Extreme Value Theory

Part I: Introduction

Gregory S. Karlovits, P.E., PH, CFM
US Army Corps of Engineers
Hydrologic Engineering Center

US Army Corps of Engineers

Hello everyone, I'm Greg Karlovits from the Hydrologic Engineering Center. Welcome to our course on statistical methods in hydrology. This video is part one of four on the topic of extreme value theory and will cover a short introduction to extreme value theory. Let's get started.

Extreme Value Theory

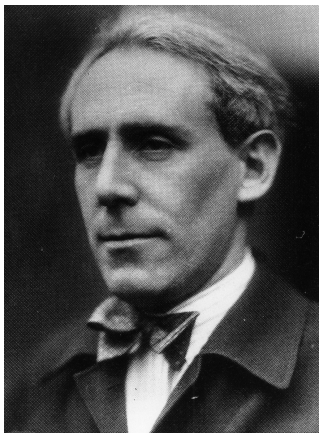
- Civil engineering is largely a practice of extremes
  - **minimum** shear strength and **maximum** shear stress for slope stability safety factor
  - **minimum** resistance and **maximum** load for LRFD
  - **minimum** time to collision and **maximum** traffic load
  - **minimum** and **maximum** annual flows on a river

2

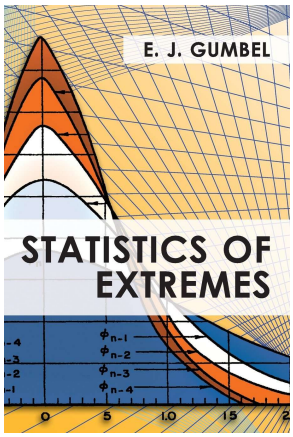
Civil engineering in general, but especially hydrology, is a practice of extremes. Most disciplines figure extremes into their design considerations. In geotechnical engineering, slope stabilities can be thought of as affected by the minimum shear strength of a soil and the maximum shear stress applied to it. In structural engineering, load and resistance factor design, also called limit state design, makes assumptions about the minimum resistance of a structure and the maximum load that is applied. Transportation engineers can look at driver behaviors and highway capacities and consider the extreme observations. And finally a topic we will see the most of in this class, minimum and maximum river discharges.

These minima and maxima are what extreme value theory is all about. The extremes we are talking about are usually the largest or smallest values in a sample. Oftentimes we are interested in what to expect from the largest or smallest values in a sample.

Emil Julius Gumbel



"It's impossible that the improbable will never happen."



3

Emil Julius Gumbel is responsible for much of what we know today about the behaviors of these extremes. Gumbel was a German mathematician forced out of Germany by the Nazis in 1932, after which he moved to France and then the United States where he taught until he died in 1966.

His 1958 book "Statistics of Extremes" is a true classic. It's not an easy read, but it is foundational for the topics we will see today. What Gumbel documented is that extremes, like the sample minimum, maximum, and so on, have a regular and predictable behavior that can be modeled in a meaningful way.

Gumbel synthesized the work of other prominent statisticians, such as Maurice Fréchet, Ronald Fisher, and others, and added the unifying details that built this comprehensive text that serves as the foundation for the analyses we do today.

Gumbel's EV Questions

- Does an individual observation in a sample taken from a distribution, alleged to be known, fall outside what may reasonably be expected?
- Does a series of extreme values exhibit a regular behavior?

4

Gumbel sought to answer two questions, which have impacts beyond just statistical curiosity.

The first question is another way of asking, "if I have a data point, how can I tell if this value is unreasonably large or small based on what I know about the population it came from?"

The second question is a little more interesting for design purposes, especially for risk-based design in hydrology. It asks, can we build a model for the extremes of a sample so that we can make predictions or inferences about them?

Extreme Value Theory

- We seek models for the behaviors of extremes
- Models are applied for making decisions
- Floods and droughts (hydrologic extremes) drive investment

5

Here are the three key reasons we study extreme value theory.

We want to be able to model the behavior of extreme events. These models help us make decisions by tying the magnitude of these extremes to how likely they are to occur. In the US Army Corps of Engineers, decisions are often made using a risk-informed decision making framework, which incorporates probability and consequence to make decisions on project investment. The statistics of extremes clearly come into play in these decisions because we need to estimate the probability of consequential events, which are at the extremes; in other words, floods and droughts.

Lecture Outline

1. Order Statistics
2. First Extreme Value Theorem
3. Second Extreme Value Theorem

6

This series will cover three main topics in extreme value theory. The first is order statistics, which deals with the regular behavior of ranked data. It builds the foundation to the two extreme value theorems, which provide models for the behavior of extreme events.

Thanks, and tune into the next video in this series on extreme value theory to learn about order statistics and their role in extreme value analysis.

Extreme Value Theory

Part II: Order Statistics and More

Gregory S. Karlovits, P.E., PH, CFM
US Army Corps of Engineers
Hydrologic Engineering Center

US Army Corps of Engineers

Hello everyone, I'm Greg Karlovits from the Hydrologic Engineering Center. Welcome to our course on statistical methods in hydrology. This video is part two of four on the topic of extreme value theory and will discuss order statistics and their role in extreme value analysis. Let's get started.

Order Statistics

Sample n = 10		Sorted	Order Statistics
X_1	19	24	$X_{(10)}$
X_2	20	22	$X_{(9)}$
X_3	8	20	$X_{(8)}$
X_4	15	19	$X_{(7)}$
X_5	9	16	$X_{(6)}$
X_6	22	15	$X_{(5)}$
X_7	24	14	$X_{(4)}$
X_8	16	12	$X_{(3)}$
X_9	14	9	$X_{(2)}$
X_{10}	12	8	$X_{(1)}$

Sample minimum $x_{(1)}$ 8

Sample maximum $x_{(n)}$ 24

Sample range $x_{(n)} - x_{(1)}$ 16

Sample median 15.5

n odd

$x_{\left(\frac{n+1}{2}\right)}$ *is an order statistic*

n even

$\frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n+1}{2}\right)}}{2}$ *is **not** an order statistic*

8

Order statistics are a way of looking at data based on rank. In our sample on the left, the data are listed in the order they were recorded. The first observation was 19, and then the next observation in time was 20, and so on.

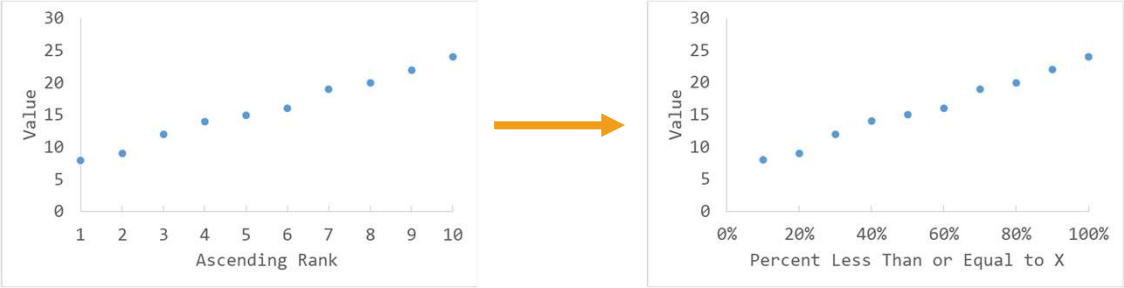
If we sort the data, we can summarize the data in a different way. Here we show the data sorted from largest to smallest, mainly because in hydrology we are most often concerned with the largest value. The order that these observations were observed is lost. Often though, that order isn't necessarily very important.

If we re-number the observations, where 1 is the smallest and n is the largest, (and n is the number of observations in the sample), we get the order statistics. Order statistics are differentiated by putting their rank index in parentheses. If you see x-subscript 1 in parentheses, that means the smallest observation, or the sample minimum. X-subscript-n in parentheses is the sample maximum. For this sample, x-sub-1 is 8, and x-sub-n is 24.

