CIVE 6361 Engineering Hydrology

Probability Estimation Modeling

Probability Estimation Modeling

Probability estimation modeling refers to the use of probability distributions to model or explain behavior in observed data. Once a distribution is selected, then the concept of risk (probability) can be explored for events (rainfalls, discharges, concentrations, etc.) of varying magnitudes. Generally two important "extremes" are important in engineering; relatively uncommon events (floods, plant explosions, etc.) and very common events (routine discharges, etc.). The concepts of analysis are the same, but the distribution models are very different. In fact, there is very little literature on the

examination of very common events, yet this area is crucial in environmental engineering, and to a lesser extent in typical civil engineering applications.

Frequency Analysis

Frequency analysis typically attempts to relate the behavior of some variable over some recurring time intervals. The time interval is assumed to be large enough so that the concept of "frequency" makes sense. If the time intervals are short, we are often dealing with a time-series that is handled using different tools. Underlying the idea of "long enough" is the concept of independence, that is the values of the variable are statistically independent, otherwise the variables are said to be serial (or auto-) correlated.

a). Concept of T-year event.

Extreme values (annual maximum or annual minimum) on some periodic time basis (usually one year) vary in magnitude (value) in an apparently random fashion. The T-year event concept is a way of expressing the probability of observing an event of some specified magnitude or smaller (larger) in one sampling period (one year).

The formal definition is: The T-year event is an event of magnitude (value) over a **long** time-averaging period, whose average arrival time between events of such magnitude is T-years.

This definition is not very useful because it is often misinterpreted as an event of such magnitude occurring ON THE AVERAGE every T years. The operational definition (one that makes more sense) is that a T-year event is an event magnitude whose probability of occurrence in a single sampling interval is 1/T (FEMA sometimes calls this the (1/T) chance event, and this is the correct way to state the probability).

The concept can be extended to other sampling intervals, but almost never is in the regulatory world. The concept assumes that the extreme values observed in sequential events are independent (serial correlation is nearly zero). Obviously if the sampling interval is short (daily) we expect strong serial correlation and the T-day event is an absurd statement.

Typically we also assume that the process is stationary over the number of sampling intervals of record. This assumption is usually not realistic in a philosophical sense (it implies nothing changes with time), but in practice it is not bad because hydrologic processes are natural integrators and tend to smooth out minor changes. Over time, or where dramatic changes occur, we do expect changes in the statistics, and these can be detected by various hypothesis tests on censored data. Also, most textbooks, underemphasize the causal relationship between rainfall and runoff. A lot of the frequency methods are developed to explain behavior of extreme flows and implicitly assume that the extreme flow is caused by the extreme precipitation event. There is often little evidence of such cause because precipitation is a distributed (spatially and temporally) variable as compared to discharge at a single location on a watershed.

b). Notation:

$P[x > X] = y$

Most probability notations are similar to the above statement. We read them as "The probability that the random variable *x* will assume a value greater than *X* is equal to *y*"

The part of the expression " $x > X$ " is the "event" so when you read probability texts they talk about universes and events but use above notation. It looks confusing at first, but with some familiarity you get used to the notation.

c). Risk/Loss

The probability in a single sampling interval is useful in its own sense, but we are often interested in the probability of occurrence (failure?) over many sampling periods.

If the individual sampling interval events are IID (independent, identically distributed) then we satisfy the requirements of a Bernoulli process.

d). Bernoulli process.

As a simple example, assume the probability that we will observe a cumulative daily rainfall depth equal to or greater than that of TS Allison is 0.10 (Ten percent).

What is the chance we would observe one or more TS Allison's in a three-year sequence?

For a small problem we can enumerate all possible outcomes.

There are eight configurations we need to consider:

	Year1	Year2	Year3	Probability
1	No TSA	No TSA	No TSA	(.9)(.9)(.9)=0.729
2	No TSA	No TSA	TSA	(.9)(.9)(.1)=0.081
3	No TSA	TSA	No TSA	(.9)(.1)(.9)=0.081
4	TSA	No TSA	No TSA	(.1)(.9)(.9)=0.081
5	No TSA	TSA	TSA	(.9)(.1)(.1)=0.009
6	TSA	TSA	No TSA	(.1)(.1)(.9)=0.009
7	TSA	No TSA	TSA	(.1)(.9)(.1)=0.009
8	TSA	TSA	TSA	(.1)(.1)(.1)=0.001

So if we are concerned with one storm in the next three years the probability of that outcome is 0.243 (outcomes 2,3,4; probabilities of mutually exclusive events add).

The probability of three "good" years is 0.729. The probability of the "good" outcomes decreases as the sampling intervals are increased. So over the next 10 years, the chance of NO STORM is $(.9)^{10} = 0.348$. Over the next 20 years, the chance of NO STORM is $(.9)^{20} = 0.121$. Over the next 50 years, the chance of NO STORM is $(.9)^{50} = 0.005$ (almost assured of a storm in the next 50 years). To pick the chances of *k* storms in *n* sampling intervals we use the binomial distribution.

$$P[k-events, n-samples, p_T] = \frac{n!}{(n-k)!k!} p_T^k (1-p_T)^{n-k}$$

This distribution enumerates all outcomes **assuming** unordered sampling without replacement.

There are several other common kinds of counting:

1. ordered with replacement (order matters), samples are replaced
2. unordered with replacement
3. ordered without replacement

Once we have probabilities we can evaluate risk.

Insurance companies use these principles to determine your premiums.

In the case of insurance one can usually estimate the dollar value of a payout – say one million dollars. Then the actuary calculates the probability of actually having to make the payout in any single year, say 10%. The product of the payout and the probability is called the expected loss. The insurance company would then charge at least enough in premiums to cover their expected loss.

They then determine how many identical, independent risks they have to cover to make profit (this is the basic concept behind the flood insurance program, if enough people are in the risk base, the probability of all of them having a simultaneous loss is very small, so the losses can be covered plus some profit).

If we use the above table (let the Years now represent different customers), the probability of having to make one or more payouts is 0.271.

	Customer 1	Customer 2	Customer 3	Probability	E(loss)
1	No Loss	No Loss	No Loss	0.729	0
2	No Loss	No Loss	Loss	0.081	$81,000
3	No Loss	Loss	No Loss	0.081	$81,000
4	Loss	No Loss	No Loss	0.081	$81,000
5	No Loss	Loss	Loss	0.009	$9,000
6	Loss	Loss	No Loss	0.009	$9,000
7	Loss	No Loss	Loss	0.009	$9,000
8	Loss	Loss	Loss	0.001	$1,000

So the insurance company's expected loss is $271,000. If they charge each customer $100,000 for a $1million dollar policy, they have a 70% chance of collecting $29,000 for doing absolutely nothing. Now there is a chance they will have to make three payouts, but it is small – and because insurance companies never lose, they would either charge enough premiums to assure they don't lose, increase the customer base, and/or misstate that actual risk. The first two are accepted business approaches; the last approach would be considered unethical and probably is illegal.

In engineering we use these same concepts to evaluate engineering risk. Often the risk determines the size of the engineered solution, which in turn determines cost.

If the engineered solution is an important component of infrastructure (an airport, hospital, highway, fuel depot, refinery, etc.) then society either directly or indirectly incurs a large cost to ensure service. Alternatively if the risk involves some kind of failure (plant explosion, chemical release, etc.) then the owner/operator either buys insurance to cover the financial portion of the loss and/or invest in safety procedures and structural controls to change the probability of the failure – in both instances, the business practice is called "risk management" and it is an essential part of modern engineering practice and should not be left solely in the hands of the MBAs and actuaries.

