Journal of Data Science 22 (4), 524–536
October 2024

DOI: 10.6339/24-JDS1120
Statistical Data Science

Discrete Extremes

Adrien S. Hitz[1], Richard A. Davis[2], and Gennady Samorodnitsky[3]

[1]*Department of Statistics, University of Oxford, 24-29 St Giles, Oxford OX1 3LB, UK*
[2]*Department of Statistics, Columbia University, 1255 Amsterdam Avenue, New York, NY, US*
[3]*School of Operations Research and Information Engineering, Cornell University, 220 Rhodes Hall, Ithaca, NY, US*

Abstract

Our contribution is to widen the scope of extreme value analysis applied to discrete-valued data. Extreme values of a random variable are commonly modeled using the generalized Pareto distribution, a peak-over-threshold method that often gives good results in practice. When data is discrete, we propose two other methods using a discrete generalized Pareto and a generalized Zipf distribution respectively. Both are theoretically motivated and we show that they perform well in estimating rare events in several simulated and real data cases such as word frequency, tornado outbreaks and multiple births.

Keywords *count data; discrete distribution; extreme value theory; generalized Pareto distribution; peaks over threshold; tail approximation; Zipf distribution*

1 Introduction

Extreme quantile estimation is an important but difficult problem in statistics, especially when the quantile is beyond the range of the data. In the univariate case, an approach that often works well in practice is to model observations above a large threshold with a parametric family of distributions, the generalized Pareto distribution. We will illustrate however that this approach can be problematic when data takes discrete values. In the context of discrete data, quantifying uncertainty may be crucial, as exemplified by hospital bed occupancy, discussed in Ranjbar et al. (2022).

Let X be a random variable (continuous or discrete) taking values in $[0, x_F)$ for $x_F \in (0, \infty]$, and suppose that there exists a strictly positive sequence a_u such that

$$a_u^{-1}(X - u) \mid X \geqslant u \ \rightarrow \ Z, \tag{1}$$

in distribution as $u \rightarrow x_F$, for some Z following a non-degenerate probability distribution on $[0, \infty)$. Then, Z follows a generalized Pareto distribution, defined by its survival function

$$\bar{F}_{\text{GPD}}(x; \sigma, \xi) = \left(1 + \xi \frac{x}{\sigma}\right)_+^{-1/\xi}, \quad x \geqslant 0,$$

with $\sigma > 0$ and $(1 + \xi x)^{1/\xi} = e^x$ if $\xi = 0$ (Pickands, 1975). For $\xi < 0$, $\bar{F}_{\text{GPD}}$ has support on $[0, \sigma/|\xi|]$. Condition (1), written as $X \in \text{MDA}_\xi$, means that X is in the maximum domain of attraction of an extreme value distribution with shape parameter ξ (see Resnick, 1987). In this case, the sequence of cumulative distribution functions of $a_u^{-1}(X - u) \mid X \geqslant u$ converges

Received January 28, 2024; Accepted January 31, 2024

uniformly to $1 - \bar{F}_{\mathrm{GPD}}$ on $[0, \infty)$. Thus, the distribution of exceedances above a large threshold u (also called "peaks-over-threshold") can be approximated in the following manner:

$$\mathrm{pr}(X - u > x \mid X \geqslant u) = \mathrm{pr}\left\{a_u^{-1}(X - u) > a_u^{-1}x \mid X \geqslant u\right\} \approx \bar{F}_{\mathrm{GPD}}(x; \sigma a_u, \xi), \tag{2}$$

(Davison and Smith, 1990). This approximation, called the generalized Pareto approximation, is convenient in practice because it does not rely on a specific distributional assumption; X is only required to belong to some maximum domain of attraction, which holds for most common continuous distributions.

If the observations are discrete, however, one may want to preserve and utilize the discreteness in the extreme estimation. It is not, however, clear how discrete exceedances over high threshold should be modeled. The generalized Pareto approximation is often applied ignoring the discrete nature of the data. This poses two issues: first, a necessary condition for a discrete random variable X to be in some maximum domain of attraction in the case $x_F = \infty$ is that X is long-tailed, i.e., $\bar{F}_X(u+1)/\bar{F}_X(u) \to 1$ as $u \to \infty$ (Shimura, 2012), and many common discrete distributions, including geometric, Poisson and negative binomial distributions, are not long-tailed. (All long-tailed distributions are heavy-tailed, but the converse is false.) Specific convergence results for maxima of discrete observations have thus been derived (Anderson, 1970, 1980; Dkengne et al., 2016), but the limit is always a continuous distribution, which leads to the second issue: treating discrete data as continuous introduces a bias in the likelihood function. Since the shape and location parameters ξ and σ of the generalized Pareto approximation are unknown in practice, they must be estimated from the exceedance data. We will see that the bias may render the approximation inadequate — even when X is long-tailed, that is, when (2) is valid in theory.

Our contribution is to overcome these limitations by proposing two peaks-over-threshold methods, each relying on a parametric family of discrete distributions: the discrete generalized Pareto and the generalized Zipf distribution. The latter distributions exist in the literature but have not been justified for modeling extremes. As we will show, these new approximations can be theoretically motivated for X belonging to a broad class of discrete distributions, and they match or outperform the generalized Pareto approximation. They deliver similar results but it is still unclear if one of them should be preferred.

