Lecture 24—Floods and flood frequency

One of the things we want to know most about rivers is "what's the probability that a flood of size *x* will happen this year? In 100 years?" There are two ways to do this—empirically, and parametrically.

First, empiricism. Let's take a bunch of data. For now, we'll take flood data—the maximum flood for each year for some number of years:

Year	Flow, cfs
1945	2290
1946	1470
1947	2220
1948	2970
1949	3020
1950	1210
1951	2490
1952	3170
1953	3220
1954	1760
1955	8800
1956	8280
1957	1310
1958	2500
1959	1960
1960	2140
1961	4340
1962	3060
1963	1780
1964	1380
1965	980
1966	1040
1967	1580
1968	3630

To plot this data empirically, we need to order these according to *rank*. That is, the highest flow comes first, and then the next highest, on down.

Year	Flow, cfs	Rank
1955	8800	1
1956	8280	2
1961	4340	3
1968	3630	4
1953	3220	5

1952	3170	6
1962	3060	7
1949	3020	8
1948	2970	9
1958	2500	10
1951	2490	11
1945	2290	12
1947	2220	13
1960	2140	14
1959	1960	15
1963	1780	16
1954	1760	17
1967	1580	18
1946	1470	19
1964	1380	20
1957	1310	21
1950	1210	22
1966	1040	23
1965	980	24

Incidentally, you can get Excel to do this for you. Select the data, then go to Data|Sort, and it will order all the data! From here, use the formula:

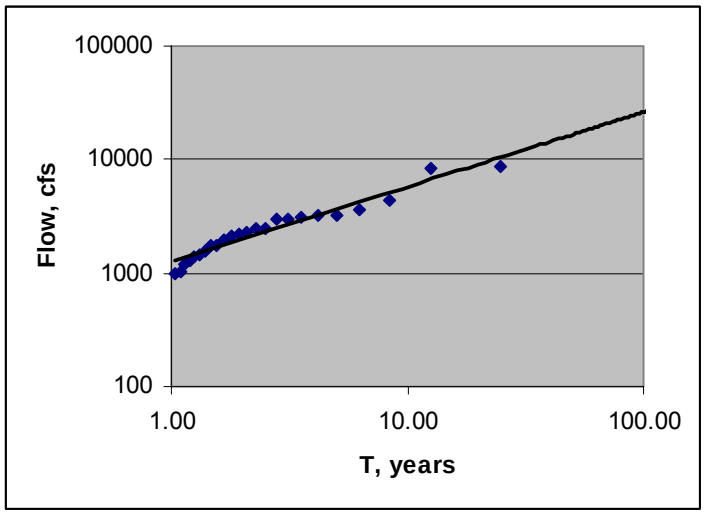
$$T = \frac{n+1}{m}$$

Where *T* is the *recurrence interval*, *n* is the number of years in the record, and *m* is the rank. Thus, for our data:

Year	Flow, cfs	Rank	T
1955	8800	1	25.00
1956	8280	2	12.50
1961	4340	3	8.33
1968	3630	4	6.25
1953	3220	5	5.00
1952	3170	6	4.17
1962	3060	7	3.57
1949	3020	8	3.13
1948	2970	9	2.78
1958	2500	10	2.50
1951	2490	11	2.27
1945	2290	12	2.08
1947	2220	13	1.92
1960	2140	14	1.79
1959	1960	15	1.67
1963	1780	16	1.56
1954	1760	17	1.47

1967	1580	18	1.39
1946	1470	19	1.32
1964	1380	20	1.25
1957	1310	21	1.19
1950	1210	22	1.14
1966	1040	23	1.09
1965	980	24	1.04

All that's left is to plot T on the horizontal axis and Flow on the vertical, and shoot a best-fit line through the whole mess. Oh, and typically we plot it on log-log paper:



By extrapolating the trend line, you can determine the "100-year" or "500-year" flood.

What is actually *meant* by "100-year flood," by the way, is that it has a 1% chance of happening every year, not that it only happens every 100 years. Here's a nifty formula for determining *frequency* or *probability* rather than recurrence.

$$F = 1 - \frac{m}{n+1}$$

In other words, $F = 1 - \frac{1}{T}$.

Here's the basic problem, though. Especially with small data sets, each new data point will significantly alter the rank of all the other points, and therefore change the whole curve. As a result, it's often

easier to perform this analysis *parametrically*, with a few important statistics. Let's talk.

Let's consider the height of every person in the room. The result of this (assuming we have adults and we've got enough people) is an oddly shaped curve. Most everyone fits between about 150 cm and 190 cm, with a pronounced hump around 170 cm or so. However, the distribution tails off to include the Shaq's and the Billy Bartletts of our population. This curve is called many things, and is of vital importance to lots of the natural world. It's called the normal distribution, the bell curve, the Gaussian distribution, or the binomial distribution (after an easy way to create it). Turns out lots of natural distributions look like this—especially if there's some sort of control making an "average." To talk about these distributions, we have a number of *parameters* that describe normal distributions. Here they are.

Average (*aka* 1st moment)—the average value of all the individual values.

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

Variance (*aka* 2nd moment)—the average of the *difference* between each individual and the mean. This is a measure of the spread of the data

$$s^2 = \frac{\sum_{1}^{n} (x - \bar{x})^2}{n-1}$$

where the square is placed to ensure a positive number. To regain the units of the mean (e.g. the units of variance in our height example are cm^2), we have

Standard deviation—which is just the positive square root of variance.