My personal favorite in risk management is the operation of an aircraft carrier during flight operations. The chance for any single failure turning either a single aircraft or the entire ship into a smoking hole in the water is huge; the Navy over years of operations has learned to do these operations in really crummy weather relatively safely (safe enough for a sitting President to "land" an aircraft on the carrier) through the use of technology, and redundant use of human resources. Essentially the system works because of a huge investment in safety protocols to change the "collective probability" of failure to a very small value. While carrier operations are a good example, the Navy doesn't have to make a profit, and can invest in the human resources needed to make the system work.

2. Typical Types of Data Needs/Determinations

Most textbooks discuss the kinds of questions that may be asked with frequency analysis methods, generally many kinds of questions can be posed in terms of probability that often are not. Additionally most texts focus frequency analysis strictly on discharges, but the approach is generic and can be used for many different kinds of variables. The construction of depth-duration-frequency curves is in-fact simply a frequency analysis in two-dimensions (depth-duration).

Data availability may include:

1. long record of the variable of interest at location of interest
2. long record of the variable of interest near the location of interest.
3. short record of the variable of interest at location of interest; short records are fine, but extrapolation to rare events will be suspect.
4. short record of the variable of interest near location of interest.
5. no records near location of interest.

Situations 1-4 can be directly analyzed; situation 5 will require use of regional methods.

Usually frequency analysis is used to produce estimates of

T-year discharges for regulatory or actual flood plain delineation.

T-year; 7-day discharges for water supply, waste load, and pollution severity determination. (Other averaging intervals are also used)

T-year depth-duration-frequency or intensity-duration-frequency for design storms (storms to be put into a rainfall-runoff model to estimate storm caused peak discharges, etc.).

3. Probability Distributions

Most presentations define various analytical probability distributions, usually in the density functional form. Oddly enough the distributions are more useful in cumulative form, but are often not presented in cumulative form.

a. Normal distribution

$$pdf(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}) \text{ [Normal Density]}$$

$$cdf(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(t-\mu)^2}{2\sigma^2})dt = \frac{1}{2}(1 + erf(\frac{x-\mu}{\sigma\sqrt{2}})) \text{ [Normal Cumulative Distribution]}$$

EXCEL has these two functions built-in.

The normal distribution is commonly used in hydrology for variables that can take any real values (negative and positive).

b. Gamma distribution

$$pdf(x) = \frac{\lambda}{\Gamma(n)}(\lambda x)^{n-1} \exp(-\lambda x) \text{ [Gamma Density]}$$

$$cdf(x) = \int_{0}^{x} \frac{\lambda}{\Gamma(n)}(\lambda t)^{n-1} \exp(-\lambda t)dt \text{ [Gamma Cumulative Distribution]}$$

The CDF is usually called the incomplete Gamma distribution and is evaluated numerically. The special case of n=1 is called the Exponential distribution. Γ(n) satisfies the special recursion relationship Γ(t+1)=tΓ(t). Gamma distribution is lower bounded by zero. EXCEL has Gamma functions and distributions built-in.

Curvilinear unit hydrographs are structurally (the arithmetic is the same) and gamma-family distributions.

c. Lognormal

$$pdf(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp(\frac{(\ln(x) - \mu)}{2\sigma^2})$$ [Log Normal Density]

Log-normal is undefined at zero.

d. Extreme value (Gumbel)

$$pdf(x) = \frac{1}{\beta} \exp(\frac{-(x-\alpha)}{\beta} - \exp(\frac{-(x-\alpha)}{\beta}))$$

$$cdf(x) = \exp(-\exp(\frac{-(x-\alpha)}{\beta}))$$

This distribution is used as a limiting distribution in extreme-value statistics. This distribution is used for modeling very rare events.

e. LaPlace distribution

$$pdf(x) = \frac{1}{2\beta} \exp(\frac{-(|x-\alpha|)}{\beta})$$

Also called the double exponential distribution.

f. Weibull distribution

$$pdf(x) = abx^{b-1} \exp(-ax^b)$$

For b=1 this distribution is an exponential distribution.

g. Other distributions.

Logistic, Pareto, Pearsonian etc.

h. Discussion.

The more common distributions in hydrology are the normal, lognormal, and the generalized Gamma family (Wiebull, Gumbel, etc.).

The one Pearsonian distribution in common use is the Log-Pearson Type III. Much of frequency analysis is aimed as using small samples of observations to infer an underlying distribution and make decisions from this distribution.

If the sample is IID and ranked, then the empirical sample CDF will represent relative frequencies that are assumed to be associated with the actual probabilities of the governing distribution.

The classical approach to “fit” to observations is based on relationships of sample means, variances and skew to population means, variances, and skews for different distributions for extrapolation of behavior.

An alternative approach is to simply fit the distribution to the data then extrapolate.

4. Procedures

Plotting positions

This is the name given to the way data are sorted and relative frequency is assigned for each data element. The formulas for plotting positions vary depending on the underlying type of probability they are supposed to explain, and for large numbers of data, they asymptotically approach each other.

The steps to follow regardless of plotting formula are:

1. rank data from small to large magnitude. (This ordering is non-exceedence; reverse order is exceedence)
2. compute the plotting position (really just a relative frequency of occurance) by selected formula. p is the "position" or relative frequency.
3. plot the observation on probability paper (some graphics packages have probability scales; Excel does not, DPlot Jr. an add-in has probability scales).

The plot does not necessarily have to be on probability scales, I find that I can the magnitude versus probability on a log-log or log-linear scale and get the same kind of information I need.

Some plotting (frequency) formulas in common use:

All use: p = relative frequency (probability);
n=number of observations;
m=rank order (1 = small, 2= larger, 3= larger …)

Weibull

$$p = \frac{m}{n+1}$$

California

$$p = \frac{m}{n}$$

Hazen

$$p = \frac{2m-1}{2n}$$

National Environment Research Council (UK)

$$p = \frac{m-0.44}{n+0.12}$$

Example of plotting position

As an example consider the data for Bear Grass Creek in somewhere.