From now on, we assume that X is a discrete random variable with non-negative values, and $\xi \geqslant 0$. The first method adapts the condition $X \in \mathrm{MDA}_\xi$ to the discrete case as follows. Suppose that there exists a random variable $Y \in \mathrm{MDA}_\xi$ with survival function $\bar{F}_Y$ on $[0, \infty)$ such that $\mathrm{pr}(X \geqslant k) = \mathrm{pr}(Y \geqslant k)$ for $k = 0, 1, 2, \ldots$, that is, the equality in distribution, $X = \lfloor Y \rfloor$, holds. In this case, we say that X is in the discrete maximum domain of attraction, which we write as $X \in \text{D-MDA}_\xi$. We call Y an extension of X and such an extension is not unique. Shimura (2012) proved that $X \in \mathrm{MDA}_\xi$ if and only if $X \in \text{D-MDA}_\xi$ and X is long-tailed. (When X is long-tailed, an extension of X, which takes values in $\mathbb{N}$, can be X itself, if seen as taking values in $\mathbb{R}$.) It was also shown by Shimura (2012) that geometric, Poisson and negative binomial distributions belong to the discrete maximum domain of attraction. Therefore, $\mathrm{MDA}_\xi \subsetneq \text{D-MDA}_\xi$ for discrete distributions. If $X \in \text{D-MDA}_\xi$ and $Y \in \mathrm{MDA}_\xi$ is a corresponding extension satisfying $X = \lfloor Y \rfloor$ in distribution, then, for large integers u, we use (2) to obtain

$$\begin{aligned}\mathrm{pr}(X - u = k \mid X \geqslant u) &= \mathrm{pr}(Y - u \geqslant k \mid Y \geqslant u) - \mathrm{pr}(Y - u \geqslant k + 1 \mid Y \geqslant u) \\ &\approx p_{\text{D-GPD}}(k; \sigma a_u, \xi),\end{aligned} \tag{3}$$

where $p_{\text{D-GPD}}$ is the probability mass function of the discrete generalized Pareto distribution defined by

$$p_{\text{D-GPD}}(k;\sigma,\xi) = \bar{F}_{\text{GPD}}(k;\sigma,\xi) - \bar{F}_{\text{GPD}}(k+1;\sigma,\xi),$$

for $k = 0, 1, 2, \ldots$. Equation (3) provides a method for modeling discrete exceedances over threshold that we call the discrete generalized Pareto approximation. The latter distribution has been applied in the context of discrete extremes to model road accidents (Prieto et al., 2014), network data (Charpentier and Flachaire, 2019), avalanche occurrences (Evin et al., 2021), hospital cases for flu (Ranjbar et al., 2022) and wildfire counts (Koh, 2023). Ahmad et al. (2022) proposed smooth extensions to avoid threshold selection, and various aspects of discrete Pareto-type distributions were studied in Krishna and Pundir (2009), Buddana and Kozubowski (2014) and Kozubowski et al. (2015),

Whereas the first method is based on an extension of $\bar{F}_X$ by a survival function in the maximum domain of attraction, the second method assumes instead an extension of p_X, the probability mass function of X. Suppose that there exists a non-negative random variable $Y \in \text{MDA}_{\xi/(1+\xi)}$ with survival function $\bar{F}_Y$ on $[0,\infty)$ such that $p_X(k) = c\,\bar{F}_Y(k)$ for $k = d, d+1, d+2, \ldots$, for some $c > 0$ and $d \in \mathbb{N}_0 = \{0, 1, \ldots\}$. In this case, we say that p_X is in the discrete maximum domain of attraction which is denoted by $p_X \in \text{D-MDA}_{\xi/(1+\xi)}$, and call $\bar{F}_Y$ an extension of p_X. We will show that $p_X \in \text{D-MDA}_{\xi/(1+\xi)}$ implies $X \in \text{MDA}_\xi$ (under a mild condition in the case $\xi = 0$), and that geometric, Poisson and negative binomial satisfy $p_X \in \text{D-MDA}_0$. It follows from (2) that, for large integers u,

$$\text{pr}(X - u = k \mid X \geqslant u) = \frac{\text{pr}(Y > u+k)/\text{pr}(Y > u)}{\sum_{i=0}^{\infty} \text{pr}(Y > u+i)/\text{pr}(Y > u)} \approx p_{\text{GZD}}\big\{k; (1+\xi)\sigma a_u, \xi\big\}, \qquad (4)$$

where

$$p_{\text{GZD}}(k;\sigma,\xi) = \frac{(1+\xi\frac{k}{\sigma})^{-1/\xi-1}}{\sum_{i=0}^{\infty}(1+\xi\frac{i}{\sigma})^{-1/\xi-1}}, \quad k = 0, 1, 2, \ldots, \qquad (5)$$

is the probability function of a distribution that we call the generalized Zipf distribution. In the case $\xi = 0$, the latter is a geometric distribution (and so is the discrete generalized Pareto distribution), and in the case $\xi > 0$, it is a Zipf–Mandelbrot distribution (Mandelbrot, 1953). Zipf-type families have been fitted to various discrete datasets such as word frequencies (Booth, 1967), city sizes (Gabaix, 1999), company sizes (Axtell, 2001), website visits (Clauset et al., 2009) and insurgency casualties (Patel et al., 2021). The Zipf law, arising in the case $\xi = \sigma$, is sometimes presented as the discrete counterpart of the Pareto distribution (Arnold, 1983). We refer to the approximation procedure in (4) as the generalized Zipf approximation.

We also note a few recent work in discrete extremes. Valiquette et al. (2023) studied the tail behavior of Poisson mixtures distributions and derived maximum domain of attractions results for this class. Koutsoyiannis (2023) proposed using K-moments to estimate distributions when one is interested in their tail distributions, which is applicable in the discrete case as well. Ghosh et al. (2023) analyzed a discretization of the Gamma-Lomax distribution, providing characterization and estimation results in the context of discrete extremes.

2 Theoretical Results

We start by showing that the probability density and mass functions of the generalized Pareto, discrete generalized Pareto and Zipf distributions are asymptotically equivalent as σ tends to infinity. Proofs are given in the Appendix.

Proposition 1. *For $\sigma > 0$, $\xi \geqslant 0$ and $q, \tilde{q} \in \{f_{GPD}, p_{D\text{-}GPD}, p_{GZD}\}$, it holds*

$$\lim_{\sigma\to\infty} \sup_{k=0,1,2,\ldots} \left| \frac{q(k;\sigma,\xi)}{\tilde{q}(k;\sigma,\xi)} - 1 \right| = 0.$$

This suggests that modeling a sample from $X - u \mid X \geqslant u$ by maximum likelihood using either f_{GPD}, $p_{\text{D-GPD}}$ or p_{GZD} should not differ too much if the estimated scale parameter $\hat{\sigma}$ is sufficiently large. When the sample size and u grow, $\hat{\sigma}$ only goes to infinity if the sequence a_u defined in (1) satisfies $a_u \to \infty$, which occurs if and only if X is long-tailed. Even in this case, a_u might grow too slowly for the three methods to be similar in practice, as we will see in Section 3.

