The Role of Extreme Value Theory in Statistics

Shobhit Sabharwal

Mentor: Adrien da Silva
Directed Reading Program, McGill Math and Stats

May 31st, 2026

Abstract

Extreme Value Theory, as the name would suggest, is a framework used to model extreme values of random samples of random variables (maximum and minimum). It is particularly useful in modelling rare events without much assumption on the underlying distribution. The theory relies on asymptotic arguments and approximations similar to the usage of the Central Limit Theorem for means. Key theorems include the Fisher- Tippett-Gnedenko theorem and the Pickands–Balkema–De Haan theorem.

Contents

1 Introduction

1.1 The Motivation for Extreme Value Theory (EVT)

Many real-world situations and applications deal with extremes instead of averages. For example, in finance, the goal is to maximize profit and minimize risk. In civil engineering, the goal is to build structures that are capable of handling forces far beyond what is expected for a high level of safety. As the world is full of uncertainty and risks, we are forced to use statistical practices that target extremes to ensure unlikely events don't bring catastrophic consequences. However, it may also be used to model extreme events that can bring positive outcomes. Regardless, there's a need to model unlikely events which is the primary role of EVT. Thus, it is most common to use EVT on time series data.

Famous contributors to Extreme Value Theory include Tippett, Fisher, and Gnedenko who are responsible for the first EVT theorem which is considered to be the foundation of the field. A notable historical application of EVT is its use in the aftermath of the North Sea flood of 1953, in Netherlands, that caused excessive flooding and claimed over 1800 lives largely due to the fact that the dikes weren't properly built and maintained. Following this event, there was a need to redesign a new dike that could withstand heavy storms with a comfortable margin of safety. Using EVT methods, the new dike was set to withstand a 1-in-10,000 year event.

1.2 Preliminary Knowledge about Statistics

This text will assume background knowledge at the level of a first course in probability to introduce the role of EVT in statistics. Statistics as a field can be split into inferential and descriptive counterparts. Descriptive statistics focuses purely on understanding observations. Inferential statistics, on the other hand, takes it a step further by using data to infer properties of the underlying data generating process. In this regard, if the observed data does not include all possible observations, then it is considered to be a sample from a larger population. We can preserve relevant information from the data through the use of 'statistics' which are functions of the data (not to be confused with statistics as a discipline). In a sense, statistics allow us to reduce the dimensionality of the data we are working with by compressing a

vector of data into a lower-dimensional vector, often a scalar. Some common statistics are sample mean and sample variance.

Definition 1 (Statistics)**.** Let X_1, X_2, ..., X_n be a collection of random variables representing our data. A statistic is a function T: $\mathbb{R}^n \to \mathbb{R}^m$, ($X_1$, X_2, ..., X_n) $\mapsto f(X_1, X_2, ..., X_n)$ for some function f, where $m < n$.

Definition 2 (Order Statistics)**.** The order statistic $X_{(i)}$ is a random variable whose realization is equal to the i^{th} largest observation:

$$\min\{X_1, X_2, ..., X_n\} = X_{(1)} \text{ and } \max\{X_1, X_2, ..., X_n\} = X_{(n)} = M_n.$$

EVT can be considered a subfield of inferential statistics that primarily focuses on estimating the odds of extreme events given sample data. Thus, maximum and minimum order statistics will be necessary to develop the theory. This is in contrast with the mean which is foundational subject in statistics, with key theorems such as the Law of Large Numbers (LLN) and the Central Limit Theorem (CLT) forming the backbone of the discipline. However, there is a similarity in the fact that asymptotic results can be found for extreme values.

Theorem 1 (Law of Large Numbers)**.** *Let X_1, X_2, ..., X_n be a collection of i.i.d. random variables of size n from the same distribution with finite population mean μ, then*

$$\bar{X}_n \xrightarrow{p} \mu.$$

Theorem 2 (Central Limit Theorem)**.** *Let X_1, X_2, ..., X_n be a collection of i.i.d. random variables of size n from the same distribution with finite population mean μ and population standard deviation σ, then*

$$\frac{(\bar{X}_n - \mu)\sqrt{n}}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1).$$

1.3 Maximum Likelihood Estimation (MLE)

Inferential statistics can be divided into parametric, nonparametric, and semiparametric (hybrid) frameworks. Parametric statistics works under the assumption that the data generating process comes from a known distribution parametrized by a set of parameters. Nonparametric statistics, on the other hand, makes little assumption on the data generating process. This

text will mainly stay in the realm of parametric statistics, but it is important to mention that order statistics play a crucial role in nonparametric statistics. When population parameters of the distribution are unknown, point estimation can be done using a statistic of the sample data to approximate the value of parameters. Maximum Likelihood Estimation (MLE) is a popular method of estimation that uses the likelihood function.