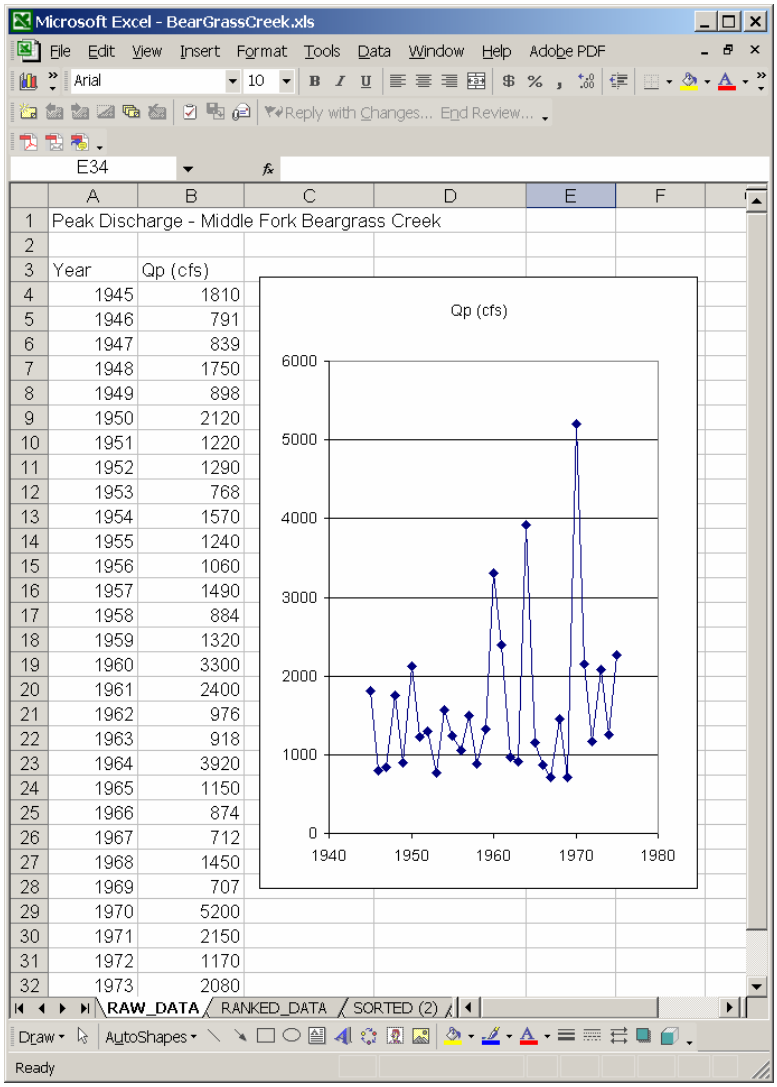
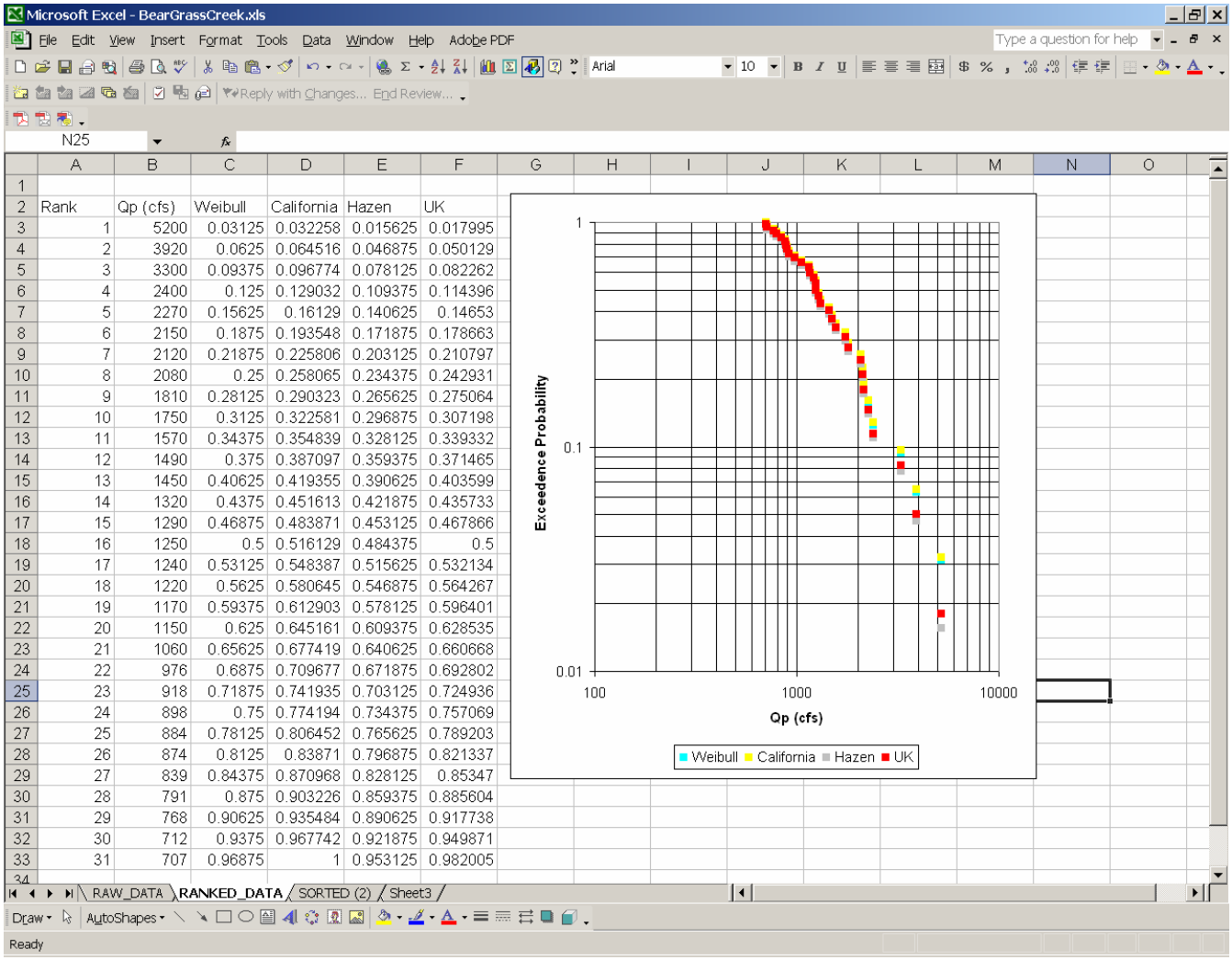


Figure 1

As arranged above the data are in serial order, each entry represents the peak flow for the year, but we have no way of knowing or expecting these peak to occur exactly 1 year apart. We do see that with the exception of three years most of the peaks are somewhere between 1000 and 2000 cfs.

If we believe that serial correlations are minimal then we can assume the data are independent and treat this as a random variable.

To analyze by some distribution the conventional approach is to reorder the data in a rank order and use the rank to assign a relative frequency (incremental probability). The next figure is a decreasing rank order thus the plot is an exceedance probability plot.