From the order statistics you can compute the range, which is always x-sub-n minus x-sub 1. For this sample it is 16.

One note is about the sample median. When n is an odd number, the sample median is an order statistic, meaning the median corresponds directly to one of the observations in the sample and can be found by looking at the x-sub-(n+1 / 2) order statistic. When n is an even number, as in this sample, the sample median is not an order statistic, but is an average of two order statistics.

Rank Plot



We can show the cumulative distribution for the sample based on the ranks, but how well do we think it reflects the population?

Do you think that another sample from this population will never exceed 24?

9

A rank plot is one way to look at a data set based on order statistics. Simply plot the rank of the data on X and the value of the data on Y, and you produce a rank plot. It is easy to convert the rank plot into an empirical quantile plot by dividing the rank by the sample size, which produces an estimate for the cumulative distribution function. However this has the limitation that it estimates the non-exceedance probability for the largest observation is 1, meaning that the largest value cannot be exceeded. It raises the question, do you really think that another sample from this population will never exceed the value 24?

Order Statistics of the Uniform Distribution

- The exceedance probabilities of a random sample of values from any population have a uniform distribution bounded on the interval (0, 1).

  $U(0, 1)$

- ***Why is this useful?***
  - We often are interested in an empirical estimate for the exceedance probabilities for values in a sample.

10

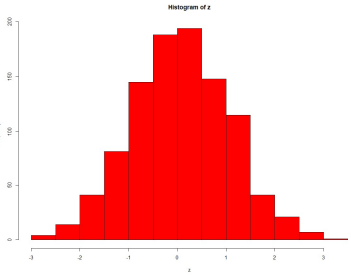
When a random sample from a population is taken, and the cumulative distribution function is computed for all of the sample values, the resulting cumulative probabilities have a uniform distribution on the interval 0, 1.

So if your sample is 10 random values, and the population cumulative distribution function is F(x), you would apply F(x) to each of the ten values getting a value between 0 and 1. The distribution of those ten values from the CDF is the uniform distribution.

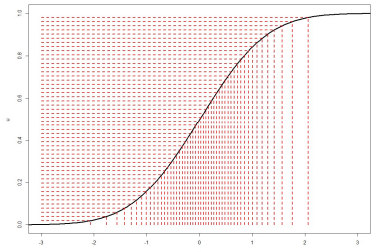
This is the opposite principle from inverse transform sampling – instead of starting with values and going to exceedance probabilities (uniform-distributed) via the CDF, we go from exceedance probabilities (uniform-distributed) to values via the *inverse* CDF.

Order Statistics of the Uniform Distribution

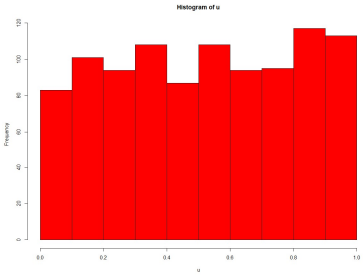
Proof:



1,000 samples from a standard normal distribution



CDF



Histogram of cumulative probabilities

11

This example shows how we go from sample values of the population to the cumulative probabilities.

First we take a large number of samples from a population. In this example we assume the population has a standard normal distribution, and we draw 1,000 samples from it. The histogram of these samples is shown on the left and it makes the familiar bell-curve shape.

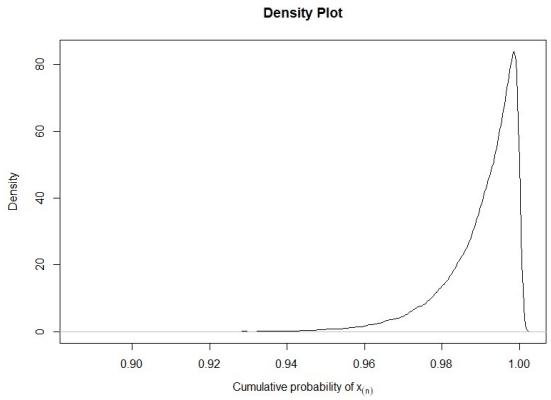
Next, we take each of those 1,000 values and we evaluate the CDF at each value. Using the CDF is a deterministic operation (not random, we are not making any new samples.) The CDF is shown in the middle as the black curve. This is the CDF for the standard normal distribution, which we have assumed is the distribution of the population. Evaluating the CDF for each point involves going from the values on the x-axis, up to the curve, and reading off the corresponding y-value. In this plot, you can see the red lines are dense in the middle and more sparse in the tails when you look along the x-axis. This is how data that are normally-distributed typically behave. Once you convert them into CDF values by reading off the y-value for each data point on the curve, you can see that the shape of the data change. They are much more regularly spread out along the y-axis. We see the histogram of the CDF values on the

right side, which have a little bit of noise, but definitely don't look bell curve-shaped anymore. With a large enough sample this histogram would have the same height in every bar, indicating a uniform distribution.

Order Statistics of the Uniform Distribution

Why care?
The same order statistic in repeated samples from the same population will not always have the same exceedance probability.

Let's examine the distribution of $F(x_{(n)})$.
- Generate a sample from the standard normal distribution of size 100
- Find the cumulative probability for sample maximum $x_{(100)}$ (= $x_{(n)}$) using CDF
- Repeat a large number of times
- Plot the empirical density



12

So why do we care about this?

When we look at the behavior of the same order statistic, for example the sample maximum (which is labeled x-sub-n in parentheses) across repeated samples from the population, it will not have the same exceedance probability. The sample maximum varies with each sample, and by taking the CDF of each sample maximum and plotting a density for it, we can see how the cumulative probability varies for the sample maximum value.

In this demo, we generate a very large number of samples of size 100 from a standard normal distribution. In each of those samples we take the sample maximum, and compute its cumulative probability. Then, we plot this huge number of exceedance probabilities on an empirical density.

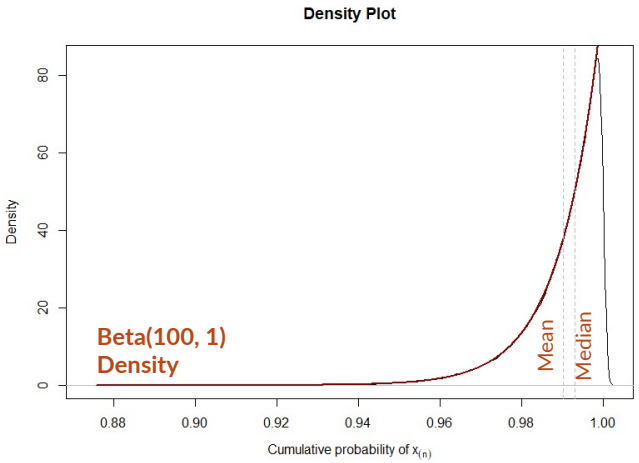
Ordinarily, we would expect the largest value in a sample of 100 to be exceeded on average 1% of the time, for a cumulative probability of 99%. In the density on the right, we can see the peak of the density around 0.99 as we would expect. However, we see that there is variation in this value, including a negative skewness and a long tail off to the left.

Order Statistics of the Uniform Distribution

What does this show?
The sample maximum has an uncertain exceedance probability.

In fact, the exceedance probability has a probability distribution, the *beta distribution*, with parameters computed from the rank and the sample size.

$$F(x_{(i)}) \sim Beta(i, n+1-i)$$



13

So what does this show?

First, it shows that the sample maximum has an uncertain exceedance probability. We can't be sure exactly what the magnitude of the sample maximum, or its corresponding cumulative probability, will be, in any sample. We see this variability in the density plot.

