CE 3354 ENGINEERING HYDROLOGY

Lecture 6: Probability Estimation Modeling

OUTLINE

- Probability estimation modeling – background
- Probability distrobutions
- Plotting positions

WHAT IS PROBABILITY ESTIMATION?

- 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.
- Two important "extremes" in engineering:
  - relatively uncommon events (floods, plant explosions, etc.)
  - very common events (routine discharges, etc.)

FREQUENCY ANALYSIS

- Frequency analysis relates 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.
  - "Long enough" is required for independence, that is the values of the variable are statistically independent, otherwise the variables are said to be serial (or auto-) correlated.
- If the time intervals are short, then dealing with a time-series; handled using different tools.

T-YEAR EVENTS

- 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).
- Also called the Annual Recurrence Interval (ARI)
- 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.
- The Annual Exceedence Probability (AEP) is a related concept

$$X-\text{year ARI} = \frac{1yr.}{Xyr.} = 1/X \ \text{AEP}$$

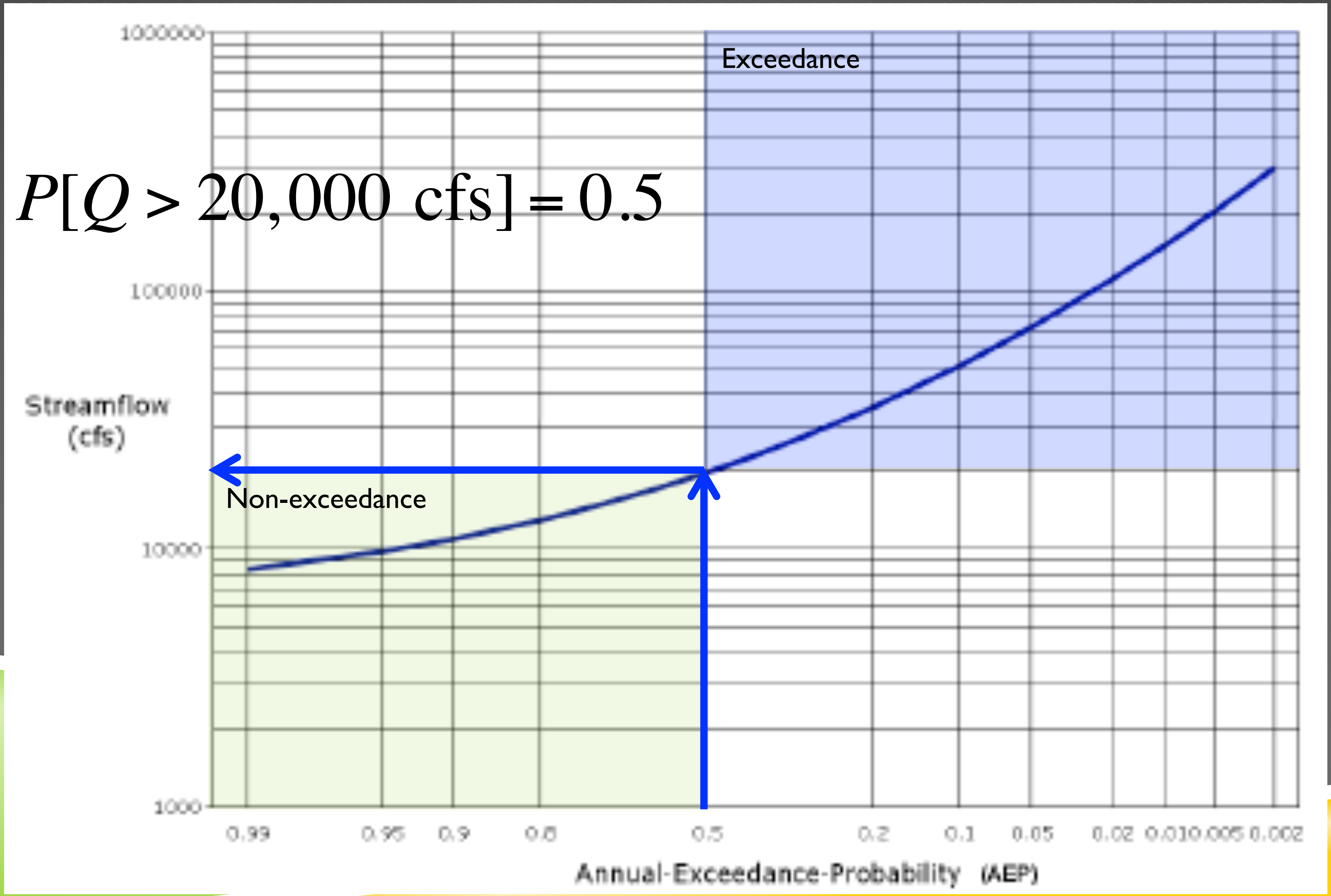
NOTATION

$$P[x>X] = y$$

- 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*"

