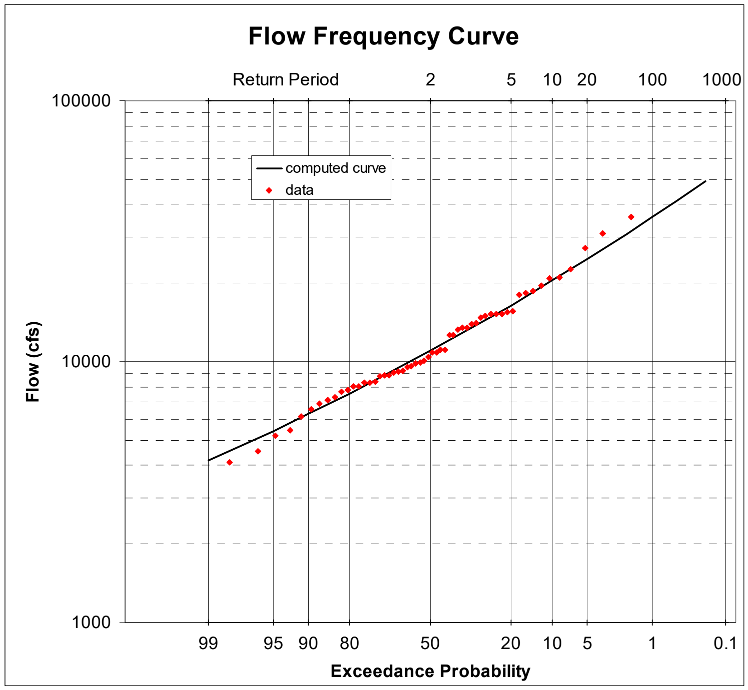
Lecture 3.4

Uncertainty in Frequency Estimates

Hydrologic Engineering Center
Beth Faber, PhD, PE
May 2022

1

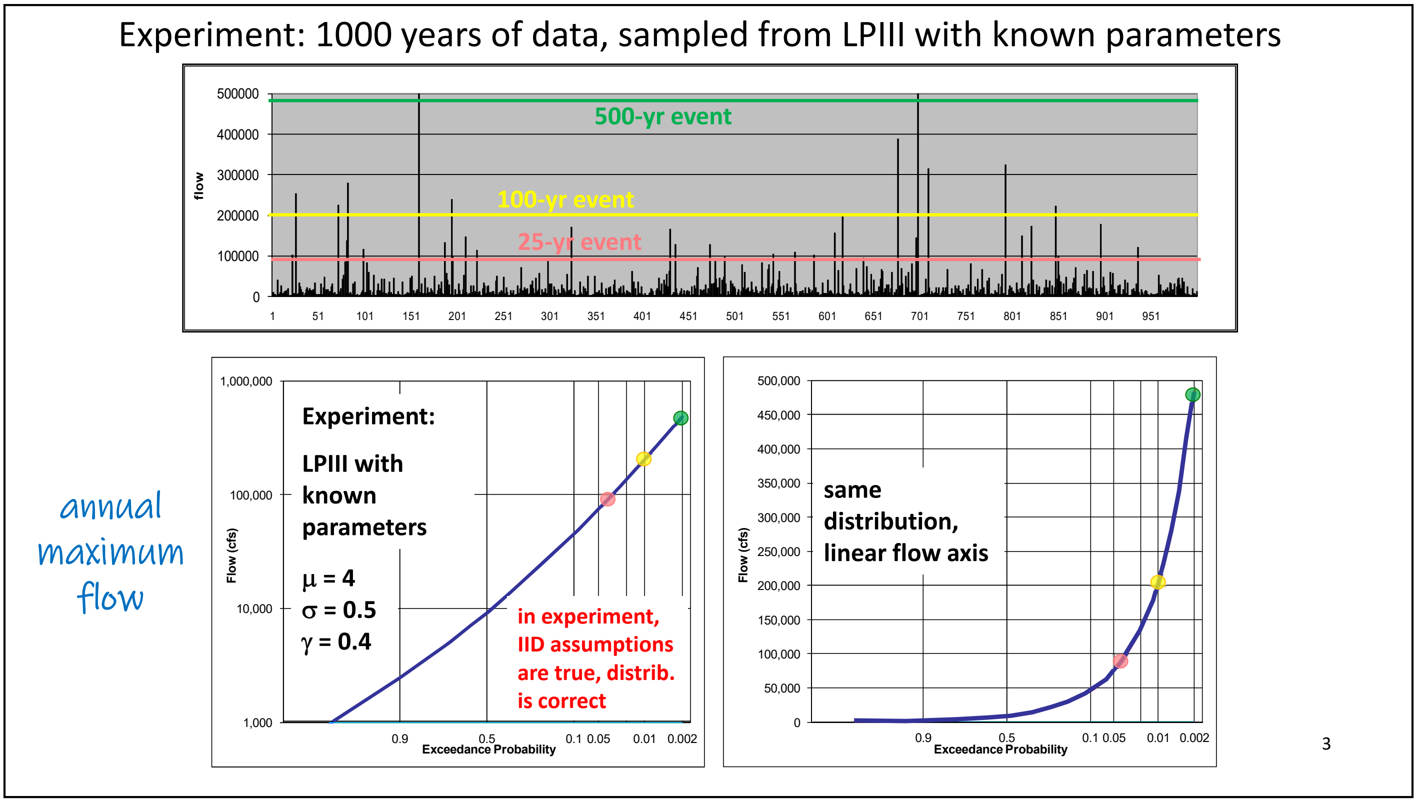
annual
maximum
flow



2

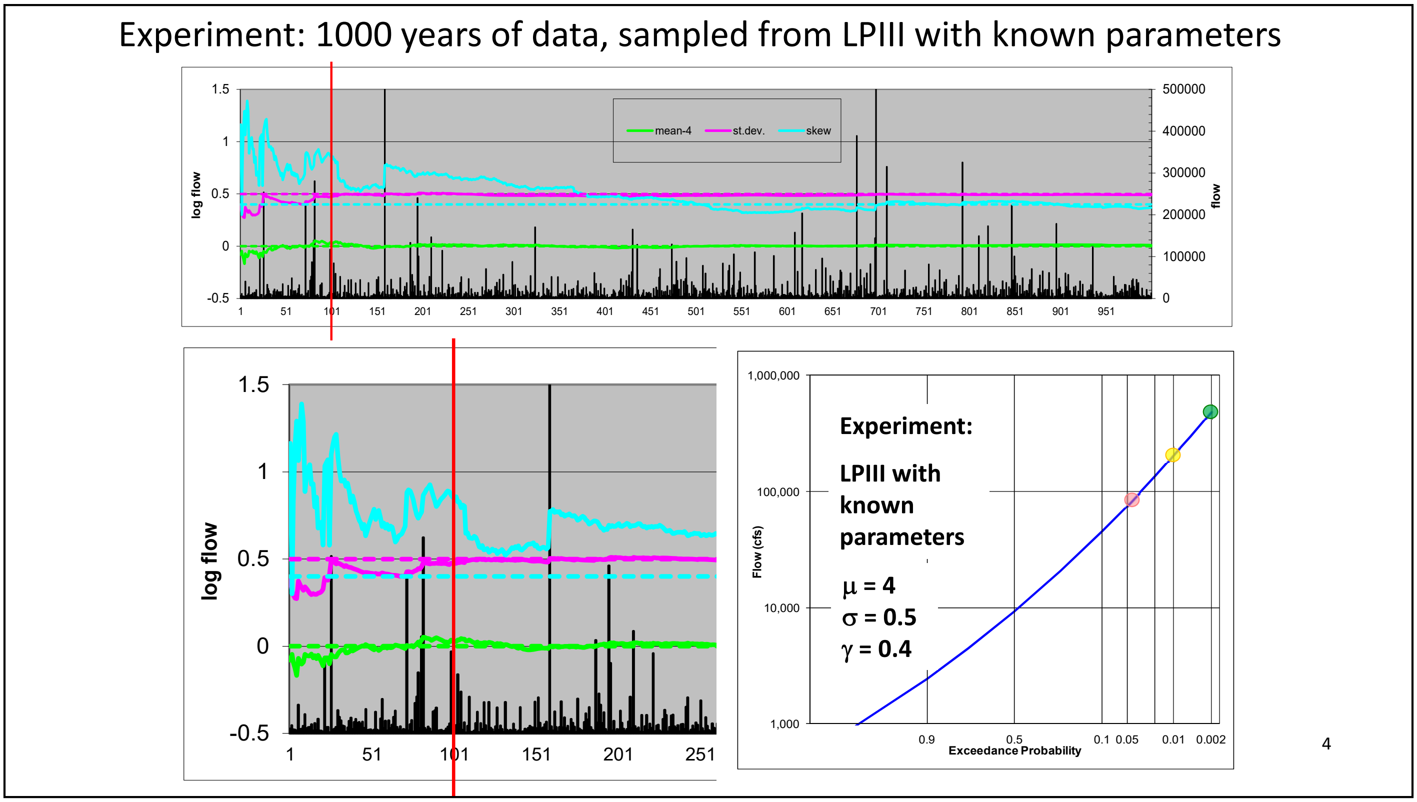
When fitting a flood frequency curve to data, the picture we're accustomed to seeing is the fitted curve and the confidence interval. This lecture explored the uncertainty in this process, both what defines the confidence interval and what exists but does not inform the interval.

56 years



This is a statistical experiment in which we start with a KNOWN probability distribution, and randomly sample many "years" of data from it. We can perform parameter estimation on that sampled data, and see how well we can predict the true values that in this case are known.

This example samples 1000 "years" of peak annual flows from an LP3 distribution with mean=4, standard dev=0.5, and skew=0.4 Note the occurrence of spans of time that are not representative of the full data set (225 to 425 has mostly very low events, and 650 to 700 has very high events.)



