# CHAPTER ONE
UNCERTAINTY AND RISK ANALYSIS

| INTRODUCTION | 1-1 |
| Definition of hydrology | 1-1 |
| Role of engineers | 1-1 |
| Additional definitions | 1-2 |
| PROBABILITY CONCEPTS | 1-3 |
| Need for probabilistic information | 1-3 |
| Types of hydrologic data series | 1-4 |
| What is probability? | 1-4 |
| Definitions | 1-5 |
| Probability density functions (pdfs) | 1-6 |
| RETURN PERIOD | 1-7 |
| Hypothetical illustration | 1-7 |
| Formal definition | 1-8 |
| BINOMIAL DISTRIBUTION | 1-9 |
| Introduction | 1-9 |
| Illustrative examples | 1-9 |
| Binomial distribution | 1-10 |
| Problem #1 | 1-11 |
| Problem #2 | 1-12 |
| RISK ANALYSIS | 1-12 |
| Background | 1-12 |
| Computational formulae | 1-12 |
| Problem #1 | 1-13 |
| Problem #2 | 1-13 |
| FREQUENCY ANALYSIS USING PLOTTING POSITION | 1-14 |
| Example data set | 1-14 |
| Exceedance probability | 1-16 |
| Graphical solution | 1-17 |
| Summary of computational steps | 1-19 |
| Disadvantages | 1-19 |
| FREQUENCY ANALYSIS USING ANALYTICAL CURVES | 1-20 |
| Introduction | 1-20 |
| Statistical characteristics | 1-20 |
| Probability density functions | 1-22 |
| EXTREME VALUE TYPE I DISTRIBUTION | 1-23 |
| Background | 1-23 |
| Probability density function - pdf | 1-24 |
| Probability for a given magnitude | 1-24 |
| Magnitude for a given return period | 1-26 |
| EXAMPLE PROBLEM | 1-26 |
| Solution with the plotting position method | 1-28 |
| Solution with the extreme value type I distribution | 1-29 |
| Uncertainty/Risk Problem Assignment | 1-32 |
| APPENDIX 1-A: RELATED MATERIALS | 1-36 |
| Average recurrence interval | 1-36 |
| Generalized extreme value distribution | 1-36 |
CHAPTER ONE
UNCERTAINTY AND RISK ANALYSIS

Prediction is hard ... especially about the future.

Niels Bohr

INTRODUCTION

Definition of hydrology

Federal Council for Science and Technology, Ad Hoc Panel (1962)

"Hydrology is the science that treats of the waters of the earth, their occurrence, circulation and distribution, their chemical and physical properties and their reaction with their environment, including their relation to living things."

Hydrologic processes of a hillslope

Role of engineers

The responsibility is to make the best hydrologic estimates or predictions that are possible for
given cost and time constraints of the project.

Comments about predictive techniques:

(a) A large number of different techniques exits to predict the same value. There is no single best technique for all applications.

(b) For a given location, the only way to truly evaluate the accuracy is to compare predicted results against observed data.

(c) Selection of modeling technique is also dependent on the scope of the project. Different techniques require more time and cost considerations. You do not want to spend $20000 on the design of a $1000 culvert.

**Additional definitions**

*Watershed (drainage basin, catchment)*

A watershed (drainage basin, catchment) is a topographically defined area drained by a channel system such that all outflow is discharged through a single outlet.

*Rainfall hyetograph*

The rainfall hyetograph is the time distribution of the rainfall intensity over the watershed. The rainfall excess hyetograph is the time distribution of rainfall excess.

*Total hydrograph*

The total hydrograph is the time distribution of the flow rate at the outlet of the watershed.

*Direct runoff or storm hydrograph*

The direct runoff hydrograph is the time distribution of the runoff (usually surface) from the watershed.
The rainfall hyetograph, the total hydrograph and the direct runoff hydrograph are shown below.

![Rainfall Intensity vs Flow Rate](image)

Time

Figure 1.3 Rainfall hyetograph and total and direct runoff hydrographs.

**PROBABILITY CONCEPTS**

**Need for probabilistic information**

*Design needs*

Engineering projects and planning projects are concerned with future events, which are dependent on the weather.

*Prediction of weather*

Weather is a complex system in which the prediction of absolute time and magnitude of future events can not be predicted. Therefore probabilistic or stochastic modeling techniques are needed.

*Key assumption*

Future events will have the same statistical properties (e.g., mean, variance) as past events. This is called a stationary time series. This concept is shown below using annual rainfall data for the Twin Cities. The mean rainfall depth between 1885 and 1985 is 26.7 in and the standard deviation is 5.6 in.
Types of hydrologic data series

*Complete duration series*: A data series that includes all available data. For many streams, the complete duration series is the average flow rate for each day of the year.

*Partial duration series*: A data series that includes only data values greater than a predefined base value. For example, we might only select daily rainfall depths for analyses that are greater than two inches.

*Annual maximum series*: A data series that includes only the largest data value that occurs within each year of the record. Most of the analysis in BAE 5513 is using the annual maximum series. Annual minimum series includes on the smallest data value that occurs within each year of record. Although this is of interest in some studies, it will not be used in BAE 5513. To simplify terminology, annual series or annual time series will sometimes be used for an annual maximum series.

What is probability?

*Classical definition*: The classical definition is frequently used in gaming problems where the number of ways a particular random event can occur is enumerated. The number of ways of obtaining an annual rainfall depth is difficult to enumerate and therefore an alternative definition is needed.

*Relative frequency definition*

For a random event,
where $n_A$ is the number of observations of events with attribute A and N is the total number of observations.