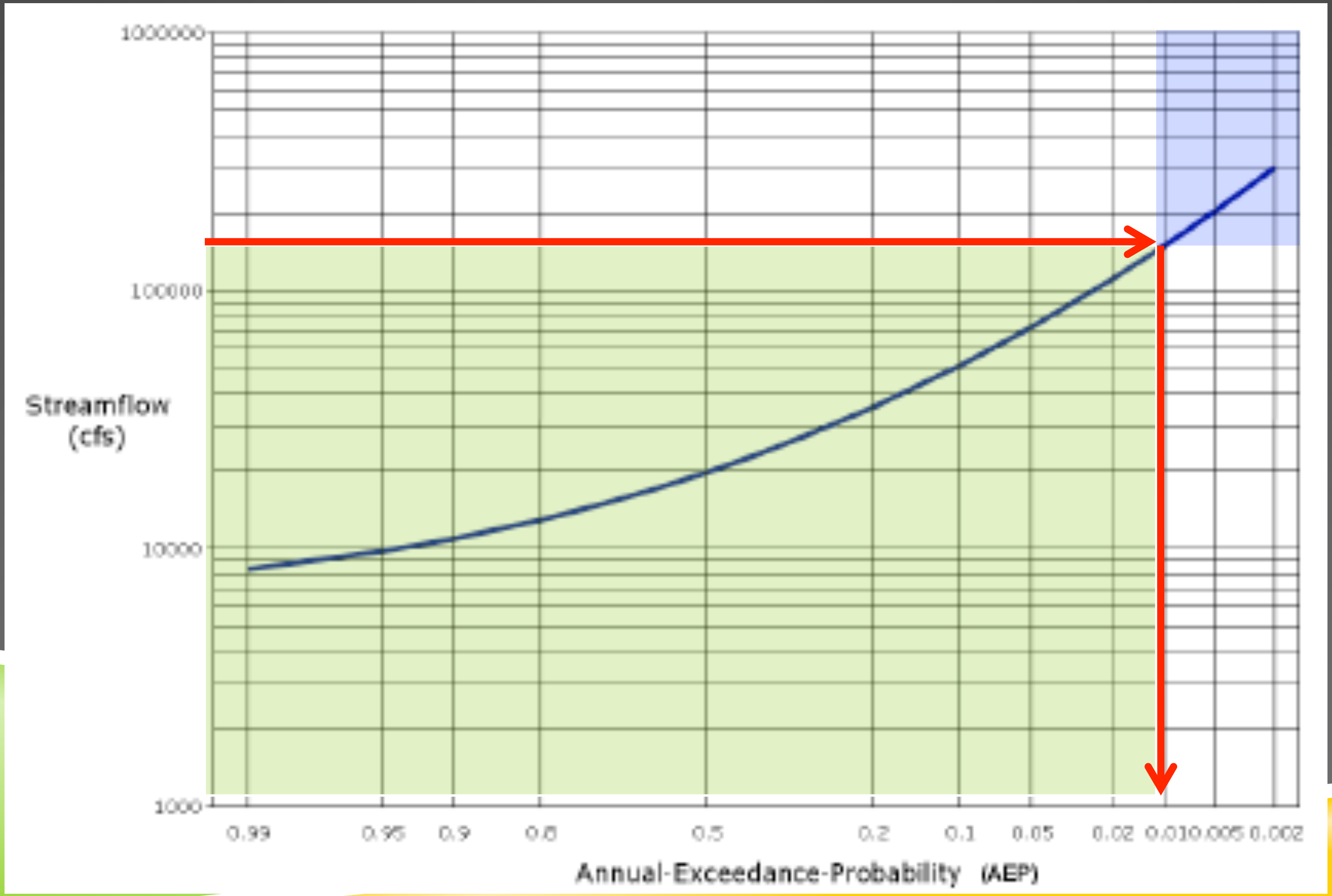
FLOOD FREQUENCY CURVE

- Probability of observing 20,000 cfs or greater in any year is 50% (0.5) (2-year).



FLOOD FREQUENCY CURVE

- Probability of observing 150,000 cfs or greater in any year is ??



PROBABILITY MODELS

- 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 independent, identically distributed then we satisfy the requirements of a Bernoulli process.

PROBABILITY MODELS

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

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

PROBABILITY MODELS

- 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

PROBABILITY MODELS

- 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 number of sampling intervals are increased.

PROBABILITY MODELS

- The probability of the "good" outcomes decreases as the number of 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 a storm).

PROBABILITY MODELS

- 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:
    - ordered with replacement (order matters), samples are replaced
    - unordered with replacement
    - ordered without replacement

USING THE MODELS