Definition 3 (Likelihood)**.** Let X_1, X_2, ..., X_n be a collection of random variables representing our data. The likelihood function is the defined as the following:

$$\mathcal{L}(\theta|X_1, X_2, ..., X_n) = f(X_1, X_2, ..., X_n|\theta) = f_\theta(X_1, X_2, ..., X_n),$$

where the data X_i are considered fixed and the parameter θ is variable. Here, f_θ is the joint distribution of the collection of random variables parametrized by θ which can be a vector of parameters. Under the independent and identically distributed assumption (i.i.d.) assumption,

$$\mathcal{L}(\theta|X_1, X_2, ..., X_n) = f(X_1, X_2, ..., X_n|\theta) = \prod_{i=1}^{n} f_\theta(X_i).$$

Maximizing the likelihood function while varying θ allows us to find the value of the parameter that best suits the data. However, this is an estimate $\hat{\theta}$ of the true population parameter θ as the sample data does not contain the full picture regarding the data generating process.

Definition 4 (Maximum Likelihood Estimator)**.** The Maximum Likelihood Estimator is defined as the following:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta|X_1, X_2, ..., X_n).$$

Maximum Likelihood Estimation is the process that allows us to find $\hat{\theta}_{\text{MLE}}$ which is commonly obtained by differentiating the likelihood with respect to θ and setting it equal to 0 to find the maximum. The log of the likelihood is commonly used for ease of algebraic manipulation as well as numerical stability as it is a monotonically increasing function of $\mathcal{L}$ that is maximized at the same value and which avoids multiplying small quantities as products are turned into sums. When it cannot be found analytically, numerical algorithms are used instead.

2 Principles of Extreme Value Theory

2.1 Distribution of Order Statistics

Maximums and minimums can be viewed as two sides of the same coin when dealing with extremes as $\max\{X_1, X_2, ..., X_n\} = -\min\{-X_1, -X_2, ..., -X_n\}$. Thus, I will mainly work with maximums when dealing with extreme values. Under the i.i.d. assumption, it is possible to find the CDF of $X_{(n)}$ for n data points:

$$F_{X_{(n)}}(x) = P(X_{(n)} < x) \quad \text{(by definition of CDF)} \tag{1}$$
$$= P(X_1 < x, ..., X_n < x) \quad \text{(by property of maxima)} \tag{2}$$
$$= \prod_{i=1}^{n} F_{X_i}(x) \quad \text{(by independence of } X_i s\text{)} \tag{3}$$
$$= F_X^n(x) \quad \text{(by identical distribution of } X_i s\text{)} \tag{4}$$

Similarly for minimums, we find that the CDF of $X_{(1)}$ is

$$F_{X_{(1)}}(x) = P(X_{(1)} < x) = 1 - P(X_{(1)} > x) \tag{5}$$

which simplifies to

$$1 - P(X_1 > x, ..., X_n > x) = 1 - \prod_{i=1}^{n}(1 - F_{X_i}(x)) = 1 - (1 - F_X(x))^n \tag{6}$$

The important property that allows us to derive the CDFs of both the maximum and minimum is that all observations must be larger than the min and smaller than the max. This may seem tautological, nonetheless, it allows for useful simplifications under i.i.d. assumptions.

When looking at the CDF of the max, it would be interesting to analyze the limiting behaviour as n approaches infinity. Given the fact that the CDF is bounded in [0,1], there are only two possibilities for the limiting behaviour. Either it converges to 0 if $F_X(x) < 1$ or equals 1 if $F_X(x) = 1$ for a certain

x. The second scenario holds if there exists $x_{\max}$ such that
$x_{\max} = \inf\{x : F(x) = 1\}$ (upper end-point of F) which entails the following:

$$\lim_{n\to\infty} (F_X(x))^n = \begin{cases} 0 & \text{if } x < x_{\max} \\ 1 & \text{if } x \geq x_{\max} \end{cases} \tag{7}$$

The result in (4) suggests that the CDF of the max converges to the indicator function $\mathbb{1}_{\{x \geq x_{max}\}}$. This would suggest that the PDF of the max converges to the Dirac delta function centered at $x_{\max}$.
Thus,

$$f_{M_n}(x) \xrightarrow[n\to\infty]{} \delta(x - x_{\max}),$$

where M_n is the maximum of n data points (M_n is the same as $X_{(n)}$ for a sample of size n). In other words,

$$M_n \xrightarrow{d} x_{\max} \quad (\iff M_n \xrightarrow{p} x_{\max} \text{ as } x_{\max} \text{ is a constant})$$

and the distribution of M_n degenerates to a point mass at $x_{\max}$.