At this point we need to make a little quibble. Whenever you measure a group of objects, like we just did with height, you are taking *samples* from some larger population. Although your sample may approximate the whole population, it may not. Statisticians,

then, draw a distinction between the mythic *population* statistic and the measurable *sample* statistic. Mean, for example, is a population statistic; *average* is a sample statistic. As defined, we have *sample* standard deviation, and we abbreviate it *s*. *Population* standard deviation is defined with *N* in place of *n*-1, and is abbreviated σ. Most of the time, however, geologists blithely ignore this distinction, and refer to average as mean, and use σ for sample standard deviation. It's all good.

These two parameters (mean and standard deviation) suffice to explain the normal curve. The equation for a normal curve is:

$$p(x) = \frac{1}{\sigma} e^{-\left[\frac{(x-\eta)^2}{2\sigma^2}\right]}$$

Oh, right, η is the population mean.

Just as it's difficult to use functions of x^n because they're hard to compare, so are normal distributions hard to look at. As a result of the whole x^n thing we came up with a set of skewed axes (log paper) that make functions of x^n look like straight lines. So it is with the binomial distribution. If you take one and *sum* the components rather than plotting them like a histogram you get something called a "cumulative frequency chart" or *s-curve*. We can *also* make skewed axes that plot s-curves as straight lines. This axis is called a probability axis. Plotting on probability paper makes things easy—normal distributions plot as straight lines, and determining standard deviation is easy—it's the distance on the line from 50% to 84% (or from 50% to 16%).

This explains the normal curve nicely, and it would be nice if that were all there was. But there's more. It turns out that there are a number of curves called *quasi-normal* curves. These involve two other statistics—skewness and kurtosis. We'll talk about these now.

Skewness (*aka* 3rd moment)—this is a measure of "lean" on the curve. Curves that lean to the left are positively skewed, and those that lean to the right are negatively skewed.

$$S_k = \frac{n}{(n-1)(n-2)} \cdot \frac{\sum_1^n (x-\bar{x})^3}{s^3}$$

Notice that this effectively takes the ratio of standard deviation to standard deviation, but allows for a sign (because it's cubed), thus allowing for contributions on one side to outweigh those on the other, and force the parameter to be positive or negative. Note that a normal distribution has a skewness of zero.

A sidebar on median. Median ($\hat{x}$) is like mean in that it shows something about where the bulk of the data lie. It is determined, however, by finding the middle value of the distribution, rather than averaging all values. That is, if we take our heights from lowest to highest, and there's 11 of us in the room, the 6th value counting up (or down) is the median value. Why do we care? In a normal distribution, the mean and median *must* lie at the same point. If the median is to the left of the mean, the bulk of the data is *also* to the left of the mean, and the distribution is positively skewed. There you have it.

Kurtosis (*aka* 4th moment)—this somewhat nebulous statistic says something about how "peaked" the curve is. High values of kurtosis represent curves more peaked than normal, and low values flatter.

$$K = \frac{n}{(n-1)(n-2)(n-3)} \frac{\sum_1^n (x-\bar{x})^4}{s^4}$$

Note that because these parameters are themselves only recombinations of average and standard deviation, effectively these are only variants of the normal curve.

Why do we care about all this? It turns out that there is no particular reason why flood data should be normally distributed, so we may have to use some OTHER distribution. What do I mean by this? Remember that I gave you a mathematical statement of the normal distribution:

$$p(x) = \frac{1}{\sigma} e^{-\left[\frac{(x-\eta)^2}{2\sigma^2}\right]}$$

Meaning that if you're given η and σ you can work out the probability of an event yourself. There are other distributions, though—one of the most popular in flood analysis is the gamma-3 or Pearson 3 distribution:

$$f(x) = \frac{x^{\alpha-1} e^{\frac{-x}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)}$$

although the extreme value distribution (EV1 or Gumbel) is often used as well:

$$f(x) = e^{-e^{\frac{-x-u}{\gamma}}}$$

To use these, simply determine your statistical parameters (namely mean and standard deviation), then convert these to the parameters used in the distributions. Here:

$$\alpha = \frac{\bar{x}^2}{\sigma^2}$$

$$\beta = \frac{\sigma^2}{\bar{x}}$$

$$\gamma = \frac{\sqrt{6}\sigma}{\pi}$$

$$u = \bar{x} - 0.5772\gamma$$

While I'm here, it's worth talking about Gamma 3 and parameters. Most probability distributions (and there are lots) have two or three parameters that are in turn functions of the elementary statistics we talked about. In Gamma 3, α is called a *shape factor*, and β is called a *scale factor*. This is because varying β just stretches the function on the y axis, but changing α changes the shape of the distribution as a whole. {graphs} You could also include a parameter that moves the whole distribution back and forth on the x-axis—that would be a *location parameter*. The important thing is that if you read about some other distribution, you may be able to determine (or be told) what the parameters do.

Enough about this. Back to floods. One quick solution would be to take the average and standard deviation of your flood data (or the log of your flood data) and use these theoretical curves to estimate peak flow instead. For example, using the data above, I got:

$x = 7.756$

$s = 0.566$

So $\alpha = 1.97$, $\beta = 1410$.