Block Maximum Techniques and River Flooding

A Sentry and Homeland Security Module

Brian Birgen, Wartburg College
Melanie Brown, Champlain College
Joyati Debnath, Winona State University

MODULE SUMMARY

The mathematical concept of Risk Assessment is fundamental in our everyday life. Almost every situation can be observed as some amount of risk involved. With an increase in extreme weather events, there is also an increased risk of property damage and loss of life. This module introduces students to Block Maximum Techniques in Statistics through the lens of river flooding. This particular application also reinforces the idea of a 100-year flood and 500-year flood to help students better understand the intersection of risk and probability in their own communities. Specific class activities are suggested to be performed that will help students become familiar with the risk assessment involved with river flooding. We also include instructions for doing the data analysis in Excel, R, and Minitab.

TARGET AUDIENCE

This module is written for any Introduction to Statistics course. The material can be covered in 1-2 class periods. In addition, it can be offered in a course that allows the instructor to include as an extension project. We suggest that this module is used after an introduction to Normal Distribution. We also note that this module could be expanded and used in some upper-level math/statistics classes. While this module introduces the Weibull Distribution, it does not explore the underlying theory and instead focuses on using technology to fit data to a particular model and then using that model for predictions.

PREREQUISITES

Students are expected to:

- Have a fundamental knowledge in Statistics.
- Understand and be familiar working with the Normal Distribution.

COMPONENTS OF THE MODULE

The module aims to introduce Block Maximum Techniques and then apply it to historical river data. We supply data for the Cedar River and provide links for students and instructors to find data on their local rivers.

Students are expected first to do the following:

a. Collect high river value for each year.
b. Fit data to Weibull Distribution.
c. Compute 100-year flood levels and 500-year flood levels.

They then repeat this problem with only recent data (last 20 years)

a. Re-calibrate the model and fit data to a new Weibull Distribution
b. Use 100- and 500-year data to determine the probability of occurrence with the new model.

ANTICIPATED NUMBER OF MEETINGS

Depending on student knowledge and course competencies, the module will require between one to two meetings. If the user is using an extension, then four meetings are recommended.

LEARNING OUTCOMES

- After completing this module, the students will be able to:
- Use the Block Maximum Technique to apply on historical data.
- Learn how to collect and use high river value for each year.
- Learn to use and fit data to Weibull Distribution.
- Solve various risk assessment related problems pertaining to flood levels.
- Develop computational and logical thinking skills, and critical understanding.

ACKNOWLEDGEMENT AND DISCLAIMER

This module was developed as a part of the ***Reconnect 2023 Workshop on Risk Assessment*** held at Omni Hotel, Providence, RI from June 18 – 23, 2022 organized by the Department of Homeland Security (**DHS**), the Center of Discrete Mathematics and Theoretical Computer Science (**DIMACS**), and **SENTRY** (**S**oft-target **E**ngineering to **N**eutralize the **T**hreat **R**ealit**Y**) and funded by National Science Foundation (**NSF**).

The authors would like to thank Dr. Midge Cozzens for her valuable suggestions and continued support. In addition, the authors want to thank the speakers Dr. Jun Zhuang, Dr. Auroop Ganguly and Mr. Jack Watson; Dr. Seth Guikema; and Dr. Allison Reilly for valuable information on Risk Assessment.

This module can be used only for educational purposes and cannot be reproduced and sold for business without the authors' permission.

I. BLOCK MAXIMUM TECHNIQUES

What is Block Maximum? Traditionally, Extreme Value Theory (EVT) uses Block Maxima (BM) techniques to estimate extreme values of an arbitrary probability values for an event [1]. The BM approach consists of dividing the observation period into **blocks**, non-overlapping periods of equal size (typically 1 year) and then obtaining a single maximum value for each period. The concept is shown in the figure below. Each block is 1 year, and the red points indicate the maximum water level within each year's block.