- 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.

USING THE MODELS

- They then determine how many identical, independent risks they have to cover to make profit.
- 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.

USING THE MODELS

- If we use the above table, 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

USING THE MODELS

- 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.

DATA NEEDS FOR PROB. ESTIMATES

1. Long record of the variable of interest at location of interest
2. Long record of the variable near the location of interest
3. Short record of the variable at location of interest
4. Short record of the variable near location of interest
5. No records near location of interest

ANALYSIS RESULTS

- 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.).

ANALYSIS RESULTS

- Data are "fit" to a distribution; the distribution is then used to extrapolate behavior

AEP

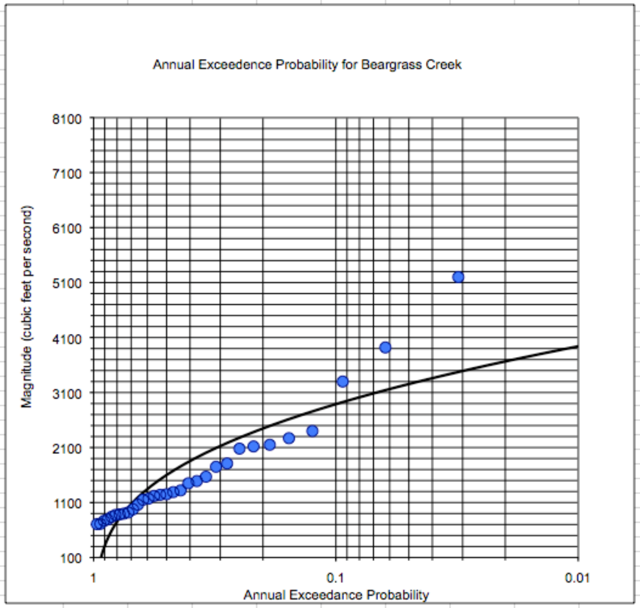
$$F(x) = \frac{1}{2}\left(1 + erf\left(\frac{x-\mu}{2\sigma}\right)\right)$$

Magnitude

Error function
(like a key on a calculator
e.g. log(), ln(), etc.)