The results below formally justify the approximation procedures we have introduced. We start with a convergence result for the discrete generalized Pareto approximation.

Proposition 2. *If $X \in D\text{-}MDA_\xi$ for $\xi \geqslant 0$, then there exists a positive sequence $(a_u,\ u = 1, 2, \ldots)$ such that*

$$\lim_{u\in\mathbb{N}_0,\, u\to\infty} \sup_{k=0,1,2,\ldots} \left| pr(X = u + k \mid X \geqslant u) - p_{D\text{-}GPD}(k; a_u, \xi) \right| = 0. \tag{6}$$

We remark that (6) is not informative if $a_u \to \infty$ because the two terms converge to 0. Next, we consider the case $p_X \in$ D-MDA. Recall that a distribution F is in MDA_0 if and only if the survival function has a representation

$$\bar{F}(x) = c(x) \exp\left\{ -\int_0^x \frac{1}{a(y)} dy \right\}, \quad x \in \mathbb{R}, \tag{7}$$

where $a(\cdot)$, called the auxiliary function, is positive and differentiable with $a'(x) \to 0$ as $x \to \infty$; and $c(\cdot)$ is a positive function with limit $c > 0$ (Embrechts et al., 2013). If $c(x) = c$ on (d, ∞) for some $d \in \mathbb{R}$, then we say that the distribution F satisfies the von Mises condition.

Theorem 1. *If $p_X \in D\text{-}MDA_{\xi/(1+\xi)}$ and $\xi > 0$, then $X \in MDA_\xi$ and, for any sequence of nonnegative integers $(k_u)_{u\in\mathbb{N}_0}$ such that $\sup_u k_u/u < \infty$,*

$$\lim_{u\in\mathbb{N},\, u\to\infty} \frac{pr(X = k_u + u \mid X \geqslant u)}{q(k_u; \xi u, \xi)} = 1, \tag{8}$$

where $q \equiv f_{GPD}$, $p_{D\text{-}GPD}$ and p_{GZD}.

If $p_X \in D\text{-}MDA_0$ and if the auxiliary function of an extension $\bar{F}$ of p_X satisfies $\lim_{x\to\infty} a(x) = \sigma > 0$, then $X \in D\text{-}MDA_0$ and

$$\lim_{u\in\mathbb{N},\, u\to\infty} pr(X = k + u \mid X \geqslant u) = p_{D\text{-}GPD}(k; \sigma, 0) = p_{GZD}(k; \sigma, 0), \quad k = 0, 1, 2, \ldots. \tag{9}$$

The condition $p_X \in$ D-MDA is satisfied, among others, by the Zipf–Mandelbrot, geometric, Poisson and negative binomial distributions as shown below and in the Appendix.

Example 1. *The probability mass function of a Zipf–Mandelbrot distribution is proportional to $(k + q)^{-1-1/\xi}$ for $k = 0, 1, 2, \ldots$, $q > 0$, $\xi > 0$, and satisfies $p_X \in D\text{-}MDA_{\xi/(1+\xi)}$ because it can be extended by $\bar{F}_Y(y) = c(y + q)^{-1-1/\xi}$ for $y \geqslant 0$ and some $c > 0$ with $Y \in MDA_{\xi/(1+\xi)}$. The probability mass function of a geometric distribution belongs to $D\text{-}MDA_0$ as it coincides up to a constant with the survival function of an exponential distribution. The latter distribution clearly satisfies the von Mises condition and thus is a member of MDA_0. The auxiliary function is, in fact, equal (eventually) to $1/\lambda$, where λ is the rate of the exponential distribution.*

To summarize, for a discrete random variable X and $\xi \geqslant 0$, it holds $X \in \mathrm{MDA}_\xi$ if and only if $X \in \text{D-MDA}_\xi$ and X is long-tailed. If $\xi > 0$, then $p_X \in \text{D-MDA}_{\xi/(1+\xi)}$ implies $X \in \text{D-MDA}_\xi$; the same implication holds in the case $\xi = 0$ if the auxiliary function of the extension of p_X satisfies $a(x) \to \sigma \in (0, \infty)$ as $x \to \infty$.

3 Simulation Study

We assess the performance of the discrete generalized Pareto and the generalized Zipf approximations for estimating the probability of a rare event from discrete data, and illustrate why they should be preferred to the generalized Pareto approximation, whether X is long-tailed or not. Let $\alpha = 2$, $\beta = 0.75$ and

$$X = \lfloor Y \rfloor, \quad Y \sim \text{Inverse-gamma}(\alpha, \beta), \tag{10}$$

where the inverse-gamma distribution has density function $f(x) = \Gamma(\alpha)^{-1}\beta^\alpha x^{-\alpha-1}\exp(-\beta/x)$, $x > 0$. The experiment described below is repeated 500 times. An independent and identically distributed sample of size 8000 is drawn from the distribution of X. The goal is to estimate the probability of the extreme region

$$p_e = \mathrm{pr}\big(X \geqslant \lfloor q_e \rfloor\big), \quad \lfloor q_e \rfloor = 52, \tag{11}$$

where q_e is the 99.99 percentile of Y, i.e., the value exceeded once every 10 000 times on average. The strategy pursued is to select an integer threshold u as the 95th empirical percentile of the sample, fit parametric distributions to the exceedances $X - u \mid X \geqslant u$, and use them to extrapolate the tail and estimate p_e. (Selecting an appropriate threshold is crucial when estimating high quantiles and can be based on techniques such as mean residual plots, see e.g. Davison and Smith, 1990.) It clearly holds $p_X \in \text{D-MDA}_{\xi/(1+\xi)}$ for $\xi = 1/\alpha = 1/2$, thus the three approximations are justified. The generalized Pareto distribution is fitted to the observations shifted by continuity correction $\delta = 0$ or $\delta = \frac{1}{2}$. As a benchmark, we will also estimate p_e from a sample of the continuous variable Y (as opposed to its discretization X) using the generalized Pareto approximation.