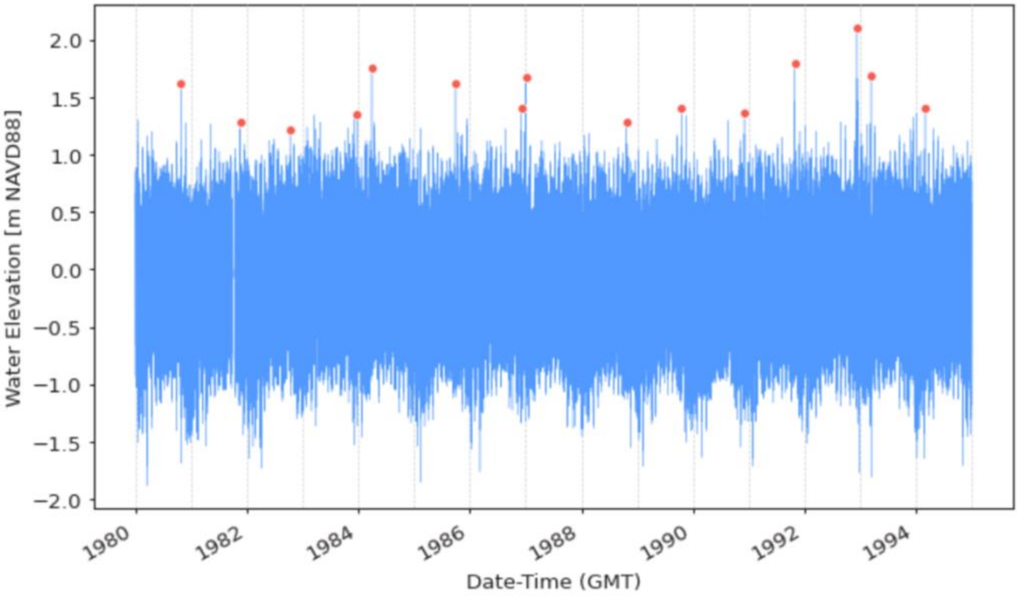


Figure 1: Illustration of BM method for water level with 1-year block size [2]

Using BM with appropriate block size can help avoid capturing seasonal data and obtain extreme behavior of the event.

Selection of BM size: Selection of block size is very important particularly in meteorological sciences with the events that are seasonal or influenced by the earth's rotation or by the climate changes like rain, snowfall, flood, earthquakes, etc. Smaller block sizes might result in poor and significant bias in the estimation and larger blocks will result in fewer blocks with larger estimation variance. Typically, 1 or 2-year blocks are used for water levels in flooding situation. In Figure 2, we see the same data as in Figure 1, but this time the block size is 2 years. There are half as many blocks, and some maxima captured in Figure 1 are now not maxima.