We will later see that with a good choice of normalizing sequence of constants $\{a_n\}$ and $\{b_n\}$, $M_n^* = (M_n - \{b_n\})/\{a_n\}$ converges to a limiting distribution that isn't degenerate. In fact, it must converge to a specific family of distributions. Generalizations can also be made to the r^{th} order statistic in a sample of n points, but the focus of this text will be on modelling maxima.

2.2 Block Maxima Approach

2.2.1 Generalized Extreme Value Distribution

The foundational theorem for Extreme Value Theory is called the Fisher-Tippett-Gnedenko theorem (also known as the Extremal Types Theorem). This theorem builds on the previously stated idea of finding an asymptotic distribution for the sample maximum using normalizing constants. There is an assumption here that the data points are i.i.d. for the maxima.

Theorem 3 (Fisher-Tippett-Gnedenko Theorem)**.** *Given normalizing sequence of constants $\{a_n\}$ and $\{b_n\}$, where $a_n > 0$, and $M_n^* = (M_n - \{b_n\})/\{a_n\}$ such that $P(M_n^* \leq z) = F_{M_n^*}(z) \to G(z)$ as $n \to \infty$ for some non-degenerate CDF G, then G must belong to one of three families:*

1. *Gumbel:* $G(z) = e^{-e^{-\frac{z-b}{a}}}, \quad z \in \mathbb{R}$ (8)

2. *Fréchet:* $G(z) = \begin{cases} 0 & z \leq b \\ e^{-(\frac{z-b}{a})^{-\alpha}} & z > b \end{cases}$ (9)

3. *Weibull:* $G(z) = \begin{cases} e^{-[-(\frac{z-b}{a})^{\alpha}]} & z < b \\ 1 & z \geq b \end{cases}$ (10)

Here, b is the location parameter and a is the scale parameter, α is the shape parameter present in Fréchet and Weibull, but not in Gumbel. These three classes of functions can be generalized into one family with an added parameter ξ.

Definition 5. (Generalized Extreme Value Distribution) Given the same conditions in Theorem 1 where $F_{M_n^*}(\mathrm{z}) \to \mathrm{G(z)}$, G must be a member of the GEV family of CDFs defined as the following:

$$G_{\xi,\mu,\sigma}(z) = \begin{cases} \exp(-[1+\xi(\frac{z-\mu}{\sigma}))]^{-\frac{1}{\xi}}) & \xi \neq 0 \\ \exp\left(\exp\left(-(\frac{z-\mu}{\sigma})\right)\right) & \xi = 0 \end{cases} \tag{11}$$

defined over $\{z : [1+\xi(\frac{z-\mu}{\sigma})] > 0\}$, where $\mu, \xi \in \mathbb{R}$ and $\sigma > 0$

In the generalized setting, ξ is the shape parameter and $\xi > 0$ corresponds to the Fréchet case, $\xi = 0$ to the Gumbel case, and $\xi < 0$ to the Weibull case. In this formula, μ is the location parameter and σ is the scale parameter. Since this is a limiting result, for large n, M_n approximately follows a $G_{\xi,\mu,\sigma}(z)$ distribution.

2.2.2 Interpretation and Usage

Essentially, Theorem 1 states that the CDF of normalized maxima converge to one distribution if proper normalizing constants can be found which contrasts very well with the CLT and the convergence of the normalized mean to the standard normal. However, μ and σ don't represent mean and standard deviation in the GEV distribution as they represent location and scale, respectively. Additionally, the choice of sequences doesn't matter as long as the

CDF converges to a non-degenerate limit, ξ will remain the same regardless of the choice of constants $\{a_n\}$ and $\{b_n\}$, but location and scale parameters may differ. If a GEV exists for a density function F_X (from which F_{M_n} is obtained), then we say that F_X is in the Maximum Domain of Attraction (MDA) of a GEV. A similar distribution can be found for minima. However, as previously stated, we can use the duality between minima and maxima and use the GEV for maxima on minima by negating the value of the data points.

