The Little Hero of Haarlem

A peek at Extreme Value Theory and its Applications

Dakshesh Vasan

Dept. of Mathematics and Statistics
Indian Institute of Science Education and Research (IISER) Kolkata

DMS Day, February 25, 2023

The Hero of Haarlem

The Dykes of Netherlands

Netherlands in an Alternate Universe

The Height (is)of the Problem

Question

How high should the dikes be built then, in order to satisfy budget constraints and safety requirements?

The Height (is)of the Problem

Question

How high should the dikes be built then, in order to satisfy budget constraints and safety requirements?

- Could take a Probabilistic approach. Define suitable random variable, and estimate using empirical methods.

IISER KOLKATA

The Height (is)of the Problem

Question

How high should the dikes be built then, in order to satisfy budget constraints and safety requirements?

- Could take a Probabilistic approach. Define suitable random variable, and estimate using empirical methods.
- However, high tides are uncommon.

IISER KOLKATA

The Height (is)of the Problem

Question

How high should the dikes be built then, in order to satisfy budget constraints and safety requirements?

- Could take a Probabilistic approach. Define suitable random variable, and estimate using empirical methods.
- However, high tides are uncommon.
- There are records of storms and high tides for the town of Delfzijil for over the past 100 years. 1877 storm surges and ZERO floods.

The Height (is)of the Problem

Question

How high should the dikes be built then, in order to satisfy budget constraints and safety requirements?

- Could take a Probabilistic approach. Define suitable random variable, and estimate using empirical methods.
- However, high tides are uncommon.
- There are records of storms and high tides for the town of Delfzijil for over the past 100 years. 1877 storm surges and ZERO floods.
- What can we do then?

Extreme Value Theory - Introduction

- Extreme Value Theory is an area of probability theory dedicated to understanding the maxima of a random sample.

Extreme Value Theory - Introduction

- Extreme Value Theory is an area of probability theory dedicated to understanding the maxima of a random sample.
- Over the past 50 years, extreme value theory has been developed extensively to study rare events and extremities and solve associated problems in day-to-day life.
- The Theory provides a solid theoretical basis for extrapolation of whatever little information we can get from an empirical distribution function near the boundary of the sample.

Extreme Value Theory - Introduction

Say $\{X_n\}$ is a sequence of independent and identically distributed random variables with distribution F.

Central Limit Theorem (CLT)
$\sum_{i=1}^{n} X_i \rightarrow$?

Extreme Value Theory - Introduction

Say $\{X_n\}$ is a sequence of independent and identically distributed random variables with distribution F.

Central Limit Theorem (CLT)
$\sum_{i=1}^{n} X_i \rightarrow$?
A car's tires being damaged due to some wear and tear from daily use.

Extreme Value Theory (EVT)
$max(X_1, X_2, ..., X_n) \rightarrow$?

Extreme Value Theory - Introduction

Say $\{X_n\}$ is a sequence of independent and identically distributed random variables with distribution F.

Central Limit Theorem (CLT)

$\sum_{i=1}^{n} X_i \rightarrow$?

A car's tires being damaged due to some wear and tear from daily use.

Extreme Value Theory (EVT)

$max(X_1, X_2, ..., X_n) \rightarrow$?

A car's tires being damaged by a sudden excessive damage from an accident one day.

Doing away with Degenerates!

- Now, $max(X_1, X_2, ..., X_n) \xrightarrow{p} x_*$ as n $\to \infty$, where $x_* = \{x | F(x) < 1\}$.
- Degenerate random variable!

IISER KOLKATA

Doing away with Degenerates!

- Now, $max(X_1, X_2, ..., X_n) \xrightarrow{p} x_*$ as n $\to \infty$, where $x_* = \{x | F(x) < 1\}$.
- Degenerate random variable!

Our Interest

To obtain a non-degenerate distribution G such that for some $\{a_n\} \subset (0, \infty)$ and $\{b_n\} \subset \mathbb{R}$,

$$\frac{max(X_1, X_2, ..., X_n) - b_n}{a_n} \xrightarrow{d} G; \text{ as n } \to \infty$$

.

Doing away with Degenerates!

- Now, $max(X_1, X_2, ..., X_n) \xrightarrow{p} x_*$ as n $\to \infty$, where $x_* = \{x | F(x) < 1\}$.
- Degenerate random variable!

Our Interest

To obtain a non-degenerate distribution G such that for some $\{a_n\} \subset (0, \infty)$ and $\{b_n\} \subset \mathbb{R}$,

$$\frac{max(X_1, X_2, ..., X_n) - b_n}{a_n} \xrightarrow{d} G; \text{ as n} \to \infty$$

.

- Through an appropriate normalization, we are trying to obtain a non-degenerate distribution(Similar to CLT).