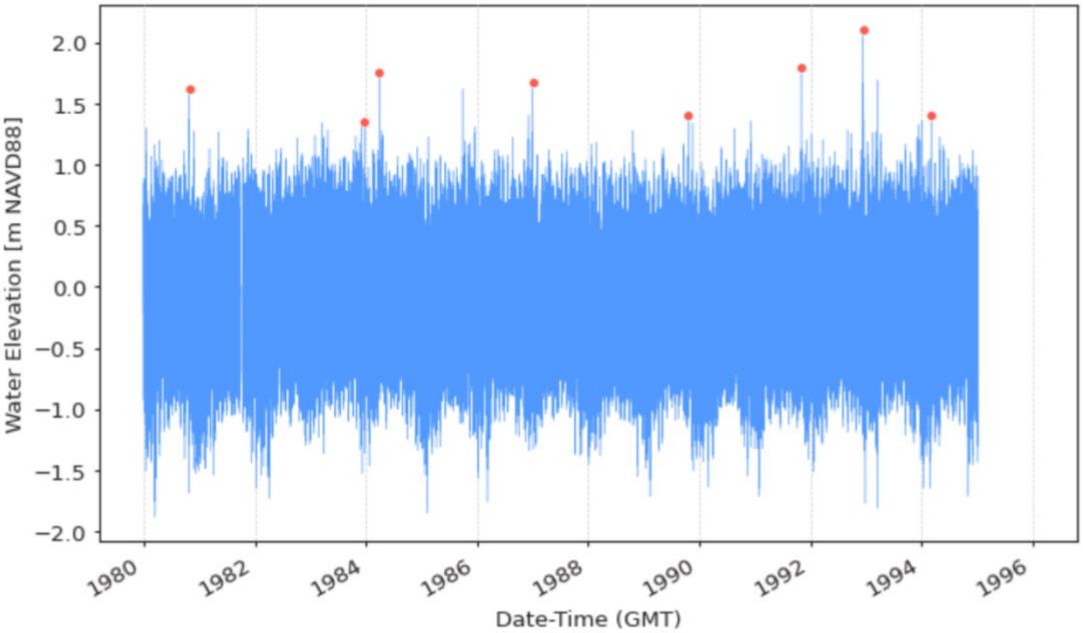


Figure 2: Illustration of BM method for water level with 2-year block size [2]

What is a 100-year flood? The term 100-year flood (or any event like storm or rain, for that matter) is used to simplify the definition of flood (or any event) that statistically has 1-percent (1/100) chance of occurring in a given year. The return period or recurrence interval is used as 100 years. This means over the course of 1 million years; these events would be expected to occur 10,000 times.

For example: Suppose the TV weatherman says, "This storm resulted in a 100-year flood on the Weibull River that crested a stage of 15 feet." This means that the probability of the Weibull River reaching a height of 15 feet is once in 100 years according to the historical data. In other words, a flood of 15 feet magnitude has a 1% chance of happening every year.

Example: The phrase "100-year flood" refers to a flood where probability of occurrence in any given year is 1%, or 1/100. Calculate the probability of a 100-flood occurring at least once within a period of 50 years.

Solution: While there may be human-related factors that influence flooding, like development on floodplains, we can generally treat the probability of flooding in different years as independent events. The probability of having no 100-year floods occur in the 50-year period is $(1 - 0.01)^{50} = 0.605$ Thus the probability of having at least one 100-year flood in the 50-year period is $1 - 0.605 = 0.395$.

What is a 500-year flood? This is just like the term 100-year flood, except that the flood has statistically 0.2 percent (1/500) chance to occur in a given year. As with the idea of a 100-year flood, it can occur multiple times at a given location within a shorter timeframe than 500 years. The flooding in New York City on Sept. 1, 2021 caused by the remnants of Hurricane Ida is classified as a 500-year flood. Ida is compared to Superstorm Sandy as that event was also a 500-year flood. That was only nine years before Ida. When Staunton, VA dealt with flash flooding in August 2020, it was classified as a 500-year flood but that also has happened multiple times such as 1985, 1996, 2002 and 2003. During the month of August, Staunton experienced 2 significant flash flood events exactly 2 weeks apart. The first occurred on Saturday August 8th, and the second occurred on Saturday August 22nd. In fact, southeast Texas had 500-year floods for 5 straight years starting in 2015.

What is Return Period or Recurrence Interval? **Return Period** or **Recurrence Interval** is another important aspect to understand. If it is incorrectly evaluated or interpreted, then it can lead to inaccurate assessment of risks. The best way to look at this is to consider a time period when a risk assessment needs to be evaluated. The Return Period or Recurrence Interval is a duration of time typically in years for flooding, corresponds to a probability that a given value will be exceeded at least once within that time period which is 1 or 2 years. If p is determined to be the return period or recurrence interval, then the probability of exceedance is $1/p$. For a 100-year flood, the recurrence interval is 100 years. For a 500-year flood, the recurrence interval is 500 years. In general, 10 or more years of data are required to perform a frequency analysis for the determination of the return period or recurrence interval.

What is River Gauge? When we talk about flooding, the data to determine flood stage is collected with a river gauge. According to American Meteorological Society, a River Gauge (also known as a Stream Gauge or a **streamgage**) is a device to measure the River Stage. Hydrologists and Environmental Scientists monitor and test bodies of water at the gauging stations. Water level surface elevations and volumetric discharge flows are generally measured. For more information look at the USGS site Gages Through the Ages: https://labs.waterdata.usgs.gov/visualizations/gages-through-the-ages/index.html#/

Note to the Instructor: The spelling "streamgage" is what the USGS uses, and they omit the "u." The authors of this module recognize the two different spellings in the previous paragraph.