A frequency plot of the exceedances of a sample of X above u is displayed on the left-hand side in Figure 1. For each model, we compute the maximum likelihood estimators $\hat{\sigma}$ and $\hat{\xi}$ by performing a two dimensional maximization using the function `optim` of R (R Core Team, 2024) with starting values $(1, 1)$. We then compute $\hat{p}_e$ and approximate 90% confidence intervals from the Fisher information matrix under asymptotic normality of the estimators. Table 1 displays: the average estimates $\hat{p}_e$, $\hat{\xi}$ and $\hat{\sigma}$ over the 500 replications of the experiment, the coverage of the confidence intervals, their average length and their true length. (Coverage indicates the proportion of time the truth lies in the confidence interval, true length is here defined as $\ell^\star = q_{0.05}(\hat{\mathbf{p}}_\mathbf{e}) - q_{0.95}(\hat{\mathbf{p}}_\mathbf{e})$, where $\hat{\mathbf{p}}_\mathbf{e}$ is the vector of maximum likelihood estimates in the 500 replicated experiments, and $q_{(\cdot)}$ is the quantile function.) The discrete generalized Pareto and Zipf approximations provide relatively accurate estimates of p_e from the discretized data with a coverage close to the correct one of 90%, and their performance is good relative to the situation of full information where the continuous data are available — notice how the estimates are very similar to one another. On the other hand, the two versions of the generalized Pareto approximation perform poorly, the worst being the case $\delta = 0$.

The ability of the discrete generalized Pareto and Zipf approximations to accurately estimate the probability of rare events is supported by complementary simulated cases covering $\xi = 0$ and $\xi < 0$ (Hitz, 2016, Chapter 2).

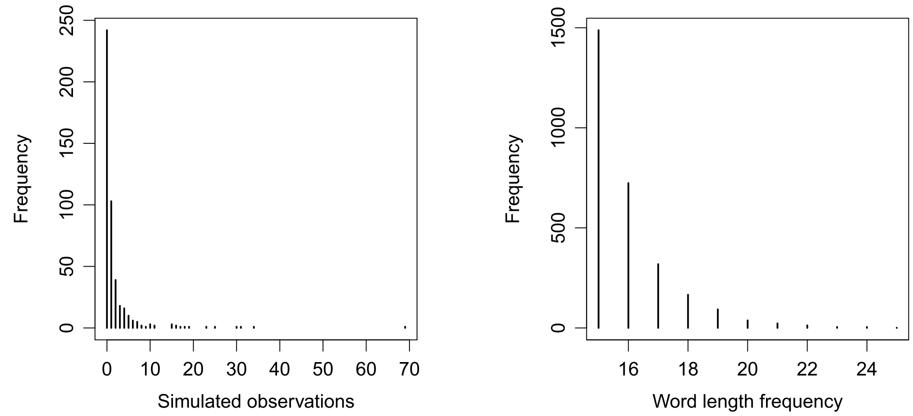

Figure 1: On the left: frequency plot of 462 exceedances of X over the 95th empirical quantile $u = 2$ simulated from (10). On the right: frequency plot of the length of the 2875 longest French words.

Table 1: Performance of several methods in estimating the probability p_e of the rare event defined in (11) from about 460 exceedances in each experiment. The table displays average maximum likelihood estimators for p_e, ξ and σ across the 500 replicated experiments. Coverage c, average length l and true length $l^\star$ of 90% confidence intervals are shown between brackets. The discrete generalized Pareto and Zipf approximations are superior in this case.

	$\hat{p}_e \cdot 10^4$ $(c, l, l^\star)$	$\hat{\xi}$ (c, l)	$\hat{\sigma}$ (l)
Truth	1.03	0.50	
Fitted to $Y - u \mid Y \geqslant u$			
Generalized Pareto distribution	1.07 (85%, 1.84, 1.81)	0.49 (94%, 0.28)	1.14 (0.36)
Fitted to $X - u \mid X \geqslant u$, $X = \lfloor Y \rfloor$			
Discrete generalized Pareto distribution	**1.09** (87%, 1.92, 1.85)	0.49 (93%, 0.29)	1.13 (0.40)
Generalized Zipf distribution	**1.11** (88%, 1.97, 1.88)	0.50 (94%, 0.30)	1.34 (0.35)
Generalized Pareto distribution, $\delta = \frac{1}{2}$	0.44 (31%, 0.86, 0.97)	0.36 (35%, 0.22)	1.38 (0.39)
Generalized Pareto distribution, $\delta = 0$	50.42 (85%, 162.32, 71.45)	8.29 (0%, 1.58)	0.00 (0.00)

4 Real Data Examples

We now illustrate the methods discussed in this article on three real datasets. The first consists of the frequency X of word length in the French lexicon (New et al., 2004); for instance, "anticonstitutionnellement" is the only word of 25 letters in French. We focus on describing the tail distribution and fit the usual models to $X - u \mid X \geqslant u$ with $u = 15$, the 98th percentile of the data. A frequency plot of the 2875 exceedances is shown on the right-hand side of Figure 1. The discrete generalized Pareto and Zipf distributions deliver a good fit and similar estimations to one another, and clearly outperform the generalized Pareto approximation as shown in Table 2 by p-values of discrete Kolmogorov–Smirnov tests based on the difference between the fitted distribution and the empirical distribution of bootstrapped data. Notice that the negative binomial also fits well in this case. The procedure for computing p-values in Table 2 for the word length data was the following: resample the data with replacement; compute the difference between the fitted and empirical distribution of this sample; get a p-value by Monte Carlo

Table 2: Fit of several distributions to the length of the 2875 longest French words, and to the number of extreme tornadoes per outbreak for the 435 outbreaks with 12 or more such tornadoes in the United States between 1965 and 2015. The table displays p-value of discrete Kolmogorov–Smirnov tests (of the discretized model in the case of continuous models), negative log-likelihood $-\ell$ and maximum likelihood estimates with 90% confidence intervals and possible temporal trend $\hat{\sigma}_t$ in the scale parameter.