Also shown on the graph is the mean and the median of the sampled values. We see the effect of the long left tail and negative skewness – the mean is to the left of the median.

Despite the uncertainty, what we do see is that the uncertainty in the exceedance probability has stable behavior. The exceedance probability of the sample maximum has a beta distribution with parameters 100 and 1, and it is shown in red on the plot on the right drawn over the empirical density from our simulation experiment. It turns out that the beta distribution can be used for the exceedance probability of any order statistic, by changing its parameters. If you use the rank as the first parameter (where 1 is the sample minimum and n is the sample maximum), and the formula n + 1 minus the rank for the second parameter, you can develop the probability

distribution for the cumulative probability of any order statistic.

Order Statistics of the Uniform Distribution

What is the consequence?
We just derived two very important *plotting position estimators*:

The median plotting position

$$F(x|i,n) = \frac{i - 0.3175}{n + 0.365}$$

approx for median of beta dist

The Weibull (mean) plotting position

$$F(x|i,n) = \frac{i}{n+1}$$

exact mean of beta dist

$F(x_{(n)})$, n = 100, 100,000 samples

Property	Sample	Exact
Median	0.9931	0.9932
Mean	0.9901	0.9901

14

We have talked about plotting positions before, but the exercise we just went through showed us exactly how these plotting position estimators are derived. Remember that plotting positions are estimators for the cumulative probability of a value in a sample based only on its rank. In other words, plotting positions are functions of order statistics. Now that we have a model for the cumulative probability of each order statistic, we can see how we get to the two most often used plotting positions in hydrology: the median and Weibull plotting positions.

The median plotting position provides the median estimate of the cumulative probability for an order statistic. The plotting position estimator is shown here, based on the rank of the data, i, and the sample size, n. The values here are an approximation for the median of a beta distribution.

The Weibull, or mean, plotting position is shown below. This is the exact mean of the beta distribution, using the rank and the sample size. This formula is probably familiar to you if you have used plotting positions before, as it is one of the most popular to use.

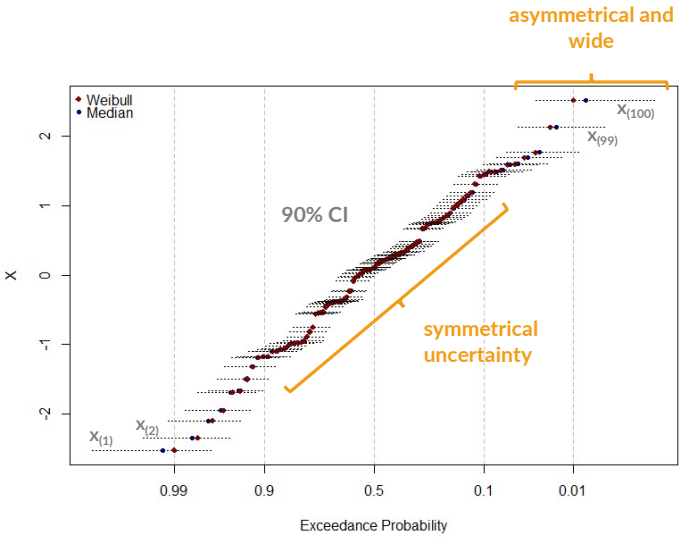
In the table in the upper right we see how these values compare based on our

sample experiment from before. We used 100,000 samples of size 100, and looked at the cumulative probability for the sample maximum. The first row compares the sample median for the cumulative probability with the value computed from the plotting position estimator. In this case they are equal out to the third decimal place, with some Monte Carlo error involved. For the mean in the second row, we see that the values are equal. It's great then theory and reality line up!

Order Statistics of the Uniform Distribution

Other consequences?
We can compute confidence limits for empirical exceedance probability and show just how uncertain it is.

- Difference in plotting position choice matters most at tails
- Uncertainty in empirical frequency large at tails



15

The consequence of this finding is that we can express the uncertainty in the true cumulative probability for each order statistic in a sample. When we plot data using an empirical quantile plot, with the exceedance probability or cumulative probability on the x-axis and the data value on the y-axis, with the estimate for the x-value coming from a plotting position estimator, we are faced with a choice: which estimator do we use? Let's consider this sample of 100 points from a standard normal distribution.

We have the median plotting position estimator in blue and the mean estimator in red. Each point is also labeled with a dotted black line, which has the 5% and 95% uncertainty range for cumulative probability using the beta distribution.

The first thing we see is that the estimators are not very different in the middle of the data. This is because the uncertainty is fairly symmetrical, and the mean and median are likely to be close to each other. *In the middle of the data, choice of plotting position estimator does not matter much.*

However, once we get out to the tails, we see that the uncertainty is asymmetrical and very wide. We also see that the plotting position estimators are quite different,

especially for the largest or smallest two or three values.

It is important to keep in mind that plotting positions are uncertain, especially in the tails, and that comparing two samples to each other, or a sample to a distribution, when the sample is plotted using plotting position estimators, has this uncertainty as well.

Relation to Extreme Value Theory

- This procedure explores the uncertainty in exceedance probability for a sample
  - We can model this with the beta distribution
- What if I want to get the distribution of the values for a particular order statistic?
  - Depends on *i*, *n*, and also f(x)
  - Straightforward for $x_{(1)}$ and $x_{(n)}$

16

The reason we cover this topic in extreme value theory is because we are concerned with the predictability of extremes, typically the sample maximum. We show here that the exceedance probability of the sample maximum has a regular behavior that we can model with the beta distribution. It turns out that this result is useful for more than just the sample maximum, which is a happy by-product.

Now we have to pivot to the opposite question. Instead of estimating the cumulative probability of an order statistic, what does the distribution of the **values** of an order statistic look like? This depends on its rank, the sample size, and also the population distribution f(x). Fortunately, we will find that this is fairly straightforward for the sample minimum and sample maximum, due to extreme value theory.

Summary

- Order statistics look at data based on their rank
- The true exceedance probability for an order statistic is uncertain
- Plotting position uncertainty can be modeled with the beta distribution

17

In this video, we discussed order statistics and their role in extreme value analysis. Three key points to take away from this video are:

- Order statistics look at data in a dataset based on their rank in that dataset
- The true exceedance probability for an order statistic is uncertain
- When we use plotting positions to represent the empirical distribution of our data, the uncertainty in those plotting positions can be modeled with the beta distribution.

Thanks, and tune into the next video in this series on extreme value theory to learn about the first extreme value theorem.

Extreme Value Theory

Part III: First Extreme Value Theorem

Gregory S. Karlovits, P.E., PH, CFM
US Army Corps of Engineers
Hydrologic Engineering Center

US Army Corps of Engineers

Hello everyone, I'm Greg Karlovits from the Hydrologic Engineering Center. Welcome to our course on statistical methods in hydrology. This video is part three of four on the topic of extreme value theory and will discuss the first extreme value theorem. Let's get started.

How do the extremes vary?