Since infinite number of observations is not possible, a working definition usually used (but not always) is,

$$P(A) = \frac{n_A}{N}$$

where the above definition becomes more accurate as N becomes larger. Examples are discussed later.

**Definitions**

**Mutually exclusive events**

Events that cannot occur at the same time, or

$$P(A \text{ and } B) = 0$$

Example: Consider a rainfall record in which each day is classified as either a wet or dry day depending on whether it rained. Since a given day cannot be both wet and dry, they are mutually exclusive.

For mutually-exclusive events:

$$P(A \text{ or } B \text{ or } C \text{ or } D) = P(A) + P(B) + P(C) + P(D)$$

If events A, B, C and D represent all possible events, then a simple result that sometimes simplifies calculations is

$$P(A \text{ or } B \text{ or } C) = 1 - P(D)$$

**Independent events**

The probability of an Event A does not dependent on the state of Event B.

$$P(A/B) = P(A)$$

Example: If temperature for a given day is independent of rainfall, then the probability of a given temperature does not depend on whether the day is wet or dry.

For independent events:

$$P(A \text{ and } B \text{ and } C \text{ and } D) = P(A) \cdot P(B) \cdot P(C) \cdot P(D)$$

**Cumulative probability**
Cumulative probability is the probability of an event being less than or equal to a given value.

\[ F = P( A \leq x_1 ) \]

**Exceedance probability**

Exceedance probability is the probability of an event being greater than or equal to a given value.

\[ G = P( A \geq x_1 ) \]

If the probability of obtaining an event exactly equal to \( x_1 \) is zero (which is the case for continuous variables), then exceedance probability and cumulative probabilities are mutually exclusive. Therefore,

\[ F = 1 - G \quad \text{or} \quad G = 1 - F \]

**Probability density functions (pdfs)**

**Definition**

The probability density function is approximately the probability that an event lies within a given range divided by the size of the range. It becomes the pdf for an infinitesimally small class size. Mathematically, it is defined as

\[ f(x) = \frac{df}{dx} = \frac{\Delta P}{\Delta x} \quad \text{[units = probability/x value]} \]

**Cumulative probability**

The cumulative probability can be obtained from a known pdf as

\[ F(x_2) = \int_{-\infty}^{x_2} f(x)dx \]

A graphic illustration of this concept is shown below.
RETURN PERIOD

Hypothetical illustration

Definition

Rainfall depth (or any other hydrologic quantity) of a 10 year event is equaled or exceeded, *on the average*, every 10 years.

Example computation

On the average, how many events in 100 years will be greater than the 10-year event?

Answer: If an event occurs every 10 years, then the solution is how many 10-year periods are in 100 years, or

\[ n_t = \frac{100}{10} = 10 \text{ events} = \frac{N}{T} \]

*Illustration*
Number of occurrences is stochastic

For instance, assume you have an infinite record (imagine-10000 yr) and you have divided the record into 100-year segments (100 segments for 10000 years of data). Some of those 100-year segments will have exactly 10 occurrences of the 10-year event, some will have 9, some 11, etc. The average number of occurrences for all segments, however, will be 10 occurrences.

Formal definition

Return period

The average interval in years in which a hydrologic quantity (such as rainfall depth) is equaled or exceeded.

Mathematical definition

Mathematically, the return period is defined as

$$ T = \lim_{N \to \infty} \frac{N}{n_T} $$

where $N$ is the total number of years (occurrences) and $n_T$ is the number of occurrences greater than or equal to the $T$-year event.

Since an infinite number of years is not possible, a working definition is
By using the relative frequency definition for exceedance probability \((G)\), the return period is also defined as

\[ T = \frac{N}{n_r} \]

A more rigorous derivation for independent events is given in Appendix 1-A using the average recurrence interval.

**BINOMIAL DISTRIBUTION**

**Introduction**

**Key assumption**

Events are independent. For example, annual rainfall depth for one year is not dependent on rainfall depth of the previous year \((P(A/B) = P(A))\).

**Possible outcome**

* E - Event exceeds the T-year event
* N - Event does not exceed (non-exceedance) the T-year event

**Review probability definitions**

* \(G\) = probability of exceeding a T-year event
* \(F\) = probability of not exceeding a T-year event

For mutually exclusive events:

\[ P + G = 1 \quad \text{and} \quad F = 1 - G \]

**Illustrative examples**

**Example #1**

What is the probability of having one event in three years that exceeded the T-year event?

**Solution**

Possibilities:

\[enn, nen, nne\]
The probabilities of ENN for independent events can be computed as

$$P(E \text{ and } N \text{ and } N) = P(E)P(N)P(N) = GFF$$

Since all possibilities are mutually-exclusive, we obtain

$$P(ENN \text{ or } NEN \text{ or } NNE) = GFF + FGF + FFG$$

or

$$f(1;3;G) = 3 \times G^2$$

This is the probability of obtaining 1 event in 3 years where the probability of exceedance is $G$.

**Example #2**

What is the probability of having two events in five years that exceeds the T-year event?

**Solution**

The possibilities are:

EENNN, ENENN, ENNEN, ENNNE, NEENN
NENEN, NENNE, NNEEN, NNENE, NNNEE

The probability of EENNN sequence for independent events is

$$P(E \text{ and } E \text{ and } N \text{ and } N \text{ and } N) = GFFFF$$

Since all possibilities are again mutually-exclusive,

$$P(ENN \text{ or } ENENN \text{ ... or } NNEEN) = GFFFE + GFGEF + GFEEF + GGFFE + GFEEF + FGEFF + FEEFF + FFFE + FFEGF + FFFGF$$

or

$$f(2;5;G) = 10 \times G^2 F^3$$

This is the probability of obtaining 2 events in 5 years where the probability of exceedance is $G$.