II. Weibull Distribution

There are many distributions that are used to model different situations, such as Normal Distributions or t-distributions. A Normal Distribution has two parameters, μ and σ. Student's t-distribution has one parameter, the degrees of freedom. The **Weibull Distribution** is another important distribution that models situations such as the time it takes until a mechanical part fails or time between events. This distribution is named for Swedish engineer and material scientist Waloddi Weibull, who used it in fatigue analysis of building materials. The Weibull Distribution has three parameters:

- Shape α: When $\alpha < 1$, the failure rate decreases over time, usually due to many early failures. When $\alpha = 1$, the failure rate stays constant over time. When $\alpha > 1$, the failure rate increases over time, usually due an aging process that makes something more likely to fail as time elapses.
- Scale β: Also called the characteristic life, is related to the mean time to failure.
- Threshold τ: Also called the location, is the lowest possible value for the distribution and represents a translation of the data. A 2-parameter Weibull Distribution assumes $\tau = 0$.

This distribution has a probability density function in the form:

$$f(x) = \frac{\alpha}{\beta^{\alpha}}(x-\tau)^{\alpha-1}e^{-((x-\tau)/\beta)^{\alpha}} \text{ for positive values of } x.$$

The process of fitting a Weibull distribution to given data is a bit complicated and uses the cumulative distribution function:

$$F(x) = 1 - e^{-((x-\tau)/\beta)^{\alpha}} \text{ for values of } x \text{ bigger than } \tau.$$

In order to find α and β, we observe that:

$$-ln(1 - F(x)) = \left(\frac{x-\tau}{\beta}\right)^{\alpha}$$

or

$$ln\left(-ln\left(1 - F(x)\right)\right) = \alpha(ln(x-\tau) - ln(\beta))$$

We fit a line between the transformed cumulative distribution function and the natural log of the data.

$$Y = aX + b$$

which yields $\alpha = a$ and $\beta = e^{-b/a}$

III. Working with Historical Data

The Google Sheet CedarRiverAtWaterloo2021 [3] contains the height of the Cedar River in feet as measured at the USGS river gauge at midnight in Waterloo, Iowa for every day in 2021. The block size is 1 day, and we have 365 blocks. Note that the data does not capture the block maxima, but simply the gauge reading at midnight each day. Copy and paste this data into the statistical software of your choice. We provide instructions for Excel, R, and Minitab. Note that different software produces slightly different results.

1. Compute the mean and (sample) standard deviation of this data.

   Answers: Mean = 6.0226 feet. Standard deviation (s.d.) = 0.6529 feet

   Excel Commands: =AVERAGE(B2:B366)
   =STDEV.S(B2:B366)
   R Commands: > mean(CedarRiverAtWaterloo2021$Height..in.feet.)
   > sd(CedarRiverAtWaterloo2021$Height..in.feet.)
   Minitab Commands: Stat → Basic Statistics

2. Find the heights of the river which are one standard deviation above the mean and one standard deviation below the mean.

   Answers: One s.d. above = 6.6755 feet. One s.d. below = 5.3697 feet

3. For the standard Normal Distribution, what is the probability a data point is less than the value one standard deviation above the mean? What is the probability a data point is less than the mean? What is the probability a data point is less than the value one standard deviation below the mean?

   Answers: $Pr(z < 1) = 0.8413, Pr(z < 0) = 0.5000, Pr(z < -1) = 0.1587$

   Excel Commands: =NORM.S.DIST(1,1)
   =NORM.S.DIST(0,1)

```
                     =NORM.S.DIST(-1,1)
R Commands:          > pnorm(1)
                     > pnorm(0)
                     > pnorm(-1)
```

Minitab Commands: Calc → Probability Distribution → Cumulative Distribution Function. Select "Normal" with Mean = 0 and Standard Deviation = 1. Use a value of -1 for one s.d. below the mean, use a value of 0 for the mean, and use a value of 1 for on s.d. above the mean.

4. What percentage of the river height data is less than the value one standard deviation above the mean? What percentage of the river height data is less than the mean? What percentage of the river height data is less than the value one standard deviation below the mean? How do these values compare to the values for a Normal Distribution?