- Usually we are most interested in $f(x_{(n)})$
- Repeated samples, n = 10 from a population:

Sample Value	Sample Number 1	2	3	4	5	6	7	8	9	10	11	12
1	48	25	40	74	26	26	76	64	74	80	95	76
2	92	48	44	80	24	94	79	32	13	48	90	31
3	4	26	6	17	62	46	23	68	44	9	24	70
4	76	13	88	47	74	99	56	2	88	29	15	93
5	63	50	92	33	6	67	44	87	91	8	26	8
6	35	81	14	62	84	81	9	86	26	35	87	94
7	27	92	63	82	79	7	72	72	98	4	33	72
8	87	96	66	21	15	71	61	37	99	74	61	65
9	12	77	37	10	33	43	45	57	15	96	66	5
10	80	85	13	32	44	35	31	65	15	31	47	39
Max	92	96	92	82	84	99	79	87	99	96	95	94

distribution of these →

19

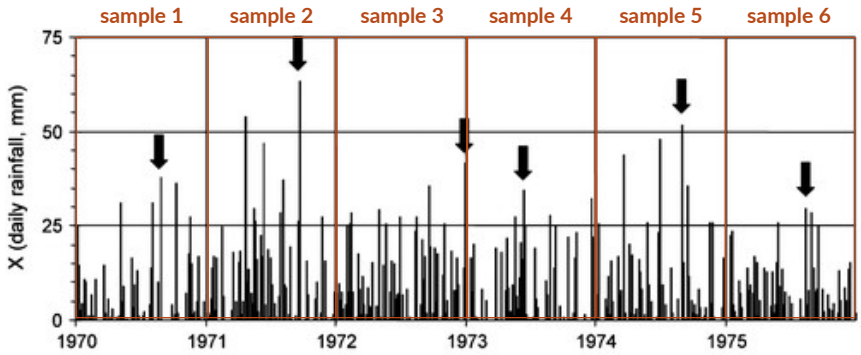
Previously, we dealt with the variability of the cumulative probability for the sample maximum. Here we will look at the variability in the value of the sample maximum.

This table shows 10 samples, each of size 10, from some population. Each column is a sample. We see the sample maximum highlighted in red for each column. Then, we record the sample maximum value in the last row. What we are ultimately concerned with is the distribution of the values in that last row.

I did say previously that this is straightforward for the sample minimum as well. If you transform these data by making each value negative and still using the sample maximum, the results still work. For example in column 1, if you make each value negative, the sample maximum is -4. The sample minimum is 4.

Block Maxima

- Non-overlapping groups
- Equal size



20

When we use non-overlapping groups of equal size to create our sample maxima, the procedure is called **block maxima.** In hydrology, we typically use the water year as our equal-sized non-overlapping groups. In the time series shown here, you can see that there are 6 water years corresponding to the 6 samples we are taking. The arrow labels the block maximum, which we also call the **annual maximum**. When the blocks we use for this kind of analysis are a year, we refer to the collection of the maxima we collect as an **annual maximum series**. You have seen this several times before, but this just reinforces exactly why the annual maximum series is used.

Modelling Extremes

- Some assumptions:
  - All of the values come from the same population, with possibly unknown density function f(x; Θ)
    - f(x; Θ) would be the distribution for *every day* of flow
  - All of the values are taken independently
    - Using block maxima with big enough blocks helps ensure this
    - Motivation for the "water year"
- We can estimate a model for $f(x_{(n)})$

21

When we go to build a model for these block maxima, we have to make some assumptions.

First we assume that all of the maxima are drawn from the same population with some distribution f(x). For annual maximum streamflow, f(x) is the distribution of every day of flow.

Second we assume that all of the maxima are taken independently of each other. This is why using blocks is important. If the blocks are large enough, the events are spaced out so that we can assume they are independent of each other. This is why water years are divided at the driest part of the year, typically the fall in North America. It maximizes the separation in time between subsequent floods.

When we meet these assumptions, we can estimate a model for the values of the sample maximum.

Fisher-Tippett-Gnedenko Theorem

- The distribution of the maximum of repeated samples of a homogeneous population converge to one of three probability distributions:
  - *EV1: Gumbel Distribution*
  - *EV2: Fréchet Distribution*
  - *EV3: Weibull Distribution*
- All three distributions can be represented with a single distribution:
  - **Generalized Extreme Value Distribution**

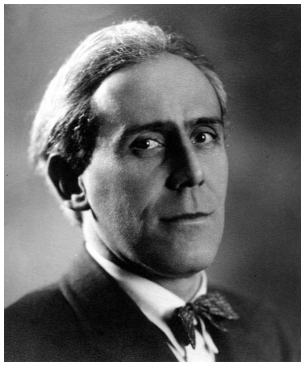
22

The Fisher-Tippett-Gnededenko theorem gives us a model for these block maxima.

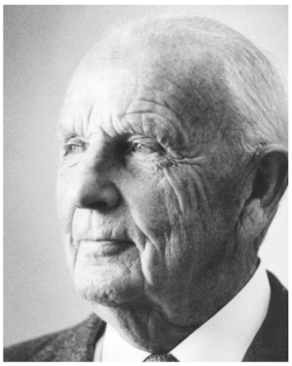
The theorem states that when we meet the assumptions previously, the values of the sample maxima will have one of three distributions: the Gumbel distribution, the Fréchet distribution, or the Weibull distribution. These are also referred to as the extreme value type I, type II, and type III distributions.

Fortunately you don't have to remember or guess which of these three distributions your maxima will converge to. If you use the **generalized extreme value distribution**, it can represent all three of those extreme value distributions with a single function.

Namesake


Emil J. Gumbel


Maurice R. Fréchet


Waloddi Weibull

23

I like to briefly explain where the names for these distributions come from. These three men were contemporaries, with most of their work coming in the first half of the 20th century. The distributions were not named for them until later.

On the left is Gumbel, who we have seen previously, was a German statistician and political scientist who was staunchly anti-Fascist, was driven from Germany in the years leading up to WWII.
In the middle is Fréchet, who was a Frenchman who seemed to contribute to every field of mathematics and is very well known.
And on the right is Weibull, a Swedish engineer who contributed greatly to the fields of reliability engineering and offshore oil exploration.

Central Limit Theorem

- Think of this as the central limit theorem (CLT) except for maxima:
  - CLT states that the sample average of repeated draws of size *n* from a population converges to a normal distribution

$$S_n = \frac{X_1 + \cdots + X_n}{n} \qquad c_n^{-1}(S_n - d_n) \xrightarrow{d} \Phi$$

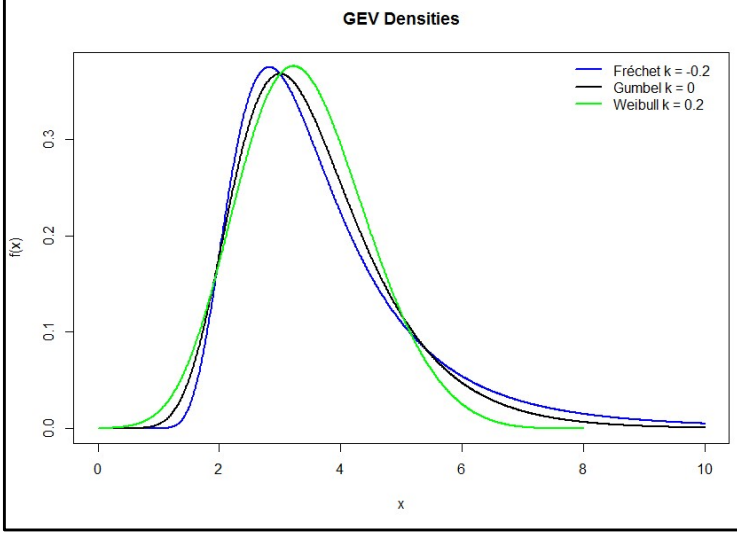
  - First EV Theorem states that the maximum of repeated draws of size *n* from a population converges to a GEV distribution

24

This result is very similar to a more familiar theorem. The central limit theorem says that if we take repeated samples from a population and compute their mean, that the sample means converge to a normal distribution. The Fisher-Tippett-Gnedenko theorem, also called the First Extreme Value Theorem, replaces the sample mean with the sample maximum, and the normal distribution with the generalized extreme value distribution.

d_n is the population mean, and c_n = the population standard deviation divided by sqrt(n) for CLT

GEV Distribution



$$F(x;\xi,\alpha,\kappa) = e^{-e^{-y}}$$

$$y = \begin{cases} -\kappa^{-1} log\left\{1 - \dfrac{\kappa(x-\xi)}{\alpha}\right\} & \kappa \neq 0 \\ \dfrac{x-\xi}{\alpha} & \kappa = 0 \end{cases}$$

25

The generalized extreme value, or GEV, distribution, changes its shape according to its third parameter, kappa. This shape parameter allows it to become each of the three types of extreme value distribution. When it is exactly equal to zero, the GEV distribution becomes the Gumbel distribution. When it is negative, it becomes the Fréchet distribution. The Weibull distribution results from a positive kappa. Sometimes in literature you will see the opposite convention with a slightly different density function. The convention shown here instead is the convention used most frequently in precipitation-frequency analysis.