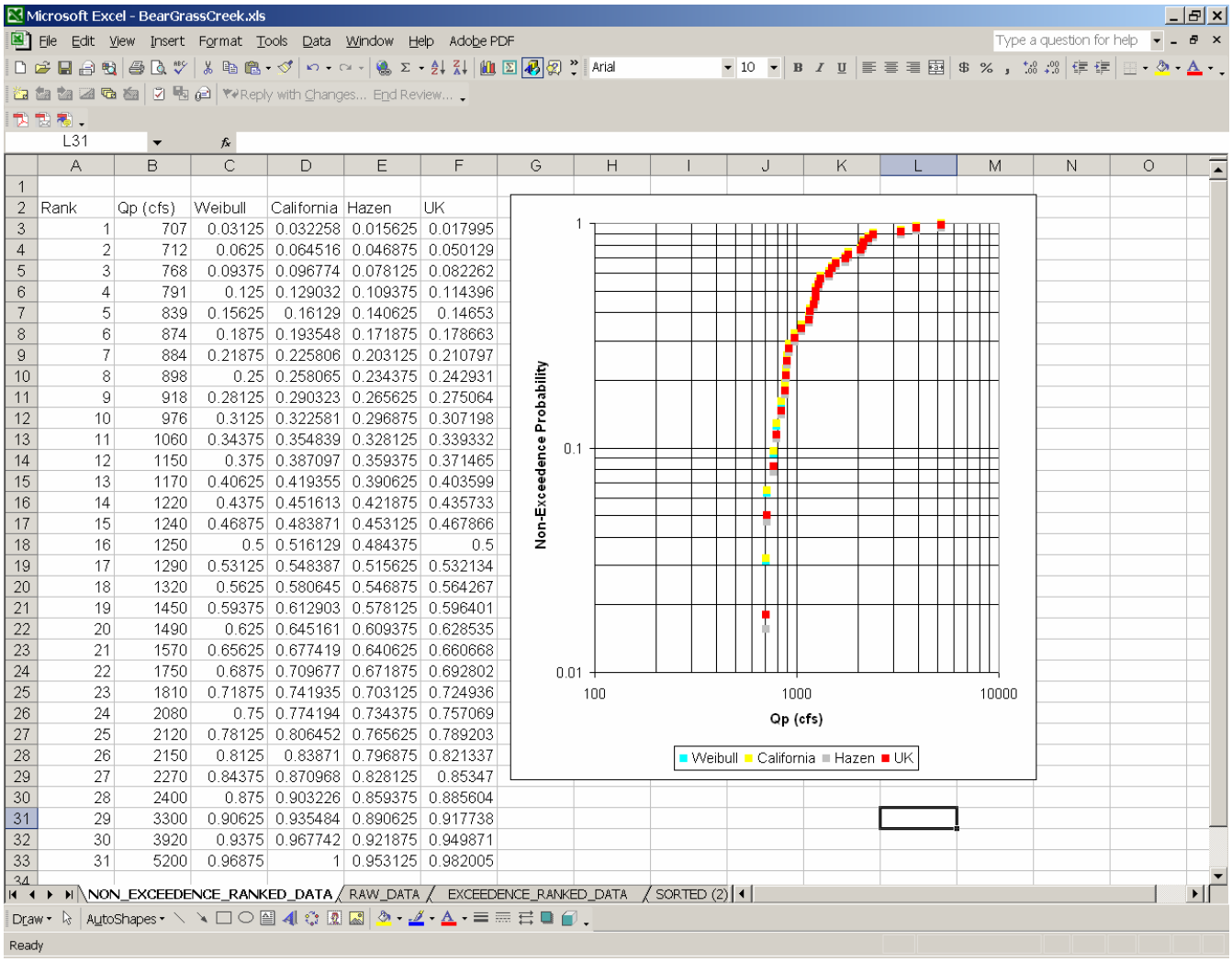
Rank	Qp (cfs)	Weibull	California	Hazen	UK
1	5200	0.03125	0.032258	0.015625	0.017995
2	3920	0.0625	0.064516	0.046875	0.050129
3	3300	0.09375	0.096774	0.078125	0.082262
4	2400	0.125	0.129032	0.109375	0.114396
5	2270	0.15625	0.16129	0.140625	0.14653
6	2150	0.1875	0.193548	0.171875	0.178663
7	2120	0.21875	0.225806	0.203125	0.210797
8	2080	0.25	0.258065	0.234375	0.242931
9	1810	0.28125	0.290323	0.265625	0.275064
10	1750	0.3125	0.322581	0.296875	0.307198
11	1570	0.34375	0.354839	0.328125	0.339332
12	1490	0.375	0.387097	0.359375	0.371465
13	1450	0.40625	0.419355	0.390625	0.403599
14	1320	0.4375	0.451613	0.421875	0.435733
15	1290	0.46875	0.483871	0.453125	0.467866
16	1250	0.5	0.516129	0.484375	0.5
17	1240	0.53125	0.548387	0.515625	0.532134
18	1220	0.5625	0.580645	0.546875	0.564267
19	1170	0.59375	0.612903	0.578125	0.596401
20	1150	0.625	0.645161	0.609375	0.628535
21	1060	0.65625	0.677419	0.640625	0.660668
22	976	0.6875	0.709677	0.671875	0.692802
23	918	0.71875	0.741935	0.703125	0.724936
24	898	0.75	0.774194	0.734375	0.757069
25	884	0.78125	0.806452	0.765625	0.789203
26	874	0.8125	0.83871	0.796875	0.821337
27	839	0.84375	0.870968	0.828125	0.85347
28	791	0.875	0.903226	0.859375	0.885604
29	768	0.90625	0.935484	0.890625	0.917738
30	712	0.9375	0.967742	0.921875	0.949871
31	707	0.96875	1	0.953125	0.982005

Figure 2

Notice the different plotting positions produce about the same plot except at the extremes. The way one can interpret the plot is the following probability statement.

The probability of observing a discharge equal to or greater than 5200 cfs is small (about 3% by Weibull position), while the probability of observing a discharge equal to or greater than 707 cfs is large (about 97% by Weibull position).

Typically we tend to like things to accumulate in increasing order so the next plot is probably the most conventional approach. In this plot the sense of the inequality is reversed.



Rank	Qp (cfs)	Weibull	California	Hazen	UK
1	707	0.03125	0.032258	0.015625	0.017995
2	712	0.0625	0.064516	0.046875	0.050129
3	768	0.09375	0.096774	0.078125	0.082262
4	791	0.125	0.129032	0.109375	0.114396
5	839	0.15625	0.16129	0.140625	0.14653
6	874	0.1875	0.193548	0.171875	0.178663
7	884	0.21875	0.225806	0.203125	0.210797
8	898	0.25	0.258065	0.234375	0.242931
9	918	0.28125	0.290323	0.265625	0.275064
10	976	0.3125	0.322581	0.296875	0.307198
11	1060	0.34375	0.354839	0.328125	0.339332
12	1150	0.375	0.387097	0.359375	0.371465
13	1170	0.40625	0.419355	0.390625	0.403599
14	1220	0.4375	0.451613	0.421875	0.435733
15	1240	0.46875	0.483871	0.453125	0.467866
16	1250	0.5	0.516129	0.484375	0.5
17	1290	0.53125	0.548387	0.515625	0.532134
18	1320	0.5625	0.580645	0.546875	0.564267
19	1450	0.59375	0.612903	0.578125	0.596401
20	1490	0.625	0.645161	0.609375	0.628535
21	1570	0.65625	0.677419	0.640625	0.660668
22	1750	0.6875	0.709677	0.671875	0.692802
23	1810	0.71875	0.741935	0.703125	0.724936
24	2080	0.75	0.774194	0.734375	0.757069
25	2120	0.78125	0.806452	0.765625	0.789203
26	2150	0.8125	0.83871	0.796875	0.821337
27	2270	0.84375	0.870968	0.828125	0.85347
28	2400	0.875	0.903226	0.859375	0.885604
29	3300	0.90625	0.935484	0.890625	0.917738
30	3920	0.9375	0.967742	0.921875	0.949871
31	5200	0.96875	1	0.953125	0.982005

Figure 3

Thus the plot suggests that the probability of observing a discharge equal to or smaller than 5200 cfs is large (about 97%, Weibull formula), while the probability of observing a discharge equal to or smaller than 707 cfs is small (about 3%, Weibull formula).

The curve of cumulative frequency (plotting position) versus magnitude is called the empirical cumulative distribution function.

Typically the next step in analysis is to determine an appropriate distribution model that explains the empirical cumulative distribution function, and use this distribution to make probability statements using methods in statistics. Once the empirical relationship is replaced with an "equivalent" distribution then one can extrapolate to very common or very rare events (large and small probabilities). For example the 0.002 probability event for Bear Grass creek is not on the chart; so we cannot realistically "guess" the magnitude, but if a distribution model is selected that passes through the data then we could from that model guess the magnitude of the 0.002 chance event (either in exceedence or non-exceedence, whichever we are interested in). Before doing this with the Bear Creek data, we cumulative distributions.

5. Cumulative Distribution Functions (CDFs).

Some of the pdf's can be directly integrated, usually using integration by parts. The resulting integrals, if they correspond to known functions are evaluated by table look-up or numerical approximation.

CDF for Normal Distribution – Analytical Result

Suppose you wanted to know the cumulative distribution function for the normal distribution.

Simply apply the CDF equation and you can plot CDF versus magnitudes (values) of the random variable of interest.

For instance the CDF for a normally distributed variable with mean = 50 and standard deviation = 15 is plotted in the next figure.