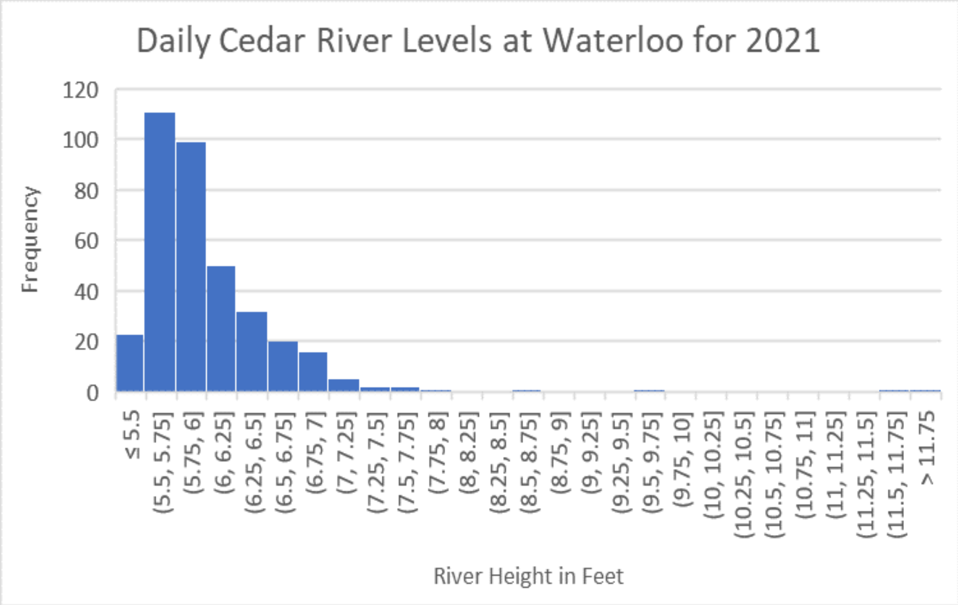
Answers: One s.d. above = 0.9014, mean = 0.6466, One s.d. below = 0.0027. These percentages are not very close to the values found in Question 3.

```
Excel Commands:      =COUNTIF(B2:B366,"<6.6755")/COUNT(B2:B366)
                     =COUNTIF(B2:B366,"<6.0226")/COUNT(B2:B366)
                     =COUNTIF(B2:B366,"<5.3697")/COUNT(B2:B366)
R Commands:          > sum(CedarRiverAtWaterloo2021$Height..in.feet.<6.6755)
                             /length(CedarRiverAtWaterloo2021$Height..in.feet.)
                     > sum(CedarRiverAtWaterloo2021$Height..in.feet.<6.0226)
                             /length(CedarRiverAtWaterloo2021$Height..in.feet.)
                     > sum(CedarRiverAtWaterloo2021$Height..in.feet.<5.3697)
                             /length(CedarRiverAtWaterloo2021$Height..in.feet.)
```

Minitab Commands: Calc → Probability Distribution → Cumulative Distribution Function. Select "Normal" and enter the mean and standard deviation from Question 1.

5. Graph a histogram of the river height data for 2021.

Answer:



Excel Commands: Under the INSERT tab, select INSERT STATISTIC CHART and then select the histogram.
R Commands: > hist(CedarRiverAtWaterloo2021$Height..in.feet.)
Minitab Commands: Graph → Histogram and select "Simple"

Note that each program will generate slightly different histograms. In particular, Minitab will generate its own recommended bins.

6. What characteristics of the data do you observe in the histogram that would indicate this data does not follow a Normal Distribution? Does this match your numerical computations?

   Answer: This data is not symmetric. It is significantly skewed to the right. This matches the computations where the proportions below various values for Normal Distributions differed from the proportions below those same values for our data.

Since our data is not symmetric and is significantly skewed to the right, the Normal Distribution is not going to model our data well. Let's try to match our data with a Weibull distribution. This means we need to use technology to find the shape, scale, and threshold parameters such that a Weibull curve is a good approximation of our data.