**Binomial distribution**

By generalizing the above two problems, we obtain the binomial distribution defined as

$$f(k;n;G) = \frac{n!}{(n-k)!k!} G^{k} (1-G)^{n-k}$$
where 0! = 1. Note that \( \frac{n!}{(n-k)!k!} = \text{combinations of selecting } k \text{ items from } n \).

**Problem #1**

**Problem statement**

What is the probability of exactly 5 occurrences of a 20-year event in 100 years? of exactly 4 occurrences? of exactly 6 occurrences?

**Common parameters**

Let's review the binomial distribution,

\[
f(k;n;G) = \frac{n!}{(n-k)!k!} \ G^k \ F^{n-k}
\]

where for this problem we can define symbols as

\[
G = \frac{1}{T} = \frac{1}{20} = 0.05 \\
F = 1 - G = 0.95 \\
n = 100
\]

**Solution for \( k = 5 \) occurrences**

The solution using the binomial distribution for \( k = 5 \) is

\[
f(5; 100; 0.05) = \frac{(100)(99)(98)(97)(96)}{(5)(4)(3)(2)} \ (0.05)^5(0.95)^95
\]

\[
f(5; 100; 0.05) = 0.180 \text{ or } 18\%
\]

In a large number of 100-year records, you would expect 18% to have exactly 5 occurrences of the 20-year event. The remaining records would have 0, 1, 2, 3, 4, 6, 7 ... or 100 occurrences.

**Solution for \( k = 4 \) occurrences**

The solution using the binomial distribution for \( k = 4 \) is

\[
f(4; 100; 0.05) = \frac{100!}{4!96!} \ (0.05)^4(0.95)^96
\]

\[
f(4; 100; 0.05) = 0.178 \text{ or } 17.8\%
\]

**Solution for \( k = 6 \) occurrences**

Likewise for \( k = 6 \), the binomial distribution can be solved as
Problem #2

Problem statement

Assume that you are considering buying a house just within the flood plain. Flood plains are defined for return periods of 100 years. What is the probability that one or more times you will be flooded within the 30-year life of a mortgage?

Solution

The probability of one or more floods can be determined as,

$$P(\text{one or more}) = 1 - f(0;30;0.01)$$

By using the binomial distribution, we obtain

$$P(\text{one or more}) = 1 - \frac{30!}{0!30!} (0.01)^0 (0.99)^{30} = 0.26$$

RISK ANALYSIS

Background

Risk is usually evaluated for one or more exceedance, that is, if one exceedance is bad then two exceedances are worse.

Computational formulae

Design confidence

Design confidence will be defined as the probability of no events within the n-year life of the project. The probability of no events is defined from the binomial distribution as

$$C = P(\text{no occurrence}) = f(0;n;G) = \frac{n!}{0!n!} G^n (1-G)^0$$

By evaluating terms and using $G = 1/T$, the design confidence is defined as

$$C = (1 - \frac{1}{T})^T$$

Design risk

Design risk will be defined as the probability of one or more events within the n-year life of
the project. Design risk is then obtained as
\[ R = P(1 \text{ or more}) = 1 - f(\theta; \mu; \sigma) = 1 - C \]

By using the above definition for \( C \), we obtain
\[ R = 1 - \left(1 - \frac{1}{T}\right)^n \]

*Design return period*

Design return periods will be defined from the level of confidence of no events greater than a T event in n years. Let’s review the definition of design confidence
\[ C = (1 - \frac{1}{T})^n \]

By solving for return period, we obtain
\[ T = \frac{1}{1 - \frac{1}{T^{th}}} \]

**Problem #1**

**Problem statement**

What is the probability of having one or more occurrence of a 10-year event within the 10-year life of a project?

**Solution**

By using the definition of design risk,
\[ R = 1 - \left(1 - \frac{1}{T}\right)^n = 1 - \left(1 - \frac{1}{10}\right)^{10} \]

we obtain
\[ R = 0.65 \]

that is, there is a 65% chance of a 10-year event within a 10-year life of the project.

**Problem #2**

**Problem statement**

What is the required return period to use to be 90% sure that the design of the project will not be exceeded within the 25 year life of the project?

**Solution**

By using the equation developed for return period
we obtain,

\[ T = 238 \text{ years} \]

*Why is this return period so large?*

The need for a large return period can be illustrated by considering a standard deck of cards (52 cards). Let’s assume that you are interested in the probability of not being dealt a red-faced card (hearts or diamonds). The probability of not getting a red-faced card is defined as

\[ 1 - \frac{26}{52} - \frac{1}{2} \]

Similar to the above problem, we might be interested in the probability of not being dealt a red-faced card in 25 consecutive tries. The probability of this happening (with replacement and reshuffling of the deck) is simply

\[ C = \left(\frac{26}{52}\right)^{25} \approx 0.000000001 \text{ or } 0.000000001% \]

This is clearly an unlikely possibility because half of the cards in the deck are red-faced. If we are instead interested in the probability of not getting a queen of hearts in 25 consecutive tries, the equivalent computational steps are

\[ C = \left(\frac{1}{52}\right)^{25} \approx 0.000000001 \text{ or } 0.000000001% \]

There is a 61.5% chance of not being dealt a queen of hearts in 25 consecutive tries. This value, however, is still considerably less than the 90% confidence criteria used in the above hydrology problem. To have a 90% chance of not having a queen of hearts dealt, we would need to mix additional decks of cards, where the queen of hearts is removed from these decks. The total number of cards would need to be 238 cards (approximately 3.5 have additional decks). For our hydrology example problem, it therefore takes a large return period of 238 years to ensure that, in 25 consecutive years, the probability of not getting-the-event is 90%.