Distribution Parameters

	A	B	C	D
1	Probability Plotting			
2	Distribution:		Normal	
3	Mean	1600		
4	Stdev	1006		
5		Y-1		X-1
6	x	Z = (x-mea	PDF(Z)	CDF(Z)
7	100	-1.49105	0.131262	0.932026
8	200	-1.39165	0.151483	0.917986
9	300	-1.29225	0.173099	0.901864
10	400	-1.19284	0.195856	0.883535
11	500	-1.09344	0.219425	0.8629
12	600	-0.99404	0.243414	0.839897
13	700	-0.89463	0.26737	0.814508
14	800	-0.79523	0.290796	0.78676
15	900	-0.69583	0.313165	0.756731
16	1000	-0.59642	0.333939	0.724553
17	1100	-0.49702	0.352589	0.690412
18	1200	-0.39761	0.368621	0.654543
19	1300	-0.29821	0.381592	0.617229
20	1400	-0.19881	0.391136	0.578793
21	1500	-0.0994	0.396976	0.539591
22	1600	0	0.398942	0.5
23	1700	0.099404	0.396976	0.460409
24	1800	0.198807	0.391136	0.421207
25	1900	0.298211	0.381592	0.382771
26	2000	0.397614	0.368621	0.345457
27	2100	0.497018	0.352589	0.309588
28	2200	0.596421	0.333939	0.275447
29	2300	0.695825	0.313165	0.243269
30	2400	0.795229	0.290796	0.21324
31	2500	0.894632	0.26737	0.185492
32	2600	0.994036	0.243414	0.160103
33	2700	1.093439	0.219425	0.1371
34	2800	1.192843	0.195856	0.116465
35	2900	1.292247	0.173099	0.098136
36	3000	1.39165	0.151483	0.082014
37	3100	1.491054	0.131262	0.067974
38	3200	1.590457	0.112622	0.055866
	3300	1.689861	0.095679	0.045527
	3400	1.789264	0.080486	0.036786
	3500	1.888668	0.06704	0.029468
	3600	1.988072	0.055291	0.023402
43	3700	2.087475	0.045152	0.018423
44	3800	2.186879	0.03651	0.014376
45	3900	2.286282	0.029232	0.011119



DISTRIBUTIONS

$$pdf(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

Normal Density

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

Cumulative Normal Distribution

DISTRIBUTIONS

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

Gamma Density

$$cdf(x) = \int_0^x \frac{\lambda}{\Gamma(n)}(\lambda t)^{n-1}\exp(-\lambda t)dt$$

Cumulative Gamma Distribution

DISTRIBUTIONS

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

Extreme Value (Gumbel) Density

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

Cumulative Gumbel Distribution

PLOTTING POSITIONS

PLOTTING POSITIONS

- A plotting position formula estimates the probability value associated with specific observations of a stochastic sample set, based solely on their respective positions within the ranked (ordered) sample set.

Bulletin 17B → Weibull (1939)

Reference	a	Formula
Weibull (1939)	0	$i / (n + 1)$
Blom (1958)	0.375	$(i - 0.375) / (n + 0.25)$
Cunnane (1978)	0.4	$(i - 0.4) / (n + 0.2)$
Gringorten (1963)	0.44	$(i - 0.44) / (n + 0.12)$
Hazen (1914)	0.5	$(i - 0.5) / n$

i is the rank number of an observation in the ordered set, *n* is the number of observations in the sample set

PLOTTING POSITION FORMULAS

- Values assigned by a plotting position formula are solely based on set size and observation position
  - The magnitude of the observation itself has no bearing on the position assigned it other than to generate its position in the sorted series (i.e. its rank)
- Weibull - In common use; Bulletin 17B
- Cunnane – General use
- Blom - Normal Distribution Optimal
- Gringorten - Gumbel Distribution Optimal

PLOTTING POSITION STEPS

1. Rank data from small to large magnitude.
   1. This ordering is non-exceedence
   2. reverse order is exceedence
2. compute the plotting position by selected formula.
   1. *p* is the "position" or relative frequency.
3. plot the observation on probability paper
   1. some graphics packages have probability scales

BEARGRASS CREEK EXAMPLE

- Examine concepts using annual peak discharge values for Beargrass Creek
- Data are on class server

NEXT TIME

- Probability estimation modeling (continued)
- Bulletin 17B