7. Match this data to a Weibull distribution.

   Answers: In Excel: Shape = 1.4045, Scale = 0.7175, Threshold = 5.359
   In R: Shape = 1.2366, Scale = 0.7158, Threshold = 5.359
   In Minitab: Shape = 1.23562, Scale = 0.71532, Threshold = 5.35923

   Note that these answers are close but not quite the same because the three programs use different approximation techniques.

   In Excel:

- In the first column, place the index (1 through 365)
- Sort the data smallest to largest and place in the second column.
- Because the smallest data in this set (5.36) is significantly above 0, we will adjust the data accordingly. We will subtract the value of 5.359, which is referred to as the threshold. The third column will be =B# - 5.359 where # is the row in use.
- In the fourth column place =LN(C#), the natural logarithm of the adjusted river height readings. This will be our X in the linear regression.
- To create the CDF use the formula =(A#-0.5)/365. This is F(x) and is placed in the fifth column.
- The sixth column is =LN(-LN(1 – E#)). This will be our Y in the linear regression.
- Compute the least squares regression line with the fourth column as input and the sixth column as the output.
  =SLOPE(F1:F365,D1:D365)
  =INTERCEPT(F1:F365,D1:D365)
- Then shape is equal to the SLOPE and the scale is EXP(-INTERCEPT/SLOPE)
- Shape = 1.4045, Scale = 0.7175, Threshold = 5.359

In R:

- Create the data modified by the threshold value and call it AdjustedHeight

```
> CedarRiverAtWaterloo2021$AdjustedHeight = CedarRiverAtWaterloo2021$Height..in.feet. - 5.359
```

- Use the *fitdistr* command, from the MASS package. You may need to download this library before you can use this command.

```
> fitdistr(CedarRiverAtWaterloo2021$AdjustedHeight, "weibull")
```

- Shape = 1.2366, Scale = 0.7158, Threshold = 5.359

In Minitab:

- Stat → Quality Tools → Individual Distribution Identification.
- Select the column you want to analyze. Minitab will generate the probability plots and run Goodness of Fit tests against multiple distributions. You can look through the output to see that the best fit is with a 3-parameter Weibull distribution.
- Shape = 1.23562, Scale = 0.71532, Threshold = 5.35923

8. In Question 4, we compared the Normal Distribution to the values obtained. We will repeat this for the Weibull distribution. Using your model from Question 7 and the data from Questions 1 and 2, compute the probability a data point is less than the value one standard deviation above the mean, the probability a data point is less than the mean, and the probability a data point is less than the value one standard deviation below the mean? How do these answers compare to the answers in Question 4?

Answers: In Excel: One s.d. above = 0.9042, mean = 0.5918, One s.d. below = 0.0027. Much closer.

In R: One s.d. above = 0.8805, mean = 0.5977, One s.d. below = 0.0055. Much closer.

In Minitab: One s.d. above = 0.8805, mean = 0.5979, One s.d. below = 0.0054. Much closer.

Excel Commands:

```
=WEIBULL.DIST(6.6755-5.359,1.4045,0.7175,1)
=WEIBULL.DIST(6.0226-5.359,1.4045,0.7175,1)
=WEIBULL.DIST(5.3697-5.359,1.4045,0.7175,1)
```

R Commands:

```
> pweibull(6.6755-5.359,1.2366,0.7158)
> pweibull(6.0226-5.359,1.2366,0.7158)
> pweibull(5.3697-5.359,1.2366,0.7158)
```

Minitab Commands: Calc → Probability Distribution → Cumulative Distribution Function. Select "Weibull" and enter the parameters from Question 7 in the appropriate fields. You will need to enter each value from Question 2 in the appropriate field.

The Google Sheet CedarRiverAtWaterlooPeaks contains the greatest height of the Cedar River in feet as measured at the USGS river gauge every year from 1973 until 2022. The block size is 1 year, and this data set does represent the block maxima.