**Summary of binomial distribution**

*Application to design*

Binomial distribution:

- Determine risk for a given return period, or
- Determine return period for a given risk or confidence

This step is often done for you by a governmental agency.

**Missing information for design**

For design and/or hydrologic analysis:

- Magnitude (e.g., rainfall depth) for a given return period, or
- Return period for a given magnitude.

This information is often obtained from a frequency analysis discussed in the next sections.

**FREQUENCY ANALYSIS USING PLOTTING POSITION**

*Example data set*
Introduction

Frequency analysis refers to procedures used to estimate the magnitude (i.e., inches or cfs) for a given return period or the return period for a given magnitude. One frequency-analysis technique is called the plotting position method. The plotting position method will be described using measured flow rates for the Minnesota River at Mankato, Minnesota. Procedures to estimate flow rates is shown schematically below. The annual peak flow rate for 1964 through 1994 will be used in the analysis. The flow record of annual peak rates at Mankato is from 1903 to the current year.

Example data set

<table>
<thead>
<tr>
<th>Year</th>
<th>Peak Flow (cfs)</th>
<th>Year</th>
<th>Peak Flow (cfs)</th>
<th>Year</th>
<th>Peak Flow (cfs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1964</td>
<td>19400</td>
<td>1975</td>
<td>24100</td>
<td>1985</td>
<td>29700</td>
</tr>
<tr>
<td>1965</td>
<td>94100</td>
<td>1976</td>
<td>5130</td>
<td>1986</td>
<td>36300</td>
</tr>
<tr>
<td>1966</td>
<td>15400</td>
<td>1977</td>
<td>7850</td>
<td>1987</td>
<td>17000</td>
</tr>
<tr>
<td>1967</td>
<td>18700</td>
<td>1978</td>
<td>13300</td>
<td>1988</td>
<td>5520</td>
</tr>
<tr>
<td>1968</td>
<td>15800</td>
<td>1979</td>
<td>30000</td>
<td>1989</td>
<td>15800</td>
</tr>
<tr>
<td>1969</td>
<td>76700</td>
<td>1980</td>
<td>15700</td>
<td>1990</td>
<td>17100</td>
</tr>
<tr>
<td>1971</td>
<td>21400</td>
<td>1982</td>
<td>15500</td>
<td>1992</td>
<td>23900</td>
</tr>
<tr>
<td>1972</td>
<td>20200</td>
<td>1983</td>
<td>33300</td>
<td>1993</td>
<td>75600</td>
</tr>
<tr>
<td>1973</td>
<td>19700</td>
<td>1984</td>
<td>41000</td>
<td>1994</td>
<td>21800</td>
</tr>
<tr>
<td>1974</td>
<td>12500</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Time series plot

It is useful to plot the flow rates for each year in a time series plot. This plot is useful in determining any trends in the data. If the plot shows a trend for the flow rates to increase or decrease with time, then the assumption of a stationary time series (future statistics = past
Exceedance probability

The exceedance probability can be estimated using the relative frequency definition previously given. The first step is to rank the data from the largest to smallest. The table below illustrates the computations of exceedance probabilities.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Flow rate (cfs)</th>
<th>Number of observations</th>
<th>G-Relative Frequency (n/N)</th>
<th>Plotting Position (n/N+1)</th>
<th>Plotting Position (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>94100</td>
<td>1</td>
<td>1/31</td>
<td>1/32</td>
<td>3%</td>
</tr>
<tr>
<td>2</td>
<td>76700</td>
<td>2</td>
<td>2/31</td>
<td>2/32</td>
<td>6%</td>
</tr>
<tr>
<td>3</td>
<td>75600</td>
<td>3</td>
<td>3/31</td>
<td>3/32</td>
<td>9%</td>
</tr>
<tr>
<td>4</td>
<td>41000</td>
<td>4</td>
<td>4/31</td>
<td>4/32</td>
<td>13%</td>
</tr>
<tr>
<td>5</td>
<td>36300</td>
<td>5</td>
<td>5/31</td>
<td>5/32</td>
<td>16%</td>
</tr>
<tr>
<td>6</td>
<td>33300</td>
<td>6</td>
<td>6/31</td>
<td>6/32</td>
<td>19%</td>
</tr>
<tr>
<td>7</td>
<td>30000</td>
<td>7</td>
<td>7/31</td>
<td>7/32</td>
<td>22%</td>
</tr>
</tbody>
</table>
### Graphical solution

#### Overview

To determine probabilities for flow rates not given in the table (or to estimate flow rates for probabilities not in the table), the plotting position method uses graphical solutions.

#### Arithmetic plot

A plot of exceedance probabilities (plotting position values) versus flow rate using arithmetic scale is shown below. Hydrologists are frequently interested in return periods of 25 years or greater. A return period of 25 year corresponds to a exceedance probability of 4%. Clearly, this type of plot will be difficult to read values corresponding to return periods of 25 years or greater.
To improve the resolution of scale in the values of greatest interest, probability plots are usually used. The scale for these plots are defined so that values are divided around the 50% value.

Probability plots can also be used to evaluate whether the data is normally or log-normally distributed. Normally distributed data plot as a straight line using an arithmetic vertical scale and log-normally data plot as a straight line using a log vertical scale.
Summary of computational steps

Step #1: Rank the data from largest to smallest.