	p-val.	$-\ell$	ξ	$\hat{\sigma}_0$	$\hat{\sigma}_t$
Word length					
Discrete generalized Pareto dist.	**0.40**	3894.0	$0.02_{[-0.01,0.06]}$	$1.36_{[1.30,1.43]}$	
Generalized Zipf distribution	**0.40**	3894.0	$0.02_{[-0.01,0.06]}$	$1.37_{[1.32,1.43]}$	
Generalized Pareto dist., $\delta = \frac{1}{2}$	0.02	3951.2	$-0.04_{[-0.06,-0.01]}$	$1.51_{[1.45,1.57]}$	
Negative binomial	0.37	3893.9			
Tornado outbreak					
Discrete generalized Pareto dist.	0.19	1439.92	$0.27_{[0.16,0.37]}$	$4.81_{[3.64,5.99]}$	$6.11_{[3.74,8.48]}$
Generalized Pareto dist., $\delta = \frac{1}{2}$	0.18	1439.93	$0.26_{[0.16,0.37]}$	$4.86_{[3.68,6.04]}$	$6.13_{[3.75,8.50]}$

simulation using `R` package `dgof` (Arnold and Emerson, 2011); repeat 200 times and take the average.

The second dataset comes from Tippett et al. (2016) who report the number X of extreme tornadoes per outbreak in the United States between 1965 and 2015, where an outbreak is a sequence of tornadoes that are high on the Fujita scale and occur close to each other in time. The authors found that the 435 observations from $X - u \mid X \geqslant u$ for $u = 12$ were well modeled by a generalized Pareto distribution with linear temporal trend in the scale parameter and continuity correction $\delta = \frac{1}{2}$. The scale parameter is modeled as $\sigma(\mathbf{t}) = \sigma_0 + \sigma_1\mathbf{t}$, where $\mathbf{t}$ is the time covariate rescaled between $[0, 1]$. The p-values in Table 2 for the tornado data were computed as follows: split the dataset into 5 groups depending on which time covariates are the nearest to $t_i = 0.1, 0.3, 0.5, 0.7, 0.9$; for each group, assume $\sigma(\mathbf{t}) = \sigma_0 + \sigma_1 t_i$ and compute the p-value of a discrete Kolmogorov–Smirnov test as explained previously; report the smallest of these 5 p-values. Maximum likelihood estimates and discrete Kolmogorov–Smirnov tests in Table 2 show that there is virtually no difference between the three approximations (only two of them are presented). This is consistent with Proposition 1 since the location parameter $\hat{\sigma}$ is larger in this case. Treating the tornado data as continuous is acceptable because there are fewer tied observations: about 38% of the data consists of values shared with no more than 20 other observations, compared to 13% for the simulated data and 1% for the word length data. Loosely, the data look less discrete (a frequency plot is displayed in the Appendix), thus the generalized Pareto approximation is appropriate here.

The third dataset counts the number X of multiple births in the United States from 1995 to 2014 and is displayed on the left-hand side of Table 3 (Hamilton et al., 2015). The observations are censored from above and only take 5 distinct values, it is thus interesting to see if the discrete generalized Pareto and Zipf distributions can still describe the tail of the data in this non-standard estimation problem. We randomly select from the dataset a sample that contains a thousand times fewer observations, and estimate from these the probability p_e that an American women delivers quintuplets or more by fitting a right-censored version of the usual models to $X^C - u \mid X^C \geqslant u$ for $u = 2$, where $X^C = \min(X, 5)$. The experiment was repeated 500 times, and each sample contained on average 9 quatruplets and 1 quintuplet or more. In the case of the generalized Zipf distribution, maximum likelihood estimates could not be computed nu-

Table 3: On the left: frequency table of multiple births in the United States from 1995 to 2014. On the right: performance of several methods in estimating the probability p_e of an American women delivering quintuplets or more at birth using only one thousandth of the dataset. In each experiment, the threshold was $u = 1$ and there were about 2600 exceedances (see Table 1 for notation). The discrete generalized Pareto and Zipf provide useful techniques for such extrapolations.

	Multiple Births
Single	78 178 588
Twin	2 500 340
Triplet	1 17 603
Quadruplet	8 108
Quint. or more	1 353

	$\hat{p}_e \cdot 10^5$ $(c, l, l^\star)$
Truth	1.7
Discrete generalized Pareto distribution	**1.4** (74%, 2.9, 2.9)
Generalized Zipf distribution	1.6 (87%, 3.3, 2.8); n/a
Negative Binomial	1.2 (65%, 2.3, 2.3)
Generalized Pareto distribution, $\delta = \frac{1}{2}$	n/a

merically in 52 out of 500 replicated experiments and the hessian matrix could not be computed numerically in 70 experiments. In the case of the generalized Pareto distribution with $\delta = \frac{1}{2}$, the log-likelihood function could not be maximized numerically. Table 3 shows that the discrete generalized Pareto and Zipf distributions outperform common alternatives, and seem to be useful techniques for inference from such limited data. The applicability of peaks-over-threshold methods when u is a particularly small integer should be more rigorously studied.

5 Discussion

In summary, we have proposed two peaks-over-threshold methods for discrete random variables, motivated by Proposition 1 and Theorem 1, and have shown that they provide accurate tail probability estimates in simulated and real data. We conclude that there is no downside to fit a discrete generalized Pareto for discrete data as opposed to a generalized Pareto distribution.

Future work could explore the use of the generalized Zipf and discrete Pareto distributions in the case $\xi < 0$, and further investigate how they relate to each other as they seem to perform similarly. The latter distribution benefits from its closed-form survival and probability mass function, allowing for exact likelihood based inference.

Supplementary Material

The data and code supporting this article are available in the GitHub repository at https://github.com/adhi1000/discrete_extremes. This archive includes the file `Simulated Data.R`, which details the simulation study discussed in Section 3, and the files `Word Frequency.R`, `Tornado.R` and `Multiple Birth.R`, which replicate the real data analysis presented in Section 4.

A Appendix

The following auxiliary lemma is elementary (as the sum can be sandwiched between two integrals).