Convergence

- EV convergence depends on three things:
  - The distribution of the parent population f(x; θ)
    - Changes to which EV distribution samples converge
  - The number of events per block *n*
  - The number of blocks forming the estimate
    - How good is the estimate of the GEV parameters?

$$\lim_{n\to\infty} f\left(x_{(n)}\right) = \text{GEV Distribution}$$

26

Our sample maxima converge to the extreme value distributions at varying rates. How quickly and how close our samples get to the GEV distribution depend on three things:

First, it depends on what the entire population that we get our maximum from looks like. Different parent population distributions result in the sample maxima converging to different EV distributions (that is, Gumbel, Fréchet, or Weibull). Some of these parent populations converge more slowly than others.

Second, the rate depends on how many events are in each block. Think of this as the number of "chances" we get to draw our maximum value. In a water year in a dry climate, there may only be a couple of significant flows during the year, so the number of events per block is small. In a humid climate, there may be many local maxima to draw our block maximum from. This is what is meant by events per block.

Finally, our convergence depends on our sample size. If we only have a short record of block maxima, then the samples may not seem to have an extreme-value distribution.

However, the most important factor in convergence is that second point: how many events per year our maximum is drawn from. As this value gets larger and larger, the distribution of the values of our maxima look more and more like the GEV distribution.

Convergence: Maximum Domain of Attraction

- EV1: Gumbel (GEV $\kappa = 0$)
  - $f(x; \Theta)$ in the exponential family
- EV2: Fréchet Distribution (GEV $\kappa < 0$)
  - $f(x; \Theta)$ is heavy-tailed
- EV3: Weibull Distribution (GEV $\kappa > 0$)
  - $f(x; \Theta)$ is light-tailed or upper-bounded

27

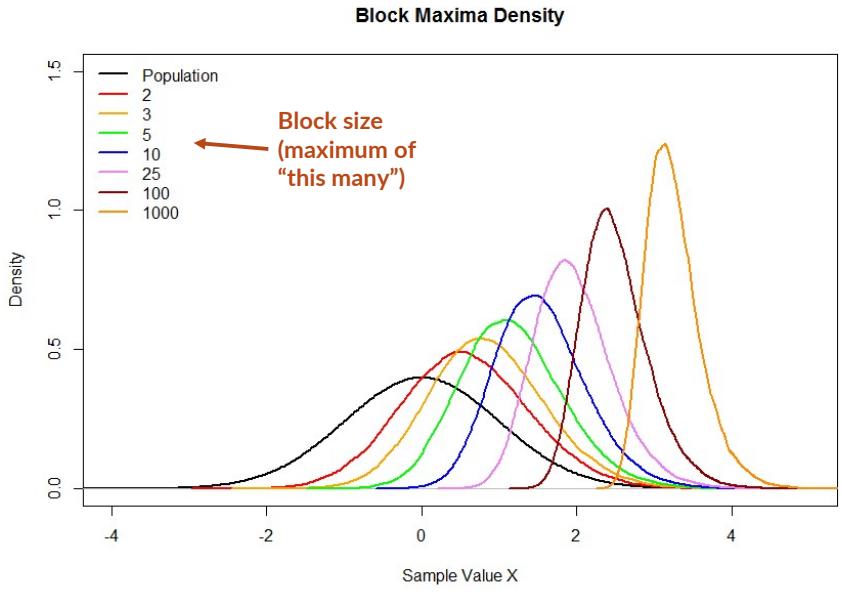
An important topic to consider is what is called the maximum domain of attraction. This describes the relationship between the distribution of our parent population, and the distribution of block maxima taken from it. Looking at the parent population first can indicate how well your maxima are converging towards the GEV distribution.

When the parent population is in the exponential family, for example the normal, exponential, and gamma distributions, the maxima are in the Gumbel MDA, and tend to have a GEV shape parameter very close to zero. Despite the modest tail weight of exponential family distributions the convergence of maxima in the Gumbel MDA can be very slow.

Heavy tailed parent populations tend to fall into the Fréchet MDA. We often find this is the case in precipitation-frequency analyses.

Light-tailed or upper bounded populations tend to fall into the Weibull MDA. These kinds of parent populations tend to converge to the GEV distribution the fastest.

Convergence of Block Maxima



This figure shows the effect of the number of events per block. The parent population, a standard normal distribution, is shown in black.

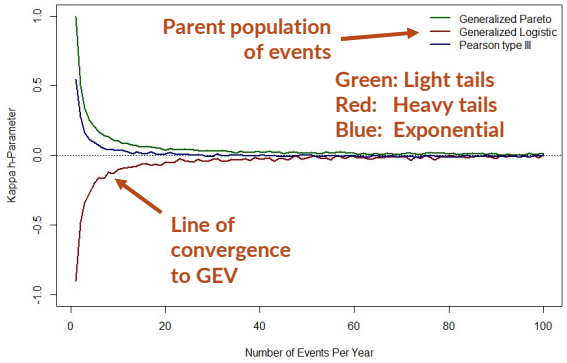
If I take repeated samples of size two from this parent, and keep the larger one, the distribution of that larger value looks like the red curve just to the right of the black one.

If I repeat this exercise for the largest of a sample of 3, 5, 10, and so on, the curves move to the right and get progressively more peaked, culminating in the maximum of 1000 samples in the orange curve to the right.

This shows two interesting things: First, that the average value of the maxima grows with the number of events per year. This makes sense: the more chances you have to see a big value, the more likely that you'll get one. Second, we see that the variability actually decreases – the peak of the density is much taller as the number of events per block increases.

Convergence: Why isn't this AMS GEV?

- Three primary factors delaying convergence:
  - Too few independent events per block
  - Too few blocks (years) creating AMS
  - Inhomogeneous parent population



$$\lim_{n\to\infty} f\left(x_{(n)}\right) = \text{GEV Distribution}$$

When *n* is small,

$$f\left(x_{(n)}\right) \sim \text{Kappa Distribution}$$

29

Sometimes when you gather annual maximum data, the sample you get doesn't look like it has a GEV distribution. There are three primary factors that can delay the onset of convergence. First is that there are too few independent events per block. From the previous slide we saw that the maximum of 2 or 3 values looks a lot like the parent population still. It takes more events for it to start looking like the GEV distribution. Second, when we have limited sample sizes, there is substantial uncertainty simply due to sample error. Third and most often, it is that the parent population is actually a mixture of several kinds of processes, so our samples are inhomogeneous.

Another way to think about it is this - imagine you live in a place where it rains once a year. If you take the maximum of that, nothing changes. The maximum of one value is the same as the population. If it rains twice, three times, etc. then when you take the maxima the resulting distribution looks more and more GEV-like.

In the situations where you are modeling maxima in this way, and you find that the number of events per block that you have means that your data haven't quite converged to the GEV distribution, consider using the kappa distribution instead. In a technical paper I co-authored, we determined that the kappa distribution should be

used to model these situations when an extreme value analysis is appropriate, but the number of events per block is too small.