Step #2: Calculate plotting position using modified relative frequency definition,

\[ PP = \frac{\text{rank}}{N+1} \]

Step #3: Plot on probability paper.

Step #4: Interpolate and/or extrapolate to determine flow rate for desired return period.

Disadvantages

Importance of largest events

The largest events have major influence on the estimation of return period values. We are uncertain as to the appropriate return period for these events.

For the Minnesota River data, the flow rate for the return period of 100 years \( (G=1\%) \) depends primarily on the three largest values (e.g., \( G= 3\%, \ 6\% \) and \( 9\% \)). If we look at the entire record at Mankato (1903-1995), these values are the largest value in this record. The
corresponding plotting position would then change to approximately 1%, 2% and 3%. Such a shift would have substantial impact on estimated flow rate for a 100-year event.

From the binomial distribution, probability of one or more 100-year event in the 31 years of record can easily be computed as

\[ P(\text{one or more}) = R = 1 - (1 - \frac{1}{T})^3 \]

\[ R = 1 - (1 - \frac{1}{100})^{31} = 0.26 \]

Subjective judgement

A visually fitted curve through observed points requires a subjective judgement. It is possible that different hydrologists could get drastically different flow rates for the same return period. This may not be desirable, especially for a large agency.

FREQUENCY ANALYSIS USING ANALYTICAL CURVES

Introduction

Goal

Analytical curves are fitted to the observed data to reduce the subjectivity of the hydrologist and, usually, to moderate the importance of two or three data points in the frequency analysis.

There are two issues in using analytical curves: (1) equation type and (2) fitting the equation to the data.

What type of equation to use?

Answer: Probability density functions

How to estimate the equations parameters?

Answer: Using statistical properties of observed data.

A key assumption in using the statistical properties of future events are equal to the statistical properties of past events. Probability density functions are fitted to the data so that the statistics (some of them) match those observed.

Statistical characteristics
Central tendency

The central tendency of data is illustrated below.

Mean is the most frequently used parameter to describe the central tendency and can be computed as

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Spread of data

The concept of spread in the data is shown below for two hypothetical watersheds.

Variance is a commonly used spread statistic and is defined as

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$
or using the following simpler computational formula as
\[
\sigma_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i^2) - \bar{x}^2
\]

The standard deviation is defined as
\[
\sigma_x = \sqrt{\sigma_x^2}
\]

**Application to probability plots**

The use of statistical properties to fit the data is shown below for a normal distribution.

![Probability density functions](image)

**Probability density functions**

**Review**

Probabilities are determined from the area under the pdf curves. As previously given, cumulative probability is defined as

\[
F(x-x_0) = \int_{-\infty}^{x} f(x) \, dx
\]

and represented graphically as
**Criteria**

There are a limitless number of functions that can be used as pdfs. They need to satisfy two criteria:

\[ f(x) \geq 0 \text{ for all } x \]

and

\[ \int_{-\infty}^{\infty} f(x) \, dx = 1 \]

Some pdfs have a domain that is bounded, for example, \( f(x)=0 \) for \( x<0 \). The domain of the distribution is usually specified along with the distribution.

Some commonly used pdfs in frequency analysis are

* Exponential
* Normal
* Log Normal
* Extreme Value Type I (Gumbel's)
* Log Pearson Type III

**EXTREME VALUE TYPE I DISTRIBUTION**

**Background**

**General**

This distribution is widely used to represent rainfall data.

Other names include Gumbel's extreme value distribution, the Fisher-Tippett type I distribution, and the double exponential distribution. A generalized extreme value
distribution has been proposed more recently using parameters estimated from L-moments. Details of this approach are given in the appendix.

Statistical characteristics

The parameters of the extreme value type I distribution is defined such that:

- Pdf's mean = Mean of past events
- Pdf's standard deviation = Standard deviation of past events
- Pdf's skew coefficient = 1.139

Probability density function - pdf

The probability density function for the extreme value type I distribution is shown below.

\[ f(x) = q \exp \left[ -q(x - w) - \exp(-q(x-w)) \right] \quad \text{for } -\infty < x < \infty \]

where

\[ q = \frac{\pi}{s \sqrt{6}} \quad \text{and} \quad w = \bar{x} - \frac{0.5572}{q} \]

The extreme value type I distribution has the following shape.

\[ C_s = 1.139 \]

Flow Rate - untransformed data

Probability for a given magnitude

Background

Hydrologists are sometimes required to determine the probability corresponding to an observed rainfall or flow rate event. Equations are presented in this section to determine the cumulative probability, the exceedance probability and the return period for a given magnitude of an event.

Cumulative distribution
The cumulative distribution function is used to compute the probability that the hydrologic quantity (e.g., rainfall, peak flow rate) is less than or equal to a given magnitude. Mathematically, it corresponds to the area under the curve from -infinity to the magnitude of interest, $x_1$, or for the extreme value type I distribution

$$F(x) = \int_{-\infty}^{x} \left[ q \exp(-q(x-w)) - \exp(-q(x-w))) \right] \, dx$$

An analytical solution for the extreme value type I distribution can be obtained by first integrating with the variable

$$t = q(x - w) \quad \text{and therefore} \quad dx = dt/q$$

$$t = -\infty \quad \text{when} \quad x = -\infty \quad \text{and} \quad t = t_2(x-w) \quad \text{when} \quad x = x_2$$