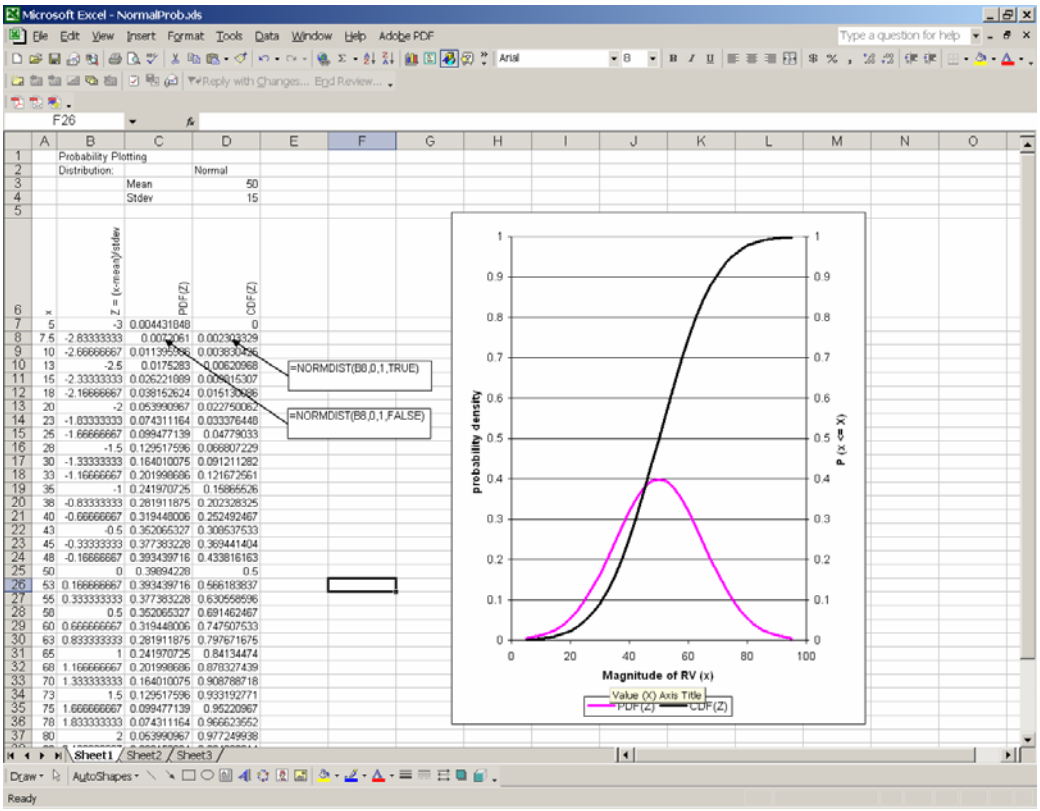


Figure 4

The "equations" in EXCEL are
=NORMDIST(B7,0,1,FALSE) (This function produces the PDF value for entries in column B)

=NORMDIST(B7,0,1,TRUE) (This function accumulates the PDF and produces the CDF value for entries in column B).

Alternatively one could simply apply the equations for the normal distribution function and you would obtain the same results.

Reference: NormProb.xls

CDF for Gamma Distribution – Analytical Result

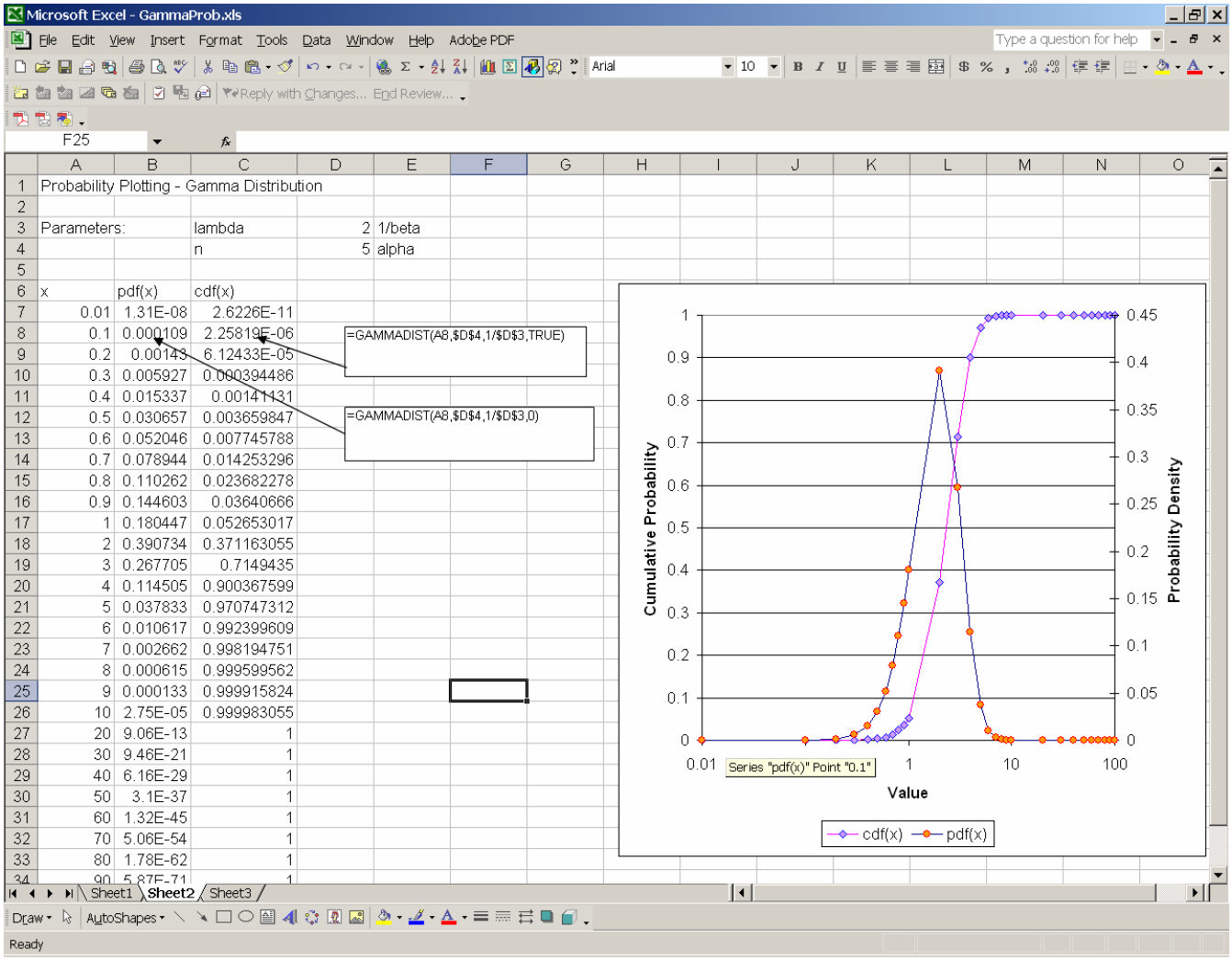
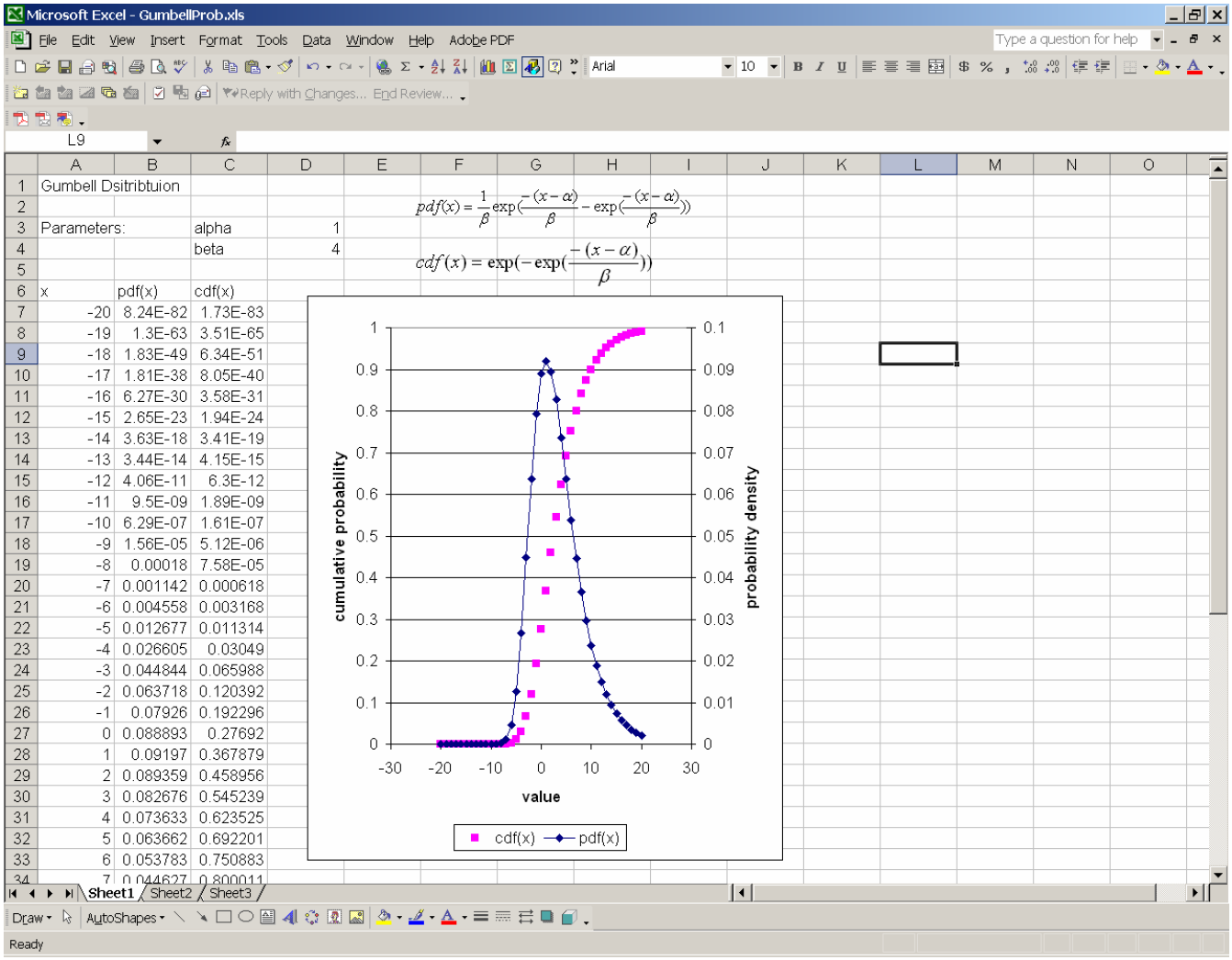