In practice, we use GEV in the block maxima approach where the data is split into periods (blocks) of equal length. From each block, we take the maxima and put them in a set, discarding the other data points. Then we fit the GEV on the maxima which is most commonly done using MLE to find the parameters that maximize the likelihood. In this context, there is generally no analytical solution, so we must use numerical optimization algorithms carefully to find the right model without violating constraints.

We can then find the return level z_p where $G(z_p) = 1 - p$. The return level z_p would have return period 1/p which means we expect on average z_p to be exceed in 1 in 1/p periods.

There is also bias-variance tradeoff which is a common theme in statistics. Bias represents how far the expected value of an estimator is from the true target value. Variance is the variability of our estimate. When blocks are of short length, there are more maxima to work with which lowers variance, but also increases bias as some maxima may not truly be "extreme" events. Similarly, with large blocks, there are very few maxima in the set which means we most likely have "extreme" events which leads to lower bias, but given the lack of data points, we may be in a high variance case. Usually, a suitable block size is given by context.

2.3 Peaks Over Threshold Approach

2.3.1 Generalized Pareto Distribution

A big shortcoming of the block maxima approach is that a lot of data is discarded which could have been significant and useful for fitting the model. This is the case if one block contains multiple extreme events, but only the single most extreme is kept. The Peaks Over Threshold (POT) approach addresses this issue by working with exceedances over a threshold instead of maxima. Fortunately for us, there exists an asymptotic model for exceedances similar to the GEV.

Theorem 4. *(Pickands–Balkema–De Haan theorem) Given independent random variables X_1, X_2, ..., X_n with the same F_x that is in the domain of attraction of a GEV, for a sufficiently large threshold u, the CDF of $(X_i - u)$ conditionned on $X_i > u$ converges to a Generalized Pareto Distribution:*

$$H(y) = \begin{cases} 1 - (1 + \frac{\xi y}{\sigma_u})^{-\frac{1}{\xi}} & \xi \neq 0 \\ 1 - e^{(-y/\sigma_u)} & \xi = 0 \end{cases} \tag{12}$$

defined over $\{y : y > 0$ and $(1 + \frac{\xi y}{\sigma_u}) > 0\}$ and $\sigma_u = \sigma_{u_0} + \xi(u - u_0)$, where $\xi \in \mathbb{R}$ and $\sigma_u > 0$.

The value of ξ is unique and the same as the corresponding GEV associated with M_n due to a duality between both approaches/distributions. Additionally, the value of σ_u is linear with respect to a change in threshold u_0 to another sufficient threshold u.

2.3.2 Interpretation and Usage

We follow a procedure similar to the one used for the block maxima approach when using the GPD, but with exceedances instead. We must first find a suitable threshold to work with. There is similarly a bias-variance tradeoff here similar to the block size issue. A low threshold breaks the asymptotic nature, leading to a higher bias, while a high threshold excludes a lot of data points leading to high variance. A good threshold would be a compromise between both sides, usually the lowest threshold that still leads to good asymptotic behaviour. We would then collect the exceedances and store the excesses $(y_i = x_i - u)$ and then fit the GPD on it using maximum likelihood estimation. Similarly to the block maxima case, there is no closed form solution, so numerical methods need to be used.

2.4 Final Remarks

Overall, EVT is a very useful statistical framework that requires only weak distributional assumptions. It applies asymptoptic behaviour similar to the CLT to find a limiting distribution, but for sample extremes rather than sample means. The first approach which is the block maxima one uses the GEV distribution, while the second POT approach uses the GPD distribution. Both approaches, however, share the same value for the parameter ξ for a given distribution of the data if that distribution is in the MDA of a GEV. Extreme Value Theory serves as a useful analogue to average-based approaches in statistics showcasing that even extremes can be modelled and accounted for without even knowing the underlying distribution.

BIBLIOGRAPHY

1. S. Coles, An Introduction to Statistical Modeling of Extreme Values, Springer Series in Statistics, Springer, London, 2004, Softcover reprint of the original 1st ed. 2001.

2. A. Siffer, P.-A. Fouque, A. Termier, and C. Largouet, Anomaly Detection in Streams with Extreme Value Theory, Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 1067–1075.

3. QRM Tutorial, Extreme value theory (QRM Chapter 5), YouTube Video, 2018, Available at `https://www.youtube.com/watch?v=IiOSxaF5oxo&t=5150s`; accessed May 28, 2026.