9. Fit a Weibull distribution to this data.

Answers:

In Excel: Shape = 0.7455, Scale = 10.9794, Threshold = 6.189
In R: Shape = 1.4894, Scale = 7.9236, Threshold = 6.189
In Minitab: Shape = 1.79957, Scale = 8.67646, Threshold = 5.75469

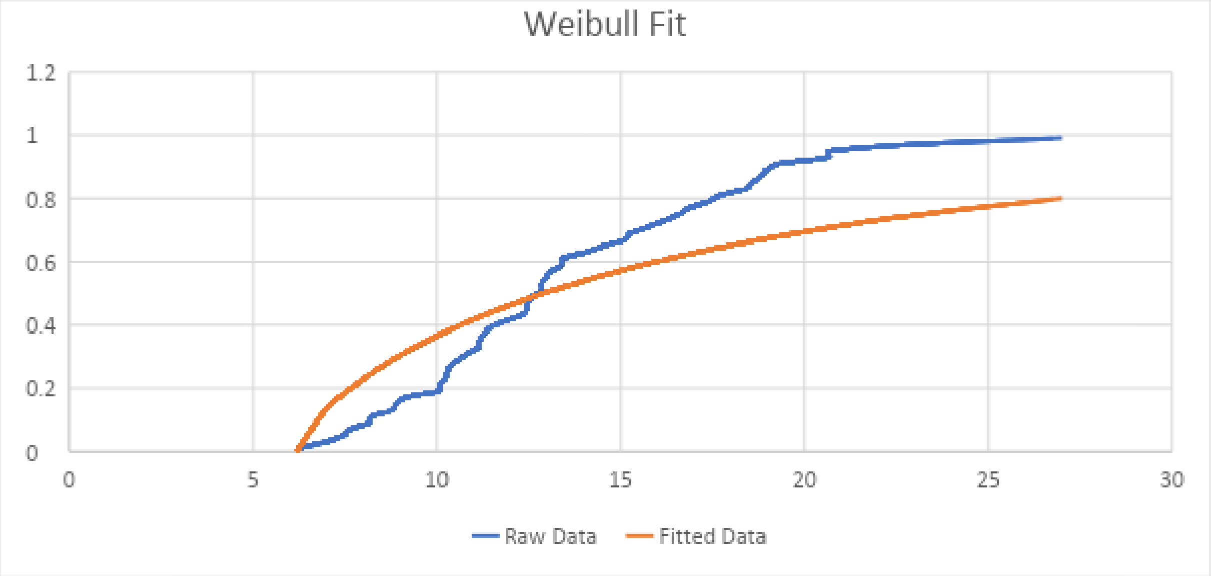


Figure 3: In Excel, comparing the data to the fitted Weibull Distribution

10. As we see in Figure 3, the Weibull Distribution we obtained in Question 9, is not a good fit for the data. This is because the high (2008) and low (1977) readings are both outliers. If we remove these data points when computing the least squares regression line, we will get a better model.

Answers:

In Excel: Shape = 1.6015, Scale = 8.2467, Threshold = 6.189
In R: Shape = 1.7817, Scale = 8.2067, Threshold = 6.189
In Minitab: Shape = 1.79401, Scale = 7.73113, Threshold = 6.46079

In Figure 4, we see that this is a much better model for our data once we trim the outliers.