Convergence – Annual Maximum Streamflow

- Although we take the maximum of 365 days of flows, what we care about most are floods
- Few real floods happen each year at most sites
  - Some years don't have any events we would consider "floods"
- Streamflow records are mixed and serially correlated
- Effectively we are taking the maxima of far fewer than 365 events
- Bottom line: Bulletin 17 procedures do not generally meet the assumptions required of EVT analysis

30

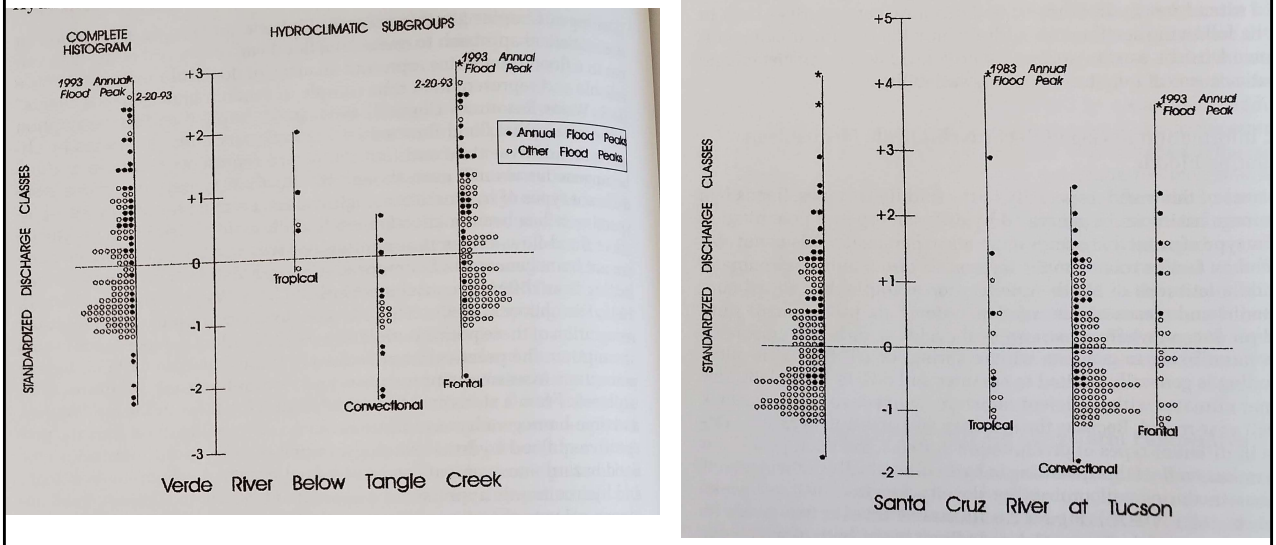
In hydrology we work with annual maximum streamflow all the time. The annual maximum is the largest instantaneous flow out of 365 days of flow observations. However, not every day is a flood event! When we are interested in the largest flow in the year it is because we want to know what big floods look like. This means that there may only be a few to no events that we think of as floods for a given site!

Analyzing streamflow also suffers from the fact that the records are mixed, with flows caused by a variety of mechanisms, and they are serially correlated, meaning observations are related in time. This makes it hard to get independent and identically-distributed samples.

The bottom line is that Bulletin 17 procedures are built to deal with some of these issues in streamflow, because the data do not always meet the assumptions we need for extreme value analysis.

Log-Pearson type III can be an excellent model for daily streamflow. The fact that we are taking the annual maximum of only a handful of true "events" could be the reason that LP3 tends to be a good model for annual maximum streamflow although theory suggests that the result should be GEV instead.

The biggest challenge is that the mechanisms that create floods are a wide spectrum – floods can depend on a long memory of highly variable meteorology. This indicates that the parent population of all floods is mixed, and any sample taken from it is inhomogeneous. This happens in almost any watershed, although rarely there can be one clear and dominant flood-causing mechanism.



Inland Flood Hazards: Human, Riparian, and Aquatic Communities (Wohl, 2011)

Some studies have been done to show how varying kinds of meteorology create floods, so that we can look at more homogeneous samples. This isn't typical practice in streamflow frequency analysis, however.

These two plots show two different gages. On the very left is a histogram of all floods for the gage, with larger values towards the top. Filled in circles are annual maximum events and open circles are other flood peaks. You can see an obvious positive skew to these data. Notable floods are shown as stars in each plot.

For both of these systems, three kinds of storm event are shown: tropical storms, convectional storms, and frontal storms. The data show two things: how frequently those types of storm occur, and how large the resulting floods tend to be. On the left, tropical floods are infrequent but severe. Frontal storms dominate the flood record, generating most of the floods at this site, including the 1993 flood of record. Convectional floods created a number of annual maxima, but are less common and less severe than the other types. The annual maxima are obviously mixed.

On the right there is a slightly different story. Tropical storms are a little more common but still uncommon. The 1983 flood of record was a tropical storm. In this

system convectional systems are more common, but most are not annual maxima. Frontal storms are less common than convectional storms but the second largest peak on record in 1993 is a frontal storm.

You can see how treating the entire histogram as a homogeneous sample could be a mistake if you wanted to use an extreme value approach to this data – there is clearly a combination of mechanisms here.

Convergence – Precipitation

- Conversely, it is much easier to isolate independent rainfall events of the same causal mechanism
  - Example: easy to identify which rainfall events in Florida are caused by tropical storms
  - Eliminates mixtures
  - Some types of storms occur several times per year at some locations
- Rainfall is much easier to analyze in traditional EVT manner
- Plus, regionalization is much easier than for streamflow

32

In precipitation-frequency analysis, it is becoming more common practice to try to isolate the mechanisms that create precipitation events. It still requires expertise in order to perform a storm type separation, but can be simpler and more straightforward than in the streamflow case.

Rainfall is much easier to analyze in a traditional extreme value theory manner, which is why precipitation-frequency analysis tends to rely on it more than streamflow frequency. Plus, as we will see in a coming topic, regionalization is easier for precipitation than for streamflow.

Summary

- The first extreme value theorem provides a model for the magnitude of block maxima
- Annual maximum series tend to converge to the generalized extreme value distribution
- Several issues can prevent sample convergence to GEV

33

In this video, we discussed the first extreme value theorem. Three key points to take away from this video are:

- The first extreme value theorem provides a model for the frequency-magnitude relationship of block maxima
- Annual maximum series tend to converge to the generalized extreme value, or GEV, distribution
- Several issues with the data may prevent that sample's convergence to the GEV distribution.

Thanks, and tune into the next video in this series on extreme value theory to learn about the second extreme value theorem.

Extreme Value Theory

Part IV: Second Extreme Value Theorem

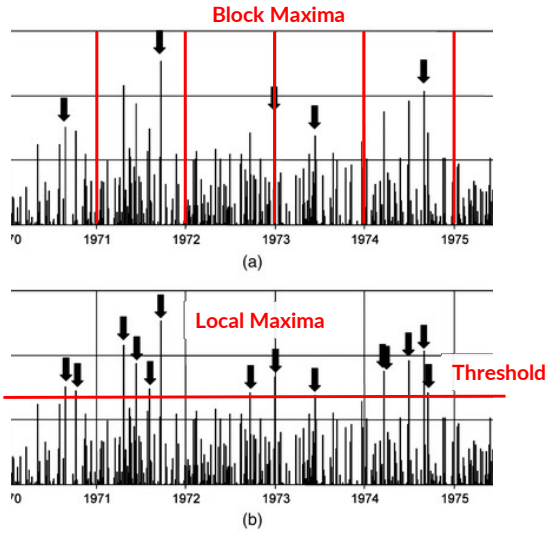
Gregory S. Karlovits, P.E., PH, CFM
US Army Corps of Engineers
Hydrologic Engineering Center

US Army Corps of Engineers

Hello everyone, I'm Greg Karlovits from the Hydrologic Engineering Center. Welcome to our course on statistical methods in hydrology. This video is part four of four on the topic of extreme value theorem and will discuss the second extreme value theorem. Let's get started.