By evaluating the cumulative distribution we respect to variable $t$, we obtain

$$F(x) = F(t_2) = \int_{-\infty}^{t_2} \exp(-t) \exp(-\exp(-t)) \, dt$$

that can be solved easily using

$$v = -\exp(-t); \quad dv = \exp(-t)dt$$

$$v = -\infty \quad \text{when} \quad t = -\infty \quad \text{and} \quad v = v_2 = -\exp(-t_2)$$

and therefore

$$F(t_2) = \int_{-\infty}^{-\exp(-t_2)} \exp(v) \, dv$$

or

$$F(t_2) = \exp(-\exp(-t_2)) - \exp(-\infty) = \exp(-\exp(-t_2))$$

In general, the cumulative distribution function can be evaluated as

$$F(x) = \exp(-\exp(-q(x-w)))$$

**Exceedance probability**

In hydrologic analysis, interest usually lies in the probability of the hydrologic quantity being greater than or equal to a value. The exceedance probability for the extreme value type I distribution is defined as

$$G(x) = 1 - F(x) = 1 - \exp(-\exp(-q(x-w)))$$

**Return period**
The return period for a given magnitude is simply obtained as

\[ T = \frac{1}{g(x)} = \frac{1}{1 - \exp\left(-\exp\left(-q(x-w)\right)\right)} \]

**Magnitude for a given return period**

**Background**

For hydrologic design, the return period is specified for the design either by a governmental agency or by using risk analysis concepts previously discussed. Interest is then in determining the magnitude \( x \) for a given return period \( T \).

**Solution**

For the extreme value type I distribution, we can arrange the above equation as

\[ 1 - \exp\left(-\exp\left(-q(x-w)\right)\right) = \frac{1}{T} \]

or by rearranging terms

\[ q(x - w) = t = -\ln[-\ln(1-1/T)] \]

which can be solved for the unknown magnitude, \( x \), as

\[ x = w + \frac{t}{q} \]

**EXAMPLE PROBLEM**

**Problem definition**

**Problem statement**

For the annual peak flow data for the Thief River watershed given below, estimate the flow rates corresponding to a return period of 10 years and 100 years using (1) the plotting position method and (2) the extreme value type I distribution.

**Peak flow rates for Thief River near Thief River Falls, Minnesota**

The observed annual peak flow rates (largest flow rate recorded for each year) is given below.

<table>
<thead>
<tr>
<th>Year</th>
<th>Discharge (cfs)</th>
<th>Year</th>
<th>Discharge (cfs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1976</td>
<td>1350</td>
<td>1986</td>
<td>2420</td>
</tr>
<tr>
<td>1977</td>
<td>403</td>
<td>1987</td>
<td>1570</td>
</tr>
<tr>
<td>1978</td>
<td>2740</td>
<td>1988</td>
<td>600</td>
</tr>
</tbody>
</table>
For the Thief River data, the statistics of observed data are

\[ \sum_{i=1}^{16} x_i = 1350 + 403 + \ldots + 2180 + 1960 = 27974 \text{ cfs} \]
and compute the mean as

$$\bar{x} = \frac{1}{19} \sum_{i=1}^{19} x_i = 27974/19 = 1472 \text{ cfs}$$

and the standard deviation as

$$s = \sqrt{\frac{\sum (x_i^2 - N \bar{x}^2)}{N - 1}}$$

$$s = \sqrt{\frac{(5.214353 \times 10^7) - (19 \times 1472)^2}{18}} = 780.2 \text{ cfs}$$

We can now compute the statistics for Extreme Value Type I as

$$q = \frac{\pi}{s \sqrt{6}} = \frac{\pi}{780 \sqrt{6}} = 0.00164 (1/\text{cfs})$$

$$w = \bar{x} - \frac{0.5572}{q} = 1472 - \frac{0.5572}{0.00164} = 1133 \text{ cfs}$$

Solution with the plotting position method

Plotting position points

A rank of the observed data and the corresponding plotting position values for the Thief River are shown below.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Flow rate (cfs)</th>
<th>pp</th>
<th>Rank</th>
<th>Flow rate (cfs)</th>
<th>pp</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2740</td>
<td>0.05</td>
<td>11</td>
<td>1500</td>
<td>0.55</td>
</tr>
<tr>
<td>2</td>
<td>2420</td>
<td>0.10</td>
<td>12</td>
<td>1350</td>
<td>0.60</td>
</tr>
<tr>
<td>3</td>
<td>2300</td>
<td>0.15</td>
<td>13</td>
<td>1350</td>
<td>0.60</td>
</tr>
<tr>
<td>4</td>
<td>2180</td>
<td>0.20</td>
<td>14</td>
<td>1220</td>
<td>0.70</td>
</tr>
<tr>
<td>5</td>
<td>2130</td>
<td>0.25</td>
<td>15</td>
<td>620</td>
<td>0.75</td>
</tr>
<tr>
<td>6</td>
<td>2130</td>
<td>0.30</td>
<td>16</td>
<td>600</td>
<td>0.80</td>
</tr>
<tr>
<td>7</td>
<td>1960</td>
<td>0.35</td>
<td>17</td>
<td>403</td>
<td>0.85</td>
</tr>
<tr>
<td>8</td>
<td>1710</td>
<td>0.40</td>
<td>18</td>
<td>211</td>
<td>0.90</td>
</tr>
<tr>
<td>9</td>
<td>1570</td>
<td>0.45</td>
<td>19</td>
<td>150</td>
<td>0.95</td>
</tr>
</tbody>
</table>
where  

$$pp = \frac{rank}{N+1} = \frac{rank}{20}$$

**Plot of data**

The values in the above table are plotted using log-probability paper. This plot is used to determine the flow rates corresponding to 10-year and 100-year events.