Lemma 1. *If* $\xi > 0$, *then*

$$u^{1/\xi} H_{1+1/\xi,u} \to \xi,$$

as $u \to \infty$*, where* $H_{s,q} = \sum_{i=0}^{\infty} (q+i)^{-s}$ *is the Hurwitz-Zeta function.*

Proof of Proposition 1. Suppose first that $\xi > 0$. Then

$$\frac{p_{\text{D-GPD}}(k;\sigma,\xi)}{f_{\text{GPD}}(k;\sigma,\xi)} = \frac{(1+\xi\frac{k}{\sigma})^{-1/\xi} - (1+\xi\frac{k+1}{\sigma})^{-1/\xi}}{\frac{1}{\sigma}(1+\xi\frac{k}{\sigma})^{-1/\xi-1}}$$
$$= \left\{1 - \left(1 + \frac{\xi}{\sigma+\xi k}\right)^{-1/\xi}\right\}(\sigma+\xi k) \to 1,$$

uniformly in $k = 0, 1, 2, \ldots$ as $\sigma \to \infty$. Furthermore,

$$\sup_{k=0,1,2,\ldots} \frac{f_{\text{GPD}}(k;\sigma,\xi)}{p_{\text{GZD}}(k;\sigma,\xi)} = \sigma^{-1}\sum_{i=0}^{\infty}(1+\xi i/\sigma)^{-1/\xi-1} \to 1,$$

as $\sigma \to \infty$ by Lemma 1. In the case $\xi = 0$,

$$p_{\text{D-GPD}}(k;\sigma,0)/f_{\text{GPD}}(k;\sigma,0) = p_{\text{GZD}}(k;\sigma,0)/f_{\text{GPD}}(k;\sigma,0) = \sigma\big(1-e^{-1/\sigma}\big) \to 1. \qquad \square$$

Proof of Proposition 2. By assumption, there exists a random variable $Y \in \text{MDA}_\xi$ for $\xi \geqslant 0$ and a positive function $(\tilde{a}_u,\ u > 0)$ such that $X = \lfloor Y \rfloor$ in distribution and the sequence of functions $\text{pr}\{\tilde{a}_u^{-1}(Y-u) \geqslant x \mid Y \geqslant u\}$, $x \geqslant 0$, converges uniformly, as $u \to \infty$, to the function $\bar{F}_{\text{GPD}}(x;\sigma,\xi)$, $x \geqslant 0$, for some $\sigma > 0$ and $\xi \geqslant 0$. For a positive integer u we let $a_u = \tilde{a}_u\sigma$. Then

$$\begin{aligned}
&\sup_{k=0,1,2,\ldots} \big|\, \text{pr}(X = u+k \mid X \geqslant u) - p_{\text{D-GPD}}(k;a_u,\xi)\,\big| \\
&= \sup_{k=0,1,2,\ldots} \big|\text{pr}\big\{\tilde{a}_u^{-1}(Y-u) \geqslant \tilde{a}_u^{-1}k \mid Y \geqslant u\big\} - \text{pr}\big\{\tilde{a}_u^{-1}(Y-u) \geqslant \tilde{a}_u^{-1}(k+1) \mid Y \geqslant u\big\} \\
&\quad - \bar{F}_{\text{GPD}}(k;a_u,\xi) + \bar{F}_{\text{GPD}}(k+1;a_u,\xi)\big| \\
&\leqslant 2\sup_{x\geqslant 0}\big|\text{pr}\big\{\tilde{a}_u^{-1}(Y-u) \geqslant x \mid Y \geqslant u\big\} - \bar{F}_{\text{GPD}}(x;\sigma,\xi)\big| \to 0
\end{aligned}$$

as $u \to \infty$ over the integers. $\square$

The proof of Theorem 1 relies on properties of regularly varying functions. Recall that a positive and measurable function f on $[1,\infty)$ is regularly varying if there exists $\alpha \in \mathbb{R}$ such that

$$\lim_{u\to\infty} \frac{f(ux)}{f(u)} \to x^\alpha, \quad x \geqslant 1,$$

and we write $f \in \text{RV}_\alpha$ (see e.g. Bingham et al. (1989)). If $f \in \text{RV}_{-\alpha}$ for $\alpha \geqslant 0$, then

$$\lim_{u\to\infty} \sup_{x\in[1,b]} \left|\frac{f(ux)}{f(u)} - x^{-\alpha}\right| \to 0, \tag{12}$$

for $b = \infty$ if $\alpha > 0$, and for any $b < \infty$ if $\alpha = 0$. If $f \in \text{RV}_{-\alpha}$ for $\alpha > 0$, then by Potter's bounds (see e.g. Resnick, 1987) for any $\epsilon > 0$ there is $u_\epsilon \in (0,\infty)$ such that

$$e^{-\epsilon}x^{-\alpha-\epsilon} \leqslant \frac{f(ux)}{f(u)} \leqslant e^{\epsilon}x^{-\alpha+\epsilon}, \quad x \geqslant 1, \tag{13}$$

for $u \geqslant u_\epsilon$. We say that X is regularly varying if $\bar{F}_X \in \text{RV}_{-\alpha}$ for some $\alpha > 0$, a necessary and sufficient condition for $X \in \text{MDA}_{1/\alpha}$.

Proof of Theorem 1. We start by proving the first part of the theorem. By assumption, there exists a survival function $\bar{F}$ such that $\bar{F}(k) = c\, p_X(k)$ for $c > 0$, k large enough and $\bar{F} \in \mathrm{RV}_{-1/\xi-1}$. The last condition is equivalent to $\bar{F}(\lfloor \cdot \rfloor) \in \mathrm{RV}_{-1/\xi-1}$ (Shimura, 2012). It follows from results on integrals of monotone regularly functions in Bingham et al. (1989) that $\bar{F}_X \in \mathrm{RV}_{-1/\xi}$, and thus $X \in \mathrm{MDA}_\xi$. We now show that (8) holds. Thanks to Proposition 1, it suffices to provide a proof for $q = p_{\mathrm{GZD}}$. We have

$$\frac{\mathrm{pr}(X = k_u + u \mid X \geqslant u)}{p_{\mathrm{GZD}}(k_u; \xi u, \xi)} = \frac{\bar{F}(u + k_u)/\bar{F}(u)}{(1 + k_u/u)^{-1/\xi-1}} \frac{\sum_{i=0}^{\infty}(1 + i/u)^{-1/\xi-1}}{\sum_{i=0}^{\infty} \bar{F}(u + i)/\bar{F}(u)}.$$