IISER KOLKATA

The Sealed Fate of any Extreme Value Distribution

Fisher and Tippet (1928), Gnedenko (1943)

Let $\{X_n\}$ be a sequence of independent and identically distributed random variables with distribution F. The class of extreme value distributions, ie, the class of non-degenerate distributions that can occur as a limit distribution for $\lim_{n\to\infty} \frac{max(X_1,X_2,....,X_n)-b_n}{a_n}$ where $\{a_n\} \subset (0,\infty)$ and $\{b_n\} \subset \mathbb{R}$ is G_γ, where :

- $G_\gamma(x) = e^{-(1+\gamma x)^{-1/\gamma}}, 1+\gamma x > 0$ when $\gamma \neq 0$ or
- $G_0(x) = e^{-e^{-x}}, x > 0$ when $\gamma = 0$.

The Sealed Fate of any Extreme Value Distribution

Fisher and Tippet (1928), Gnedenko (1943)

Let $\{X_n\}$ be a sequence of independent and identically distributed random variables with distribution F. The class of extreme value distributions, ie, the class of non-degenerate distributions that can occur as a limit distribution for $\lim_{n\to\infty} \frac{max(X_1,X_2,....,X_n)-b_n}{a_n}$ where $\{a_n\} \subset (0,\infty)$ and $\{b_n\} \subset \mathbb{R}$ is G_γ, where :

- $G_\gamma(x) = e^{-(1+\gamma x)^{-1/\gamma}}, 1+\gamma x > 0$ when $\gamma \neq 0$ or
- $G_0(x) = e^{-e^{-x}}, x > 0$ when $\gamma = 0$.

Estimating γ

There are several estimators of γ like the Moment Estimator, MLE, Pickand's, Hill and so on.

IISER KOLKATA

The Sealed Fate of any Extreme Value Distribution

Alternative Formulation

- Define U(y)= inf$\{$x$|$ $\frac{1}{1-F(x)} \geq$ y$\}$.
- This function returns the minimum value of x from amongst all those values whose c.d.f value exceeds 1-$\frac{1}{y}$.
- The theorem can be restated in terms of the 'quantile' type function as follows:
  when $\gamma \neq 0$, $\lim_{t\to\infty} \frac{U(tx)-U(t)}{a_{[t]}} = \frac{x^{\gamma}-1}{\gamma}, x > 0$
  when $\gamma = 0$, $\lim_{t\to\infty} \frac{U(tx)-U(t)}{a_{[t]}} = ln(x), x > 0$.
- This form becomes more convenient in practice. We will see it in action, right now!

Revisiting Netherlands with Extreme Value Theory

- Probability of a flood in a given year to be restricted to $\leq 10^{-4}$.

Revisiting Netherlands with Extreme Value Theory

- Probability of a flood in a given year to be restricted to $\leq 10^{-4}$.
- Delfzijil again, 1877 storm data -with high tide water levels recorded- over a 111 year period. Let the records of high tide water levels collected be the realizations of an iid sequence, $\{X_i\}_{i=1}^{1877}$ with distribution F.

Revisiting Netherlands with Extreme Value Theory

- Probability of a flood in a given year to be restricted to $\leq 10^{-4}$.
- Delfzijil again, 1877 storm data -with high tide water levels recorded- over a 111 year period. Let the records of high tide water levels collected be the realizations of an iid sequence, $\{X_i\}_{i=1}^{1877}$ with distribution F.
- We want the height at which the probability of a high tide exceeding the height $=10^{-4}$. We are okay with a dyke of this height, implying that we are okay if a flood occurs from a tide of a level higher than the dyke's height.

IISER KOLKATA

Revisiting Netherlands with Extreme Value Theory

- Probability of a flood in a given year to be restricted to $\leq 10^{-4}$.
- Delfzijil again, 1877 storm data -with high tide water levels recorded- over a 111 year period. Let the records of high tide water levels collected be the realizations of an iid sequence, $\{X_i\}_{i=1}^{1877}$ with distribution F.
- We want the height at which the probability of a high tide exceeding the height $=10^{-4}$. We are okay with a dyke of this height, implying that we are okay if a flood occurs from a tide of a level higher than the dyke's height.
- 111 years : 1877 storms :: $\frac{111}{1877}$ years : 1 storm.
- Then, we want that P(flood during one storm) $= \frac{111}{1877} x 10^{-4}$.

Revisiting Netherlands with Extreme Value Theory

- Probability of a flood in a given year to be restricted to $\leq 10^{-4}$.
- Delfzijil again, 1877 storm data -with high tide water levels recorded- over a 111 year period. Let the records of high tide water levels collected be the realizations of an iid sequence, $\{X_i\}_{i=1}^{1877}$ with distribution F.
- We want the height at which the probability of a high tide exceeding the height $=10^{-4}$. We are okay with a dyke of this height, implying that we are okay if a flood occurs from a tide of a level higher than the dyke's height.
- 111 years : 1877 storms :: $\frac{111}{1877}$ years : 1 storm.
- Then, we want that P(flood during one storm) $= \frac{111}{1877}x10^{-4}$.
- We are looking for h$=$ $\inf\{x|F(x) > 1 - (\frac{111}{1877} \times 10^{-4})\ \}$
$= \inf\{x \mid \frac{1}{1-F(x)} > (\frac{111}{1877} \times 10^{-4})\ \} \approx U(17 \times 10^4)$.