**Solution from graph**

For $T=10$ years ($G=0.1=10\%$)  

$$x_{10} = 2600 \text{ cfs}$$

For $T=100$ years ($G=0.01=1\%$)  

$$x_{100} = 4200 \text{ cfs}$$

**Solution with the extreme value type I distribution**
Computational equations

The flow rate, \( x \), has previously been shown to be computed from the integration variable \( t \) as

\[
x = w + \frac{t}{q}
\]

where the integration variable is computed for a given return period as

\[
t = -\ln[-\ln(1 - 1/T)]
\]

The extreme value type I statistics have been previously computed as

\[
q = 0.00164 \text{ (1/cfs)} \quad \text{and} \quad w = 1133 \text{ cfs}
\]

Solutions

For 10-year return period, variable \( t \) is defined as

\[
t = -\ln[-\ln(0.9)] = 2.250
\]

and the 10-year flow rate is

\[
x_{10} = 1133 + \frac{2.250}{0.00164} = 2505 \text{ cfs}
\]

For 100-year return period, we obtain

\[
t = -\ln[-\ln(0.99)] = 4.600
\]

\[
x_{100} = 1133 + \frac{4.600}{0.00164} = 3938 \text{ cfs}
\]

Comparison of methods

We will again evaluate the appropriateness of using the extreme value type I distribution by plotting these values with those obtained from the plotting position method. We will also show the results for the normal distribution.
The extreme value type I distribution and the normal distribution are both reasonable approximations to the plotting position values. The extreme value type I should be used for a conservative estimate. Sometimes the distribution is specified for the hydrologist by standard procedures of an agency or consulting firm. Greater confidence in using the extreme value type I pdf can be obtained by investigating the fit for other gaging stations in the Red River Basin.
Uncertainty/Risk Problem Assignment

Due Date: ________________________________

Format

All homework assignments must be done in a neat and organized manner. Clearly identify the solution in those with several mathematical steps by underlining or boxing in your answer. Points will be deducted from problems that are sloppy and difficult to follow.

Problems

(15 points)
1. Using U.S. Geological Survey records, you have obtained 50 years of data for annual peak flow rates at a given gaging station. Assume that these flow rates can be evaluated as a Bernoulli process. Calculate the probability that this 50-year record contains one or more events with a return period of 100 years. What is the probability that this record contains one or more events with a return period of 500 years? What is the probability that this record contains exactly 5 events with a return period of 10 years?

(10 points)
2. You are a hydrologist employed by a regulatory agency. Your responsibility is to set regulations to minimize the adverse environmental impact of construction activities and surface mining operations. After talking to biologists, sociologists, and economists on your staff, you have decided to set regulations so that you are 95% confident that pollution control devices will not fail within the 5-year life of these land disturbing projects. To be 95% confident of no failures, what return period will you select for the design of these pollution control devices?

(15 points)
3. A businessman is considering the construction of a shipping facility near a river. This facility will be designed for a life of 10 years. His engineers have determined that this facility will be flooded if the flow rate in the river exceeds 100,000 cfs. By using a frequency analysis, the return period of this flow rate has been estimated at 35 years. An economic analysis indicates that the total profit of the facility over 10 years is 1 million dollars, neglecting any flooding losses. It is estimated that the total profit can be reduced $600,000 dollars for each flooding event. What is the probability that the businessman will have a profit of 1 million dollars? What is the probability that the businessman will have a profit of at least $400,000? What is the probability that the businessman will lose money?
1-33

(15 points)
4. You have recently installed a rainfall network consisting of 30 automatic, recording rainfall gages. Your equipment vendor has now informed you that these gages were part of a larger stock in which manufacturing defects were found in some of the gages. The defective gages were randomly distributed in the stock. They have estimated that the probability of any single gage being defective is 0.1. What is the probability that two or more of your rainfall gages are defective?

(45 points)
5. For an experimental watershed located near Stillwater, OK, the maximum rainfall depth to occur in 5 minutes was determined for twenty-two consecutive years. These values are given below. You are required to:

(a) Obtain the plotting position values and plot these values using the attached probability paper. Draw you best "eyeball" curve to fit these points. What are the rainfall depths corresponding to return periods of 2 years, 10 years and 100 years?

(b) Determine the statistics necessary for the extreme value type I distribution. What are the rainfall depths corresponding to the 2-year, 10-year and 100-year return periods using the extreme value type I distribution?

(c) Which method (plotting position or extreme value type I distribution) would you use to estimate the 5-minute duration rainfall depth for the 2-year return period? the 10-year return period? the 100-year return period? Why?

Maximum Yearly 5 Minute Rainfall Depths
Rain Gage R-3 near Stillwater, OK

<table>
<thead>
<tr>
<th>Year</th>
<th>Depth (in)</th>
<th>Year</th>
<th>Depth (in)</th>
<th>Year</th>
<th>Depth (in)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1951</td>
<td>0.37</td>
<td>1960</td>
<td>0.44</td>
<td>1970</td>
<td>0.32</td>
</tr>
<tr>
<td>52</td>
<td>0.52</td>
<td>61</td>
<td>0.59</td>
<td>71</td>
<td>0.26</td>
</tr>
<tr>
<td>53</td>
<td>0.35</td>
<td>62</td>
<td>0.51</td>
<td>72</td>
<td>0.45</td>
</tr>
<tr>
<td>54</td>
<td>0.39</td>
<td>63</td>
<td>0.55</td>
<td></td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>0.46</td>
<td>64</td>
<td>0.35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>56</td>
<td>0.55</td>
<td>65</td>
<td>0.38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>57</td>
<td>0.51</td>
<td>66</td>
<td>0.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>58</td>
<td>0.50</td>
<td>67</td>
<td>0.45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>59</td>
<td>0.44</td>
<td>68</td>
<td>0.29</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>69</td>
<td>0.35</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX 1-A: RELATED MATERIALS