First, by the uniform convergence (12) and the the fact that k_u grows at most linearly fast, we conclude that

$$\frac{\bar{F}(u + k_u)/\bar{F}(u)}{(1 + k_u/u)^{-1/\xi-1}} \to 1,$$

as $u \to \infty$ over the integers. Second, Lemma 1 yields

$$u^{-1}\sum_{i=0}^{\infty}(1 + i/u)^{-1/\xi-1} \to \xi.$$

Third, it follows from (13) that for $\epsilon \in (0, 1/\xi)$, there exists $u_\epsilon > 0$ such that for $u \geqslant u_\epsilon$,

$$u^{-1}\sum_{i=0}^{\infty} \bar{F}(u + i)/\bar{F}(u) \leqslant u^{-1} e^{\epsilon} \sum_{i=0}^{\infty} \left(1 + \frac{i}{u}\right)^{-1-1/\xi+\epsilon} \to \frac{\xi e^{\epsilon}}{1 - \xi\epsilon},$$

using Lemma 1 once again. A similar lower bound can be found in the same manner. Now letting $\epsilon \to 0$, this completes the proof of (8).

Let us now prove the second part of the theorem. For large integers u,

$$\mathrm{pr}(X = k + u \mid X \geqslant u) = \frac{\bar{F}(k + u)/\bar{F}(u)}{\sum_{i=0}^{\infty} \bar{F}(i + u)/\bar{F}(u)}.$$

We have for every $i = 0, 1, 2, \ldots,$

$$\bar{F}(i + u)/\bar{F}(u) = \frac{c(i + u)}{c(u)} \exp\left\{-\int_0^i 1/a(u + y)dy\right\} \to e^{-i/\sigma}$$

as $u \to \infty$. Since $\sigma > 0$, the dominated convergence theorem gives us

$$\sum_{i=0}^{\infty} \bar{F}(i + u)/\bar{F}(u) \to \sum_{i=0}^{\infty} e^{-i/\sigma} = 1/\big(1 - e^{-1/\sigma}\big),$$

showing (9). Finally, it follows from

$$p_X(n) = c(n) \exp\left\{-\int_0^n \frac{1}{a(y)} dy\right\}$$

for all n and $a(y) \to \sigma \in (0, \infty)$ that

$$\lim_{n\to\infty} \frac{p_X(n)}{\mathrm{pr}(X \geqslant n)} = 1 - e^{-1/\sigma} \in (0, \infty),$$

which immediately implies that $X \in \text{D-MDA}_0$ as well. □

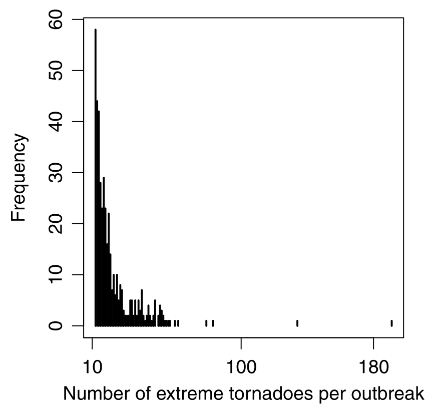

Figure 2: Frequency plot of the number of extreme tornadoes per outbreak for the 435 outbreaks with 12 or more extreme tornadoes in the United States between 1965 and 2015.

Example 1. *The probability mass function p_X of a Poisson distribution with rate $\lambda > 0$ coincides on $k = 0, 1, 2, \ldots$ with the function*

$$g(x) = \frac{\lambda^x e^{-\lambda}}{\Gamma(x+1)},$$

a continuous function on $\mathbb{R}_+$ satisfying $\lim_{x\to\infty} g(x) = 0$. Moreover,

$$\frac{d}{dx}\log g(x) = -\psi_0(x+1) + \log\lambda,$$

where ψ_0 is the polygamma function of order 0. Since $\psi_0(x) \to \infty$ as $x \to \infty$, we see that $g'(x) < 0$ for x sufficiently large. Therefore, $\bar{F}_Y(x) = g(x)/g(d)$ is a survival function on $[d, \infty)$ for some $d \geqslant 0$. Furthermore,

$$\frac{d}{dx}\left(-\frac{1}{g'(x)}\right) = -\frac{\psi_1(x+1)}{\{\psi_0(x+1) - \log\lambda\}^2},$$

where $\psi_1 = \psi_0'$ is is the polygamma function of order 1. Since $\psi_1(x) \to 0$ as $x \to \infty$, we conclude that F_Y satisfies the von Mises condition, with the auxiliary function $a(x) = \{\psi_0(x+1) - \log\lambda\}^{-1} \to 0$ as $x \to \infty$. Therefore, the Poisson probability mass function is in D-MDA$_0$.

Similarly, the probability mass function of the negative binomial distribution with probability of success $p \in (0, 1)$ and number of successes $r > 0$ is also in D-MDA$_0$ because it coincides on $\{0, 1, 2, \ldots\}$ with the function

$$g(x) = \frac{p^r}{\Gamma(r)}\frac{\Gamma(x+r)}{\Gamma(x+1)}(1-p)^x,$$

a continuous function on $\mathbb{R}_+$. It is simple to check that $\lim_{x\to\infty} g(x) = 0$, and $g'(x) < 0$ for x large enough, so that $\bar{F}_Y(x) = g(x)/g(d)$ is a survival function on $[d, \infty)$ for some $d \geqslant 0$. Furthermore, $g(x) \sim cx^{r-1}(1-p)^x$ for large x, where c is a positive constant. Therefore, $\bar{F}_Y$ is of the form (7) with the auxiliary function

$$a(x) = \frac{1}{-\log(1-p) - (r-1)/x}, \quad x \text{ large},$$

and so it converges to $-1/\log(1-p)$ as $x \to \infty$.