Revisiting Netherlands with Extreme Value Theory

- Probability of a flood in a given year to be restricted to $\leq 10^{-4}$.
- Delfzijil again, 1877 storm data -with high tide water levels recorded- over a 111 year period. Let the records of high tide water levels collected be the realizations of an iid sequence, $\{X_i\}_{i=1}^{1877}$ with distribution F.
- We want the height at which the probability of a high tide exceeding the height $=10^{-4}$. We are okay with a dyke of this height, implying that we are okay if a flood occurs from a tide of a level higher than the dyke's height.
- 111 years : 1877 storms :: $\frac{111}{1877}$ years : 1 storm.
- Then, we want that P(flood during one storm) $= \frac{111}{1877} x 10^{-4}$.
- We are looking for h$=$ inf$\{$x$|F(x) > 1 - (\frac{111}{1877}$ x $10^{-4})$ $\}$ $=$ inf$\{$x $|\frac{1}{1-F(x)} > (\frac{111}{1877}$ x $10^{-4})$ $\} \approx$ U(17 x 10^4).
- The highest order statistic available corresponds to U(19 x 10^2)!

Revisiting Netherlands with Extreme Value Theory

Problem

- The highest order statistic available corresponds to U(19 x 10^2)!

EVT to the Rescue!

However, using an approximation based on the Result stated earlier, we arrive at the following:

- Take t $=$ 19 x 10^2 and tx $=$ 17 x 10^4. Estimate U(17 x 10^4) from the empirical distribution and the highest order statistic available.
- Then, U(17 x 10^4) $\approx U(19x10^2) + a_{[t]}\frac{x^\gamma - 1}{\gamma}$ where x $= \frac{17x10^2}{19}$.
- Using estimates of $a_{[t]}$ from the order statistics, and a suitable estimator for γ, we can estimate the required height within a reliable confidence interval.

IISER KOLKATA

Other Applications of Extreme Value Theory

Business and Finance

- Actuarial Science: in an insurance firm, to avoid filing of a claim so large that it represents a threat to its solvency.
- Stock Markets: to decide on a big risky investment, while unable to afford a loss larger than a certain amount.

Natural Sciences

- Earth Sciences: to study extreme rainfall or rise of sea level along the coast, predict extreme climate changes and natural disasters.
- Chemical Sciences: to characterize food-processing systems.
- Physical Sciences: to study the physics of disordered systems.
- Life Sciences: to estimate the maximum possible life span of an individual.

Other Miscellaneous

- to estimate the ultimate sports records.
- to establish the safety of a runway.

IISER KOLKATA

A guide to choosing your major

Hypothesis

There exists a certain CGPA below 10, beyond which no CGPA can be obtained by a Mathematics Major at IISER Kolkata.

Proof Sketch

- Take the CGPA data of DMS, IISER Kolkata BS-MS graduates for over the past 5 years. Generate order statistics.
- Endpoint is finite (cannot exceed 10). It can be proved that this implies $\gamma \leq 0$. Say, it is estimated to be less than zero.
- Then, $\lim_{t\to\infty} \frac{U(\infty)-U(t)}{a_{[t]}} = \frac{-1}{\gamma}$.
- Then, estimate U(∞) $\approx$ U(t) - $\frac{a_{[t]}}{\gamma}$ using estimates from order statistics and estimators of γ to obtain a theoretically sound upper bound on a Mathematics CGPA at IISER Kolkata.

IISER KOLKATA

Acknowledgment

I am grateful to the organisers of this symposium for the opportunity to share something enjoyable and interesting, to all.

I thank Dr. Anirvan Chakraborty, for introducing me to Extreme Value Theory, and especially mentioning the Netherlands Problem and the Fisher-Tippet-Gnedenko Theorem which motivated me to look into EVT.

I also acknowledge the authors of 'Extreme Value Theory: An Introduction', Laurens Haan and Ana Ferreira. This book and its exercises are helping me grasp the basics of EVT, and I hope to continue positively.

References

- Extreme Value Theory: An Introduction, by Laurens Haan and Ana Ferreira.
- Estimating the conditional extreme-value index under random right-censoring, by Gilles Stupfler
- Records in Athletics Through Extreme-Value Theory, by John H. J. E INMAHL and Jan R. M AGNUS
- Steps in Applying Extreme Value Theory to Finance: A Review by Younes Bensalah
- Peter B. Skou, Stephen E. Holroyd, Frans Berg, Tutorial – applying extreme value theory to characterize food-processing systems, Journal of Chemometrics
- Universality classes for extreme-value statistics Jean-Philippe Bouchaud and Marc M ezard