Figure 5

Similar for the Gamma distribution. For the special case of N=1, the distribution is called the exponential distribution and the pdf and cdf are compliments of each other (i.e. they sum to one).

Reference: GammaProb.xls

CDF for Gumbel Distribution – Analytical Result



x	pdf(x)	cdf(x)
-20	8.24E-82	1.73E-83
-19	1.3E-63	3.51E-65
-18	1.83E-49	6.34E-51
-17	1.81E-38	8.05E-40
-16	6.27E-30	3.58E-31
-15	2.65E-23	1.94E-24
-14	3.63E-18	3.41E-19
-13	3.44E-14	4.15E-15
-12	4.06E-11	6.3E-12
-11	9.5E-09	1.89E-09
-10	6.29E-07	1.61E-07
-9	1.56E-05	5.12E-06
-8	0.00018	7.58E-05
-7	0.001142	0.000618
-6	0.004558	0.003168
-5	0.012677	0.011314
-4	0.026605	0.03049
-3	0.044844	0.065988
-2	0.063718	0.120392
-1	0.07926	0.192296
0	0.088893	0.27692
1	0.09197	0.367879
2	0.089359	0.458956
3	0.082676	0.545239
4	0.073633	0.623525
5	0.063662	0.692201
6	0.053783	0.750883
7	0.044627	0.800011

This worksheet is direct computation of the density and cumulative equations.

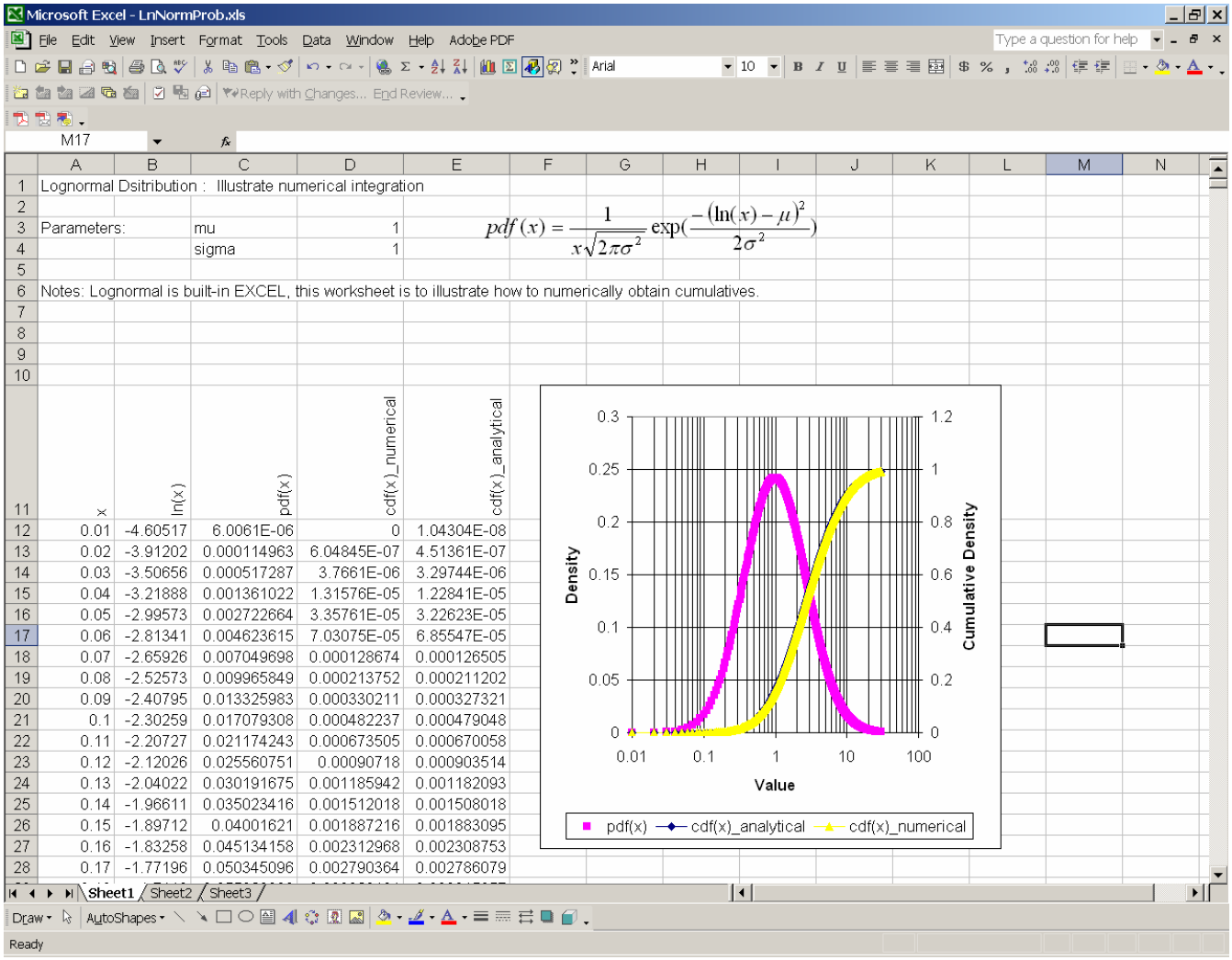
Reference: GumbellProb.xls

b. Numerical integration to construct CDFs

Numerical integration is very effective for constructing CDFs. Simpson's type rules with uniform panels is usually adequate, but the analyst needs to remember to use a lot of panels.