Peaks Over Threshold

- Block maximum approach "throws out" data
- Some blocks have small maxima
  - Smaller than non-maxima in other blocks
- What if we consider independent local maxima?



35

Previously we considered order statistics, more specifically the sample maximum from non-overlapping blocks. One issue with doing this is that we are often "throwing out" data! In each block there are sometimes events that are notable but are not the block maximum. Sometimes these secondary events are larger than some block maxima. However, when we use the first extreme value theorem, we don't consider these events.

Instead, what happens if we throw away the notion of blocks, and only look at all of the large events in our record, where large is defined as exceeding some threshold that we define. We would have to ensure that each of these events is independent, which is a little more work than when we use block maxima, which we assume are independent by design. When we collect all of the events in our record exceeding a threshold, where there may be anywhere from zero to several per year, we call this approach the "peaks over threshold" approach.

Peaks Over Threshold

- Zero or more local maxima per block
  - Count of peaks needs to be considered
- Need to ensure local maxima sufficiently independent
- No longer dealing with order statistics
  - Cumulative probability ≠ 1 – AEP

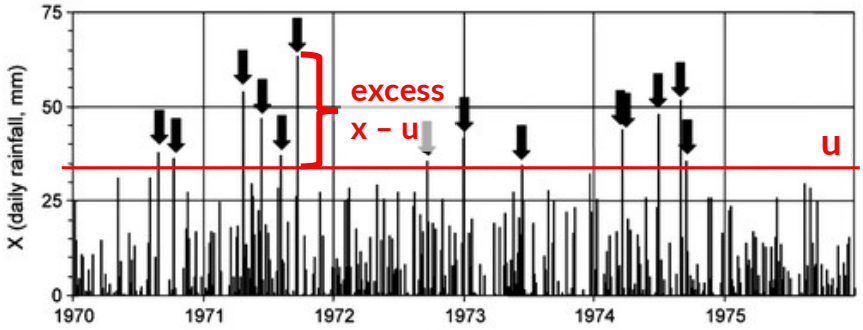
36

When we count the peaks over a threshold, we wind up with a varying number of peaks per block. These are local maxima instead of the global maximum for the block. This means that instead of having the same number of peaks as blocks, we potentially have more or less than that depending on the threshold that we pick. When getting these local maxima, it's important to make sure that they are independent. For example you wouldn't want to take both peaks of a double-peaked flood hydrograph, because they are likely to be strongly related.

One consequence of working in this realm is that we are not working with order statistics anymore, and the cumulative probabilities that we compute are not the complement to the AEP anymore. However, at the end, we will discuss how to get back to annual exceedance probabilities using a peaks over threshold model.

Peaks Over Threshold

- We are interested in the distribution of *excesses*:
  - Given a set of values that exceed a threshold,
  - What is the distribution of the excess?
  - The collection of peaks is sometimes called a ***partial duration series*** **(PDS)**



37

For all of these floods that exceed some threshold that we use to define "large events" we can build a model for how much the values exceed the threshold. The amount that the peak exceeds the threshold is called an *excess*, and it is simply the value of the peak x, minus the value of the threshold u. They will have a minimum value of zero, and most of the values are closer to zero than a large value.

In hydrology the peaks over threshold model is also sometimes called a partial duration series, or PDS.

Pickands-Balkema-de Haan Theorem

- $F_u(y) = Pr(X - u \leq y | X > u) = \frac{F(y+u)-F(u)}{1-F(u)}$
  due to conditional probability

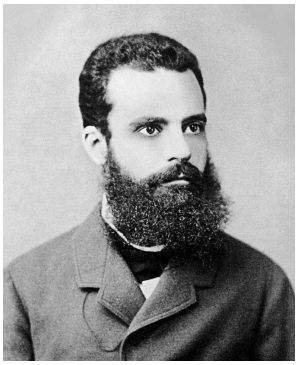
- $\lim_{u \to \infty} F_u(y) \to G(y)$
  where $G(y)$ is the ***Generalized Pareto Distribution***
  due to the Pickands-Balkema-de Haan theorem

38

The Pickands-Balkema-de Haan theorem, also called the second extreme value theorem, gives us the probability distribution for the values of those excesses.

The bottom line is that if you select a sufficiently high threshold, the values of the excesses will converge to the generalized Pareto distribution. The generalized Pareto distribution is a three-parameter distribution where one parameter is the value of the threshold selected for a PDS analysis.

Namesake



Vilfredo Pareto

39

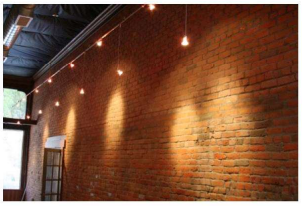
The generalized Pareto distribution is named for Vilfredo Pareto who was a civil engineer by trade, but more famously an economist who gave us the 80/20 theory of income distribution that shows up in many other applications as well.

I think he kind of looks like he might work at that trendy microbrewery downtown.

The hip new microbrewery taproom starterpack



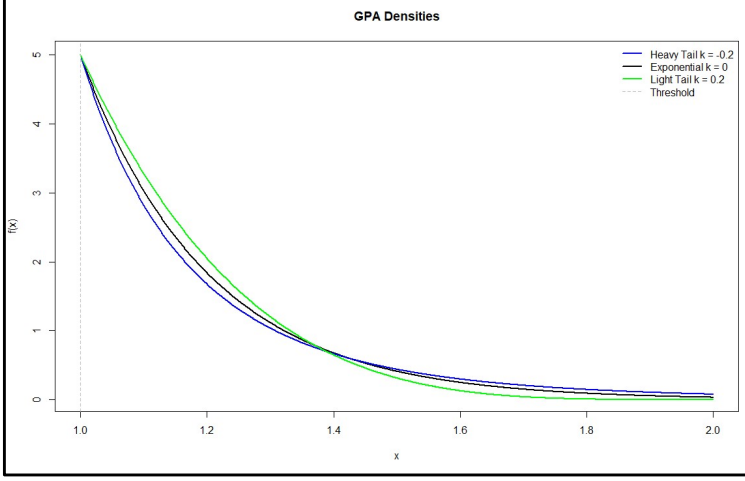





Kind of like that right? Anyway. Back to statistics.

Generalized Pareto Distribution



$$F(x;\xi,\alpha,\kappa) = 1 - e^{-y}$$

$$y = \begin{cases} -\kappa^{-1} log\left\{1 - \dfrac{\kappa(x-\xi)}{\alpha}\right\} & \kappa \neq 0 \\ \dfrac{x-\xi}{\alpha} & \kappa = 0 \end{cases}$$

41

The generalized Pareto distribution's first parameter is the threshold used to compute the excesses. It has a scale parameter, and a shape parameter as well, making it a three-parameter distribution. The shape parameter controls the tail weight of the distribution. Much like the generalized extreme value distribution, negative values of the shape parameter have a heavy tail, a shape parameter equal to zero has exponential tails, and a positive shape parameter has a light tail. Notice that all three shapes for the distribution are very similar, mainly varying in their asymptotic approach to the right, and their fixed lower support at the threshold value.

Peaks Over Threshold

- Real-life challenges:
  - Choosing a threshold
  - Ensuring peaks are independent
  - Difference between AMS and PDS results may be small

42

There are three challenges in using peaks over threshold or PDS in real life.

First is the problem of choosing a threshold. There is a tradeoff between how high the threshold is and how many sample values you get. However, a higher threshold means better agreement with the generalized Pareto distribution for the resulting excesses. This sometimes requires iterative testing of the threshold and fit to find the sweet spot. Some studies have suggested how to choose a threshold based on the resulting average number of events per year after setting it. Sometimes it is recommended to choose the threshold to be equal to the minimum annual maximum value, so that all of the data in the AMS are contained in the peaks over threshold data as well.