Average recurrence interval

Recurrence interval is defined as time interval between events of a particular exceedance probability (i.e., G). Let’s limit our discussion to recurrence intervals corresponding to integer values of years. Following logic given for the binomial distribution, the probability of having recurrence interval of \( \tau \) years corresponds to \((1-\tau)\) years of not having an event, followed by an event in the next year. This probability is therefore defined as:

\[
P(N \cap N \cap \ldots \cap N \cap E) = (1-G)^{\tau-1} \cdot G
\]

Mean (average) or expected value is typically defined as the center of mass of a pdf. For discrete probability defined by the equation, the expected value of the recurrence interval is then defined as

\[
E(\tau) = \sum_{\tau=1}^{\infty} \tau \cdot (1-G)^{\tau-1} \cdot G = G \left[ 1 + 2(1-p) + 3(1-p)^2 + 4(1-p)^3 + \ldots \right]
\]

Let’s review Taylor series expansion of the following power function around \( x = 0 \)

\[
(1 + x)^n = 1 + nx + \frac{n(n-1)}{2} x^2 + \frac{n(n-1)(n-2)}{3!} x^3 + \ldots + \frac{n(n-1)\ldots(n-k+1)}{k!} x^k + \ldots
\]

A careful inspection of this expansion with the bracket term in the previous equation allows one to conclude that they are equal for \( x = -(1-p) \) and \( n = -2 \). We are then able to evaluate the average recurrence interval as

\[
E(\tau) = \frac{G}{[1 - (1-G)]^2} = \frac{1}{G}
\]

that is, the average recurrence interval equals the inverse of the exceedance probability, which corresponds to the definition of the return period given in the chapter.

Generalized extreme value distribution

Definition

Extreme value type I, type II and type III (sometimes called the Gumbel, Frechet, and Weibull, respectively) distributions are represented by the generalized extreme value (GEV) distribution. These distributions are derived from applying statistical theory to the maxima of independent and identically distributed random variables. Consequently, GEV is often used to represent the maxima of a long sequence of random variables (such as maximum annual flow rate).

The probability density function is defined as

\[
f(x) = \frac{1}{\alpha} \left[ 1 - \frac{k(x - w)}{\alpha} \right]^{1/\alpha-1} \exp \left\{ - \left[ 1 - \frac{k(x - w)}{\alpha} \right]^{1/\alpha} \right\} \text{ for meaningful } x \text{ values}
\]

The cumulative distribution is readily obtained using integration variable of
\( u = [1 - k(x - w)/q]^{1/k} \). The cumulative distribution is then defined as

\[
P(u) = \int_{-\infty}^{u} \exp(-u) \, du = \exp(-u)
\]

and therefore

\[
P(x) = \exp \left\{ -\left[ 1 - \frac{k(x - w)}{q} \right]^{1/k} \right\}
\]

The above equation can be rearranged to solve for \( x \) for a given cumulative distribution by first

\[
[- \ln(P)]^{k} = 1 - \frac{k(x - w)}{q}
\]

and therefore

\[
x = w + \frac{q}{k} \left[ 1 - [- \ln(P)]^{k} \right]
\]

Parameter values using L-moments

In contrast to the classical moment definitions given in the chapter for the extreme value type I distribution, the parameters for GEV are often estimated using the L-moments. The L-moments are defined using theoretical concepts from ordered statistics (see Landwehr et al. 1979. Probability Weighted Moments Compared with Some Traditional Techniques in Estimating Gumbel Parameters and Quantiles. Water Resource Research, Vol. 15(5): 1055-1064, and Hosking. 1990. L-moments: Analysis and Estimation of Distributions using Linear Combinations of Order Statistics, J. R. Statist. Soc. B, Vol. 52(1): 105-124.) The data are first ordered from smallest to largest, that is \( x_1 < x_2 < x_3 < \ldots < x_n \). Conceptually, we can consider the total number of possible ways to obtain a subsample of a particular size from these \( n \) values. We can also compute the number of ways that these subsamples contains \( x_i \). Likewise, we can compute the number of ways of containing \( x_2 \) and not \( x_1 \) and so forth for each observation. These relationships are used to weigh the contribution of each \( x \) value to the computation of the L-moment.

The definition of the first three L-moments are

\[
L_1 = \bar{x} \\
L_2 = 2b_1 - \bar{x} \\
L_3 = 6b_2 - 6b_1 + \bar{x}
\]

\( t_3 = L_3/L_2 \)

where \( b_1 \) and \( b_2 \) are defined using the ordered \( x \) values as

\[
b_1 = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{i - 1}{n - 1} \right) x_i \quad \text{and} \quad b_2 = \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{(i - 1)(i - 2)}{(n - 1)(n - 2)} \right] x_i
\]
The parameter values for the GEV can be computed from the L-moments as

\[ z = \frac{2}{3 + t_3} - \frac{\ln(2)}{\ln(3)} \quad \text{and} \quad k = 7.859 z + 2.9554 z^2 \]

\[ q = \frac{L_2 k}{(1 - 2^{-k}) \Gamma(1 + k)} \quad \text{and} \quad w = \bar{x} + q \left( \frac{\Gamma(1 + k) - 1}{k} \right) \]

**Comparison of results**

A comparison of the results using the extreme value type I and the GEV distributions for the 5-minute and 60 minute rainfall depths are shown below.