as in §5.2.

We have not discussed statistical aspects associated with the various complicating scenarios that could be envisaged, for instance if F_Z needs to be estimated. If there was an auxiliary sample in which the class labels were known, the simplest nonparametric estimator is $\hat{F}_Z(z) = n^{-1} \sum_{i=1}^{N_Z} 1\!\!1\{X_i \leq z\}$, leading to

$$\hat{U}_i \triangleq 1 - \hat{F}_Z(Z_i) = U_i + F_Z(Z_i) - \hat{F}_Z(Z_i).$$

The final term is bounded in absolute value by $\sup_{z \in \mathbb{R}} |\hat{F}_Z(z) - F_Z(z)|$. We have

$$\mathrm{pr}(\sup_{z \in \mathbb{R}} |\hat{F}_Z(z) - F_Z(z)| < \varepsilon) \geq 1 - 2e^{-2N_Z \varepsilon^2}$$

(Dvoretzky, Kiefer and Wolfowitz, 1956; Massart, 1990) implying (see §3.8 of Barndorff-Nielsen and Cox, 1989) that $\hat{U}_i = U_i + O_p(N_Z^{-1/2})$. The same argument applies for the $V_1, \ldots, V_{N_X}$ components.

Acknowledgement: the work was supported by the EPSRC (EP/P002757/1). I am grateful to the referees for helpful comments and Nicholas Beale of Sciteb Ltd. for asking a question that led to some of these calculations.

REFERENCES

Barndorff-Nielsen, O. E. and Cox, D. R. (1989). *Asymptotic Techniques for Use in Statistics.* Chapman and Hall, London.

Bairamov, I. and Parsi, S. (2011). Order statistics from mixed exchangeable random variables. *J. Comput. Appl. Math.*, 235, 4629–4638.

Bayramoglu (Bairamov), I. and Eryilmaz, S. (2015). Order statistics of dependent sequences consisting of two sets of exchangeable variables. *J. Comput. Appl. Math.*, 286, 1–6.

Dvoretzky, A. Kiefer, J. and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.*, 27, 642–669.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest of a sample. *Proc. Camb. Phil. Soc*, 24, 180–90.

Gnedenko, B. V. (1943). Sur la distribution limite du terme maximum d'une série aléatoire, *Ann. Math.*, 44, 423–53.

Gumbel, E. J. and von Schelling, H. (1950). The distribution of the number of exceedances. *Ann. Math. Statist.*, 21, 247–262.

Hall, P. G. (1979). On the rate of convergence of normal extremes. *J. App. Probab.*, 16, 433–39.

Massart, P. (1990). The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *Ann. Probab.*, 18, 1269–1283.

Sarkadi, K. (1957). On the distribution of the number of exceedances. *Ann. Math. Statist.*, 28, 1021–1023.