Second, it can be difficult to ensure that peaks are independent, especially for streamflow. For precipitation this is generally much easier, because storm events can be separated by periods of no rainfall. In streamflow it generally requires setting a minimum amount of time between subsequent events, and a value that the flow must go below between events in order for them to be considered independent.

Finally, for many studies, the difference between an AMS and PDS analysis for the

frequencies of interest may be small. This means that for some studies the extra work is not worth it. However, depending on the kinds of data you are analyzing, PDS can have benefits. One example is in a stream with annual maximum flows that are zeros; that is, the river sometimes goes dry for more than a year. This means that there are zeros in the annual maximum series. Using the partial duration series technique instead is only concerned with the situations where there are flood events.

A Fishing Example

- You are fishing in a lake over several days.
- Assuming you catch the fish at random,
  - The parent population f(x; Θ) is the length of all fish in the lake
  - You catch an average of λ fish each day
    - λ = total fish / total days
  - The largest fish you catch each day is asymptotically GEV-distributed
  - The length of all of your keepers is asymptotically generalized Pareto-distributed (GPA)

43

Let's look at a non-hydrology example of how the two extreme value theorems relate.

Imagine you are out on a lake, fishing for several days. Assume that the fish you catch from the lake are a random sample of the fish in that lake. There is a parent population f(x) that describes the distribution of the length of all fish in the lake. Over those several days of fishing, you catch some number of fish in total, resulting in an average of lambda fish per day. If you track the length of the largest fish you catch each day, the distribution of the lengths of your biggest fish is asymptotically GEV-distributed. If the department of natural resources says you can only keep fish over a certain long length, then if you record the length of all of your "keepers", the keeper length will be asymptotically generalized Pareto distributed.

A Fishing Example

- Your fish samples may not look GEV/GPA because:
  - There is a mixture of fish species in the lake
    - f(x; Θ) is not homogeneous!
  - The fish you catch probably aren't independent
  - The number of fish you catch per day (λ) is small
  - You are only fishing for a few days

44

Now what you might also find is that your fish samples aren't well modeled by the GEV or GPA distributions. This could be for a number of reasons. First, the parent population of all fish in the lake is probably a mixture: there are probably several species of fish in the lake, each with their own distribution of lengths. Second, it's likely that the fish you catch aren't independent – you probably go to the parts of the lake with the best chances to catch a fish, which probably have some species more than others. And sometimes, you catch the same fish twice! You may also only catch a few fish each day, so the convergence in distribution hasn't quite happened yet – your fish samples look more like the parent distribution than the GEV or GPA distributions. Finally, you are only fishing for a few days so your resulting samples are probably small.

Going Between AMS and PDS

- Assuming mean rate of events per year λ:
  - $F(x) = e^{-\lambda(1-g)}$

  where g is the CDF of the PDS distribution
  - This is usually called the "Langbein Adjustment"

  See Langbein (1949)
- Simulation experiment:
  - Generate 100 "years" of 50 events each from a standard normal distribution
  - Look at the distribution of the AMS
  - Look at the distribution of the PDS
  - See how the adjustment performs

45

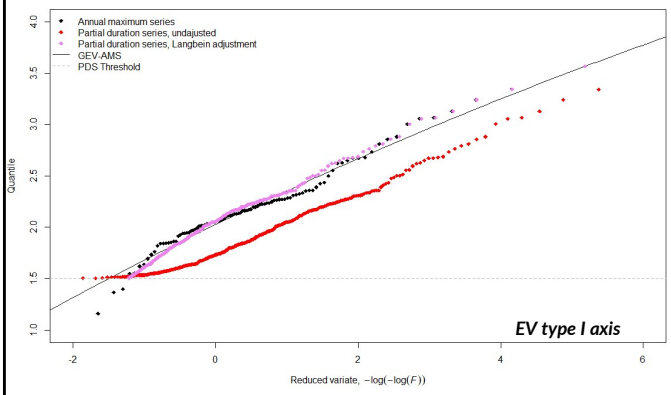
An important question is that if you use a partial duration analysis, how do you get annual exceedance probabilities, which are needed for risk-informed decision-making?

We recorded the average rate of events lambda because it can be used to convert between PDS and AMS analyses. This comes from an old paper by Walter Langbein in 1949 which still gets great usage today.

You can get the value of the cumulative probability for an annual maximum distribution, also known as the non-exceedance probability, using the cumulative probability from the PDS distribution, and the mean rate of events.

A simulation experiment will show this adjustment in action. We will generate 100 synthetic years each containing 50 events from a standard normal distribution. Then we will fit an AMS model to it using the GEV distribution, and a PDS model to it using the GPA distribution. Then, we will use the Langbein adjustment to convert the PDS result to its equivalent AMS frequencies.

Langbein Adjustment



Test equivalence of x = 3

$G(3.0) = 0.9883$	GPA fit to PDS
$\lambda = 3.38$	Sample value
$e^{-3.38*(1-0.9883)} = 0.9613$	Langbein adj.
$F^{-1}(0.9613) = 3.03$	GEV fit to AMS

46

The AMS extracted from our sample is shown in black. The PDS for the exact same sample is plotted in red. When the Langbein adjustment is applied to each of the points in the PDS, they are plotted as magenta.

The procedure showing that the results work is shown at the right. Suppose we are interested in knowing the AEP for the value x = 3, except we have only the generalized Pareto model based on the partial duration series. In this sample, the cumulative probability is 0.9883. In the PDS fit, the annual rate of events is about 3.4 events pear year. When using the Langbein adjustment, the CDF of the AMS is 0.9613. In this situation, we can check to see how well this agrees with an AMS/GEV fit to the same sample. By plugging the adjusted CDF value of 0.9613 into the inverse CDF for the AMS/GEV model, we get 3.03, which is close to 3, but slightly different due to sample error. The samples aren't big enough to perfectly estimate both the PDS/GPA and AMS/GEV models, so there is a slight discrepancy in the conversion.

Madsen et al. 1997

- $\xi^* = \xi + \alpha ln(\lambda)$ when $\kappa = 0$
- $\xi^* = \xi + \frac{\alpha}{\kappa}(1 - \lambda^{-\kappa})$ otherwise

- $\alpha^* = \alpha\lambda^{-\kappa}$

- $\kappa^* = \kappa$

At $x = \xi$, the CDF of the AMS is equal to $exp(-\lambda)$ which is the probability of no exceedances in a year

47

There is a more direct way to go between the PDS/GPA model and AMS/GEV model that bypasses the adjustment. It requires fitting the PDS model and getting the mean rate of events, and then converting the parameters from the GPA to the equivalent GEV/AMS model, as above.

Xi, alpha, and kappa are the parameters of the PDS/GPA model. Lambda is the mean rate of events for the partial duration sample. The same symbols with the star are for the AMS/GEV distribution. It is easy to convert between the two, especially because the way that these distribution functions were formulated results in the kappa parameter being equal for the two.

Summary

- The second extreme value theorem provides a model for all independent extremes above a threshold
- Partial duration series tend to converge to the generalized Pareto distribution
- Annualized estimates can be made from PDS models

48

In this video, we discussed the second extreme value theorem. Three key points to take away from this video are:

- The second extreme value theorem gives us a model for the frequency-magnitude relationship for all independent extremes exceeding a threshold
- Partial duration series tend to converge to the generalized Pareto distribution
- We are able to use a PDS model to make inferences about annual frequencies, which is important for when we need annual exceedance probabilities.

Thanks, and come back soon to learn about other topics in our statistical methods in hydrology course.