The pdf's evaluate quickly, so there is little need to worry about computation speed.

CDF for Log-Normal Distribution – Numerical Result



In this example the cumulative density is from:

CDF(x) = 0.5*(PDF(x+Dx)+PDF(x))*Dx

(Trapezoidal rule – need to use a lot of panels).

Reference: LnNormProb.xls

c. Probability plotting without probability paper

Historically, one would use probability paper to analyze data to extract the distribution parameters.

Today, one will most likely use a computer program to do the work and use the computer to plot the data.

If the graphics program has probability plotting built-in, by all means use it, otherwise logarithmic plotting often is helpful. Several programs have probability scales built-in (to my knowledge EXCEL does not!).

D-PlotJr. (free download), and D-PLot ($48) have Excel add-ins that allow plotting on probability scales. D-Plot is an excellent piece of software and well worth the $48.

If you have a FORTRAN compiler, and know how to use dynamic link libraries, then the free download is a viable option (you have to write code to generate plots, but you get complete control).

GRAPHER (Golden Software also has probability scales and is relatively cheap).

Example:

Estimate the 20-year peak flow based on an extreme value Type-1 distribution for the annual peak flow data below:

20,45,13,80,12,30,18,22,17,32,22

First the relevant equations:

EV-1 (Gumbel)

$$cdf(x) = \exp(-\exp(\frac{-(x-\alpha)}{\beta}))$$

Next,

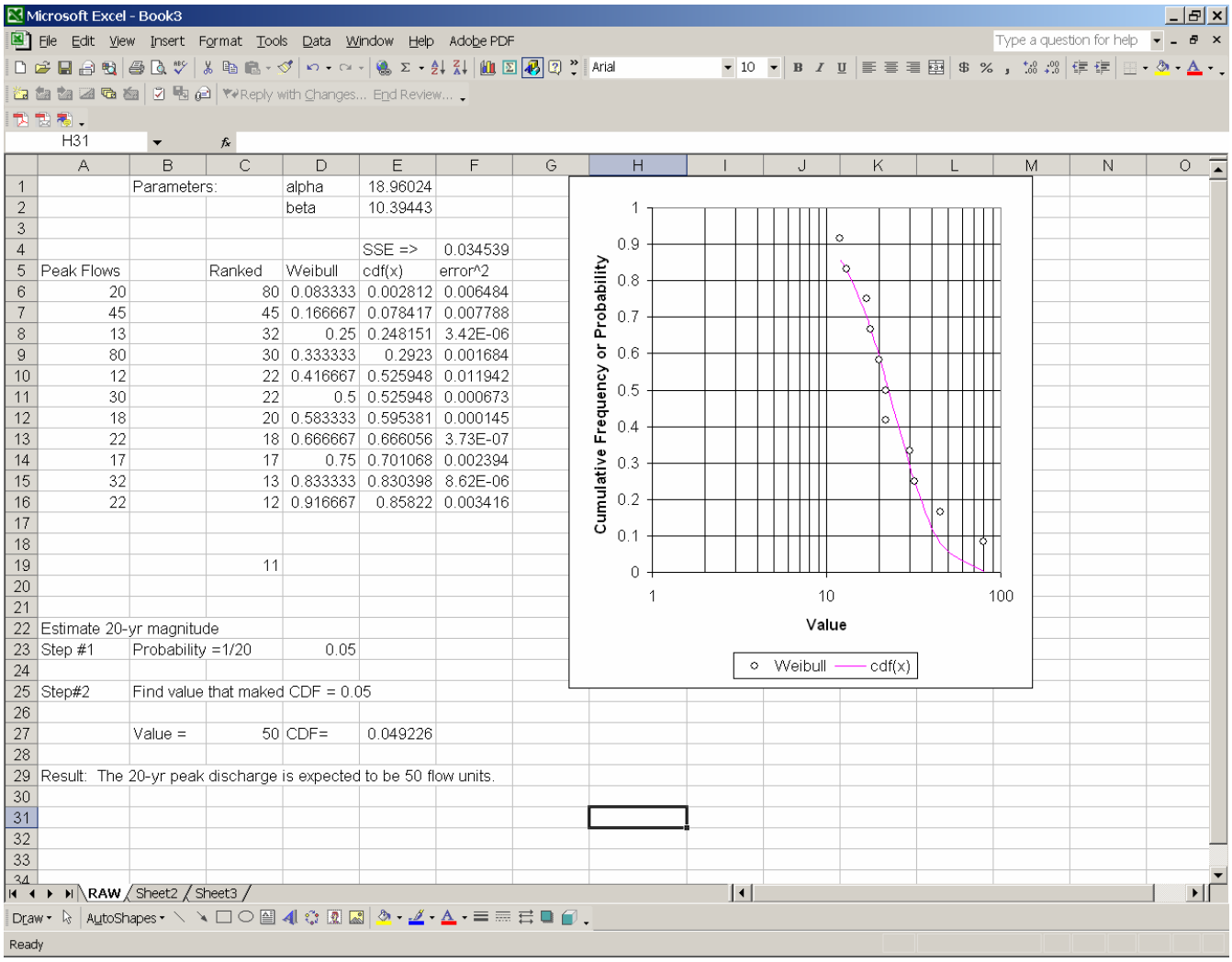
Rank the data into exceedence or non-exceedence structure.

Choose a plotting formula, I used Weibull in the example.

Then by trial-and-error or using some systematic method "fit" the Gumbell CDF to the plotting position values using the ranked magnitudes as the "x" values in the distribution. You will be changing the parameters to accomplish this.

Once you have distribution parameters, reading from the plots to find the 20-year value. While reading from the plots is reasonably straight forward, the value of the exercise is to use the distributions just fitted to interpolate/extrapolate for us. Simply evaluate the distribution to find the magnitude that produces the desired probability value. This example is the "essence" of probability estimation modeling – it is no more complicated than this example, although for unusual data and distributions the effort can be overwhelming.

Result is in figure below.



	A	B	C	D	E	F
1		Parameters:		alpha	18.96024	
2				beta	10.39443	
3						
4					SSE =>	0.034539
5	Peak Flows		Ranked	Weibull	cdf(x)	error^2
6	20		80	0.083333	0.002812	0.006484
7	45		45	0.166667	0.078417	0.007788
8	13		32	0.25	0.248151	3.42E-06
9	80		30	0.333333	0.2923	0.001684
10	12		22	0.416667	0.525948	0.011942
11	30		22	0.5	0.525948	0.000673
12	18		20	0.583333	0.595381	0.000145
13	22		18	0.666667	0.666056	3.73E-07
14	17		17	0.75	0.701068	0.002394
15	32		13	0.833333	0.830398	8.62E-06
16	22		12	0.916667	0.85822	0.003416
17						
18						
19			11			
20						
21						
22	Estimate 20-yr magnitude					
23	Step #1	Probability =1/20		0.05		
24						
25	Step#2	Find value that maked CDF = 0.05				
26						
27		Value =	50	CDF=	0.049226	
28						
29	Result: The 20-yr peak discharge is expected to be 50 flow units.					