Acknowledgement

The first author would like to thank Robin Evans, Gesine Reinert, David Steinsaltz and Jonathan Tawn for valuable advice and encouragement. We also thank Michael Tippett for providing the tornado data, and anonymous referees for their comments.

Funding

The first author is grateful to the Berrow Foundation for financial support. This research was partially supported by the ARO grant W911NF-12-10385.

References

Ahmad T, Gaetan C, Naveau P (2022). Modelling of discrete extremes through extended versions of discrete generalized Pareto distribution. arXiv preprint: https://arxiv.org/abs/2210.15253.

Anderson CW (1970). Extreme value theory for a class of discrete distributions with applications to some stochastic processes. *Journal of Applied Probability*, 7: 99–113. https://doi.org/10.2307/3212152

Anderson CW (1980). Local limit theorems for the maxima of discrete random variables. *Mathematical Proceedings of the Cambridge Philosophical Society*, 88(1): 161–165. https://doi.org/10.1017/S0305004100057443

Arnold BC (1983). *Pareto Distribution*. International Cooperative Publishing House, Maryland.

Arnold TB, Emerson JW (2011). Nonparametric goodness-of-fit tests for discrete null distributions. *The R Journal*, 3(2): 34–39. https://doi.org/10.32614/RJ-2011-016

Axtell RL (2001). Zipf distribution of US firm sizes. *Science*, 293: 1818–1820. https://doi.org/10.1126/science.1062081

Bingham NH, Goldie CM, Teugles JL (1989). *Regular Variation*. Cambridge University Press.

Booth AD (1967). A law of occurrences for words of low frequency. *Information and Control*, 10(4): 386–393. https://doi.org/10.1016/S0019-9958(67)90201-X

Buddana A, Kozubowski TJ (2014). Discrete Pareto distributions. *Economic Quality Control*, 29(2): 143–156. https://doi.org/10.1515/eqc-2014-0014

Charpentier A, Flachaire E (2019). Extended scale-free networks. arXiv preprint: https://arxiv.org/abs/1905.10267.

Clauset A, Shalizi CR, Newman MEJ (2009). Power-law distributions in empirical data. *SIAM Review*, 51(4): 661–703. https://doi.org/10.1137/070710111

Davison AC, Smith RL (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society, Series B, Methodological*, 52(3): 393–442.

Dkengne PS, Eckert N, Naveau P (2016). A limiting distribution for maxima of discrete stationary triangular arrays with an application to risk due to avalanches. *Extremes*, 19(1): 25–40. https://doi.org/10.1007/s10687-015-0234-0

Embrechts P, Klüppelberg C, Mikosch T (2013). *Modelling Extremal Events: For Insurance and Finance*. Springer, New York.

Evin G, Sielenou PD, Eckert N, Naveau P, Hagenmuller P, Morin S (2021). Extreme avalanche cycles: Return levels and probability distributions depending on snow and meteorological conditions. *Weather and Climate Extremes*, 33: 100344. https://doi.org/10.1016/j.wace.2021.100344

Gabaix X (1999). Zipf's law and the growth of cities. *American Economic Review*, 89(2): 129–132. https://doi.org/10.1257/aer.89.2.129

Ghosh I, Alzaatreh A, Hamedani G (2023). A new class of discrete distribution arising as an analogue of gamma-lomax distribution: Properties and applications. In: *G Families of Probability Distributions*, 181.

Hamilton BE, Martin JA, Osterman MJ, Curtin MA, Mathews TJ (2015). Births: Final data for 2014. *National Vital Statistics Reports*, 64(1): 12.

Hitz AS (2016). Modelling of extremes, Ph.D. thesis, University of Oxford.

Koh J (2023). Gradient boosting with extreme-value theory for wildfire prediction. *Extremes*, 26: 273–299. https://doi.org/10.1007/s10687-022-00454-6

Koutsoyiannis D (2023). Knowable moments in stochastics: Knowing their advantages. *Axioms*, 12(6): 590. https://doi.org/10.3390/axioms12060590

Kozubowski TJ, Panorska AK, Forister ML (2015). A discrete truncated Pareto distribution. *Statistical Methodology*, 26: 135–150. https://doi.org/10.1016/j.stamet.2015.04.002

Krishna H, Pundir PS (2009). Discrete Burr and discrete Pareto distributions. *Statistical Methodology*, 6(2): 177–188. https://doi.org/10.1016/j.stamet.2008.07.001

Mandelbrot B (1953). Contribution à la théorie mathématique des jeux de communication. *Publications de l'Institut de statistique de l'Université de Paris.*

New B, Pallier C, Brysbaert M, Ferrand L (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, 36(3): 516–524. https://doi.org/10.3758/BF03195598

Patel L, Shand L, Tucker JD, Huerta G (2021). Spatio-temporal extreme event modeling of terror insurgencies. arXiv preprint: https://arxiv.org/abs/2110.08363.

Pickands J III (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, 3: 119–131.

Prieto F, Gómez-Déniz E, Sarabia JM (2014). Modelling road accident blackspots data with the discrete generalized Pareto distribution. *Accident Analysis and Prevention*, 71: 38–49. https://doi.org/10.1016/j.aap.2014.05.005

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Ranjbar S, Cantoni E, Chavez-Demoulin V, Marra G, Radice R, Jaton K (2022). Modelling the extremes of seasonal viruses and hospital congestion: The example of flu in a swiss hospital. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, 71(4): 884–905. https://doi.org/10.1111/rssc.12559

Resnick SI (1987). *Extreme Values, Regular Variation, and Point Processes*. Springer, New York.

Shimura T (2012). Discretization of distributions in the maximum domain of attraction. *Extremes*, 15(3): 299–317. https://doi.org/10.1007/s10687-011-0137-7

Tippett MK, Lepore C, Cohen JE (2016). More tornadoes in the most extreme US tornado outbreaks. *Science*, 354(6318): 1419–1423. https://doi.org/10.1126/science.aah7393

Valiquette S, Toulemonde G, Peyhardi J, É Marchand , Mortier F (2023). Asymptotic tail properties of poisson mixture distributions. *Stat*, 12(1): e622. https://doi.org/10.1002/sta4.622