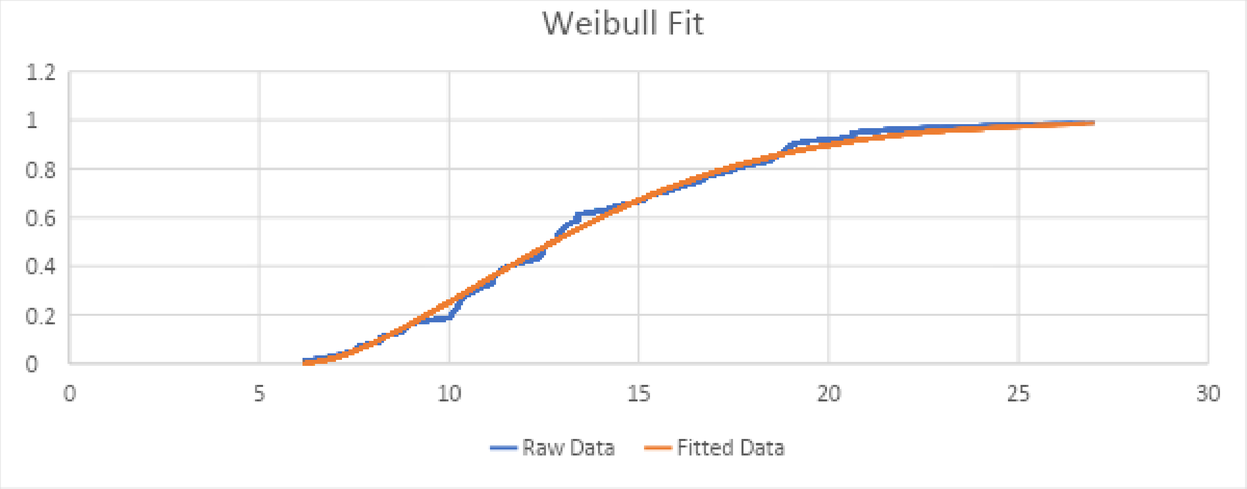


Figure 4: In Excel, comparing the trimmed data to the fitted Weibull Distribution

11. Using the answers in Question 10, compute the 100-year flood level and the 500-year flood level.

    Answers:

    In Excel:

    100-year flood level = 27.59 ft,
    500-year flood level = 31.99 ft

    In R:

    100-year flood level = 25.52 ft,
    500-year flood level = 29.07 ft

    In Minitab:

    100-year flood level = 24.57 ft,
    500-year flood level = 27.87 ft

    Excel Commands:

    ```
    =6.189 + EXP(LN(8.2467)+LN(-LN(0.01))/1.6015)
    =6.189 + EXP(LN(8.2467)+LN(-LN(0.002))/1.6015)
    ```

    R Commands:

    ```
    > qweibull(0.99,1.7817,8.2067) + 6.189
    > qweibull(0.998,1.7817,8.2067) + 6.189
    ```

    Minitab Commands:

    Calc → Probability Distribution → Inverse Cumulative Distribution Function. Select “Weibull” and enter the parameters as given.

12. The effects of climate change and residential development have caused an increase in flood levels. There is concern that the flood models have changed, and the previous models are less accurate. For that reason, adjust the model to only use the **last 20 years of data** (2003-2022), again removing the high (2008) and low (2012) readings to get a better fit of data (so you will only be using 18 observations from this 20-year period). What is the Weibull distribution obtained in this case? What are the 100- and 500-year flood levels?

Answers:

In Excel:

- Shape = 1.22877, Scale = 7.41299, Threshold = 8.209
- 100-year flood level = 33.90 ft,
- 500-year flood level = 41.00 ft

In R:

- Shape = 1.62315, Scale = 7.11512, Threshold = 8.209
- 100-year flood level = 26.44 ft,
- 500-year flood level = 30.14 ft

In Minitab:

- Shape = 1.25879, Scale = 5.93652, Threshold = 9.01523
- 100-year flood level = 28.99 ft,
- 500-year flood level = 34.36 ft

Excel Commands:

```
=8.209 + EXP(LN(7.41299)+LN(-LN(0.01))/1.22877)
=8.209 + EXP(LN(7.41299)+LN(-LN(0.002))/1.22877)
```

R Commands:

```
> qweibull(0.99,1.62315,7.11512) + 8.209
> qweibull(0.998,1.62315,7.11512) + 8.209
```

Minitab Commands:

Calc → Probability Distribution → Inverse Cumulative Distribution Function. Select “Weibull” and enter the parameters as given.

EXTENSION:

We encourage instructors and students to look up the historic data for rivers and streams in their communities and repeat these exercises. For example, the Winooski River at Essex Junction, VT during 2023 is particularly interesting since they had two 100-year floods over a 6-month period. Students can also compare the data regarding historic floods with the impacts on their communities to further reflect on the risk associated with weather events.

REFERENCES

[1] Ferreira, A and Laurens, D. H., (2015). *On the Block Maxima Method in Extreme Value Theory: PWM Estimators*. The Annals of Statistics. Vol 43(1), pp 276-298.

[2] https://georgebv.github.io/pyextremes/user-guide/2-extreme-value-types/

[3] https://waterdata.usgs.gov/monitoring-location/05464000/#parameterCode=00065&showMedian=false&startDT=2021-01-01&endDT=2021-12-31