Reference: GumbellExample.xls

Finally interpret the 20-year value correctly as the magnitude of the 20-year discharge is 50 flow units. That is we expect the discharge to be 50 units or more with probability 0.05.

The classical approach based on sample moments for this distribution is

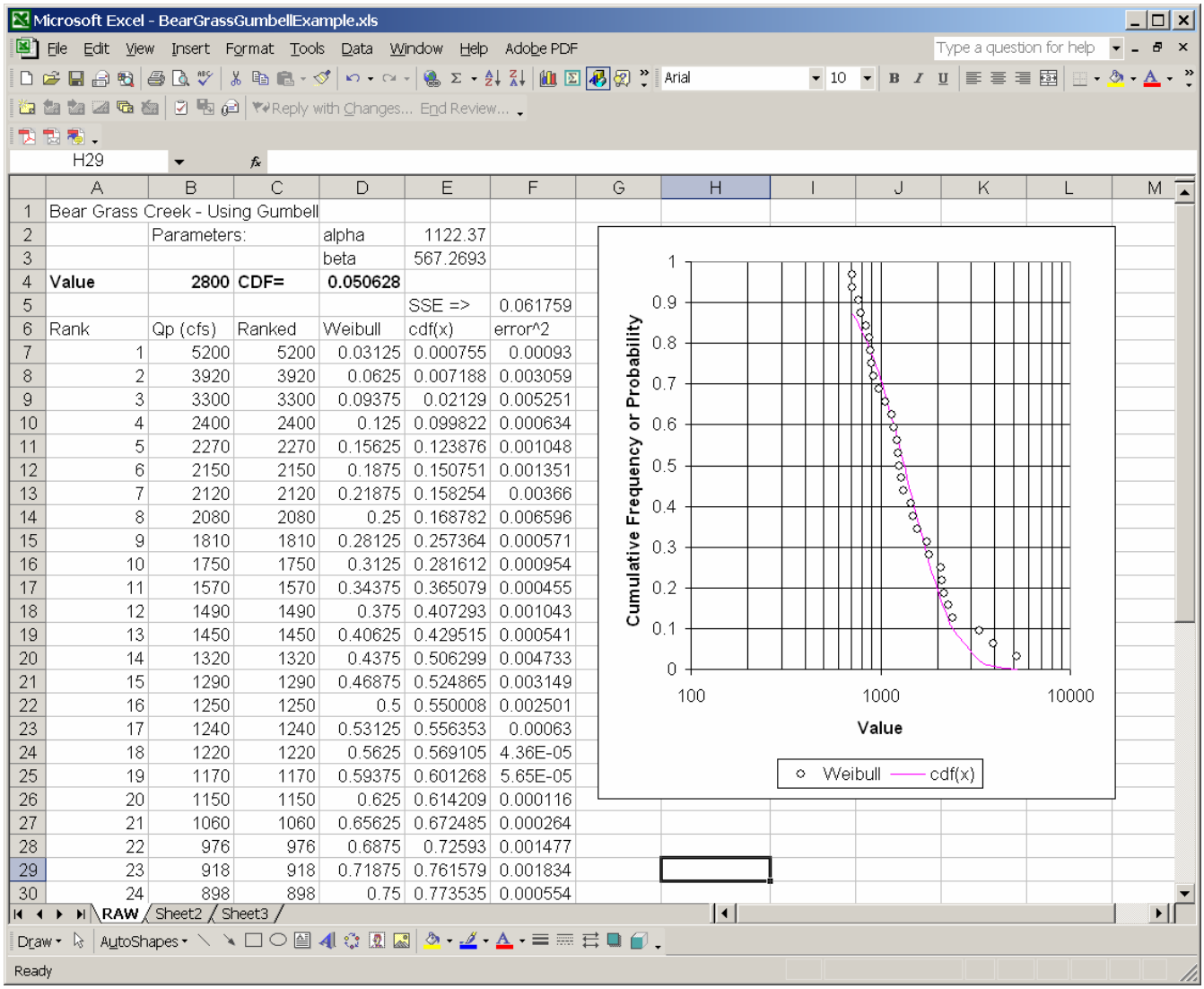
$$\alpha = \frac{\mu\sqrt{6}}{\pi}; \beta = 0.45\sigma$$

Which produces the following results for alpha, beta, and Q_{20} as

22, 8.8, and 49.7 (essentially the same result).

Now lets repeat the analysis for the Bear Grass Creek Flows and find the 5% chance event. Set up the data set in same fashion. Rank and choose plotting formula. Then "fit" the distribution.

Finally use the fitted distribution to find Q with CDF=0.05. In the case of the Bear Creek data, the value is 2800 cfs. That is a discharge of 2800 cfs or larger will be observed on average with probability 0.05 (or 5%); 95% of the time the discharges will be smaller than this value.



	A	B	C	D	E	F
1	Bear Grass Creek - Using Gumbell					
2		Parameters:		alpha	1122.37	
3				beta	567.2693	
4	**Value**	**2800**	**CDF=**	**0.050628**		
5					SSE =>	0.061759
6	Rank	Qp (cfs)	Ranked	Weibull	cdf(x)	error^2
7	1	5200	5200	0.03125	0.000755	0.00093
8	2	3920	3920	0.0625	0.007188	0.003059
9	3	3300	3300	0.09375	0.02129	0.005251
10	4	2400	2400	0.125	0.099822	0.000634
11	5	2270	2270	0.15625	0.123876	0.001048
12	6	2150	2150	0.1875	0.150751	0.001351
13	7	2120	2120	0.21875	0.158254	0.00366
14	8	2080	2080	0.25	0.168782	0.006596
15	9	1810	1810	0.28125	0.257364	0.000571
16	10	1750	1750	0.3125	0.281612	0.000954
17	11	1570	1570	0.34375	0.365079	0.000455
18	12	1490	1490	0.375	0.407293	0.001043
19	13	1450	1450	0.40625	0.429515	0.000541
20	14	1320	1320	0.4375	0.506299	0.004733
21	15	1290	1290	0.46875	0.524865	0.003149
22	16	1250	1250	0.5	0.550008	0.002501
23	17	1240	1240	0.53125	0.556353	0.00063
24	18	1220	1220	0.5625	0.569105	4.36E-05
25	19	1170	1170	0.59375	0.601268	5.65E-05
26	20	1150	1150	0.625	0.614209	0.000116
27	21	1060	1060	0.65625	0.672485	0.000264
28	22	976	976	0.6875	0.72593	0.001477
29	23	918	918	0.71875	0.761579	0.001834
30	24	898	898	0.75	0.773535	0.000554

Reference: BearGrassGumbell.xls

In this example using this distribution the 5% chance event is about 2805 cfs.

References

Asquith, W.H., 1998. Depth-Duration Frequency Analysis of Precipitation for Texas. U.S. Geological Survey, WRIR 98-4044. U.S. Geological Survey, Branch of Information Services, Box 25286, Denver CO 80225-0286.

National Weather Service [http://www.nws.noaa.gov/]

Haan, C.T., Barfield, B.J., Hayes, J.C., 1994. Design Hydrology and Sedimentology for Small Catchments. Academic Press, San Diego, 588p. (Chapter 2)

Wurbs and James, 2002. Water Resources Engineering. Prentice-Hall, New Jersey. Chapter 7.