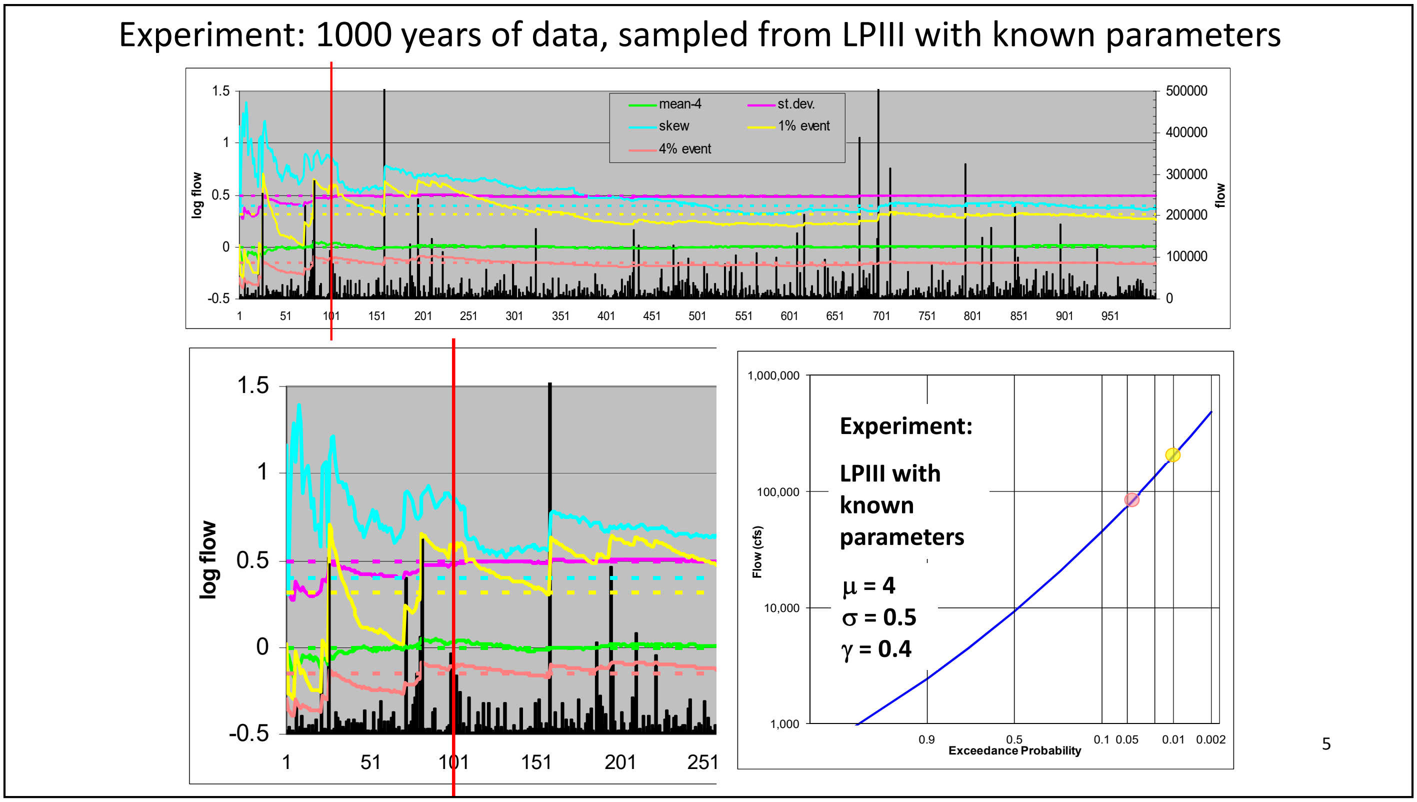
This is a statistical experiment in which we start with a KNOWN probability distribution, and randomly sample many "years" of data from it. We can perform parameter estimation on that sampled data, and see how well we can predict the true values that in this case are known.

This example samples 1000 "years" of peak annual flows from an LP3 distribution with mean=4, standard dev=0.5, and skew=0.4. Parameters are re-estimated each year, with each additional value. Mean and standard deviation are estimated well with 50 to 100 years of data. However, skew does not stabilize until 300 to 400 years of data are available. The 1% (100-yr) and 4% (25-yr) events are also shown, with the 1% event being much more sensitive to the erratic skew value.

Note how the occurrence of a large flood event impacts the skew estimate. Note how a span of years without large events affects the skew estimate.

Note the occurrence of spans of time that are not representative of the full data set (225 to 425 has mostly very low events, and 650 to 700 has very high events.)

This example is meant to display our uncertainty in our frequency curve estimate, even with 100 years of data available to us. The example actually underestimates our real uncertainty because it represents an ideal case in which the distribution truly is a LogPearson III. In reality, the choice of LP3 is a simplification that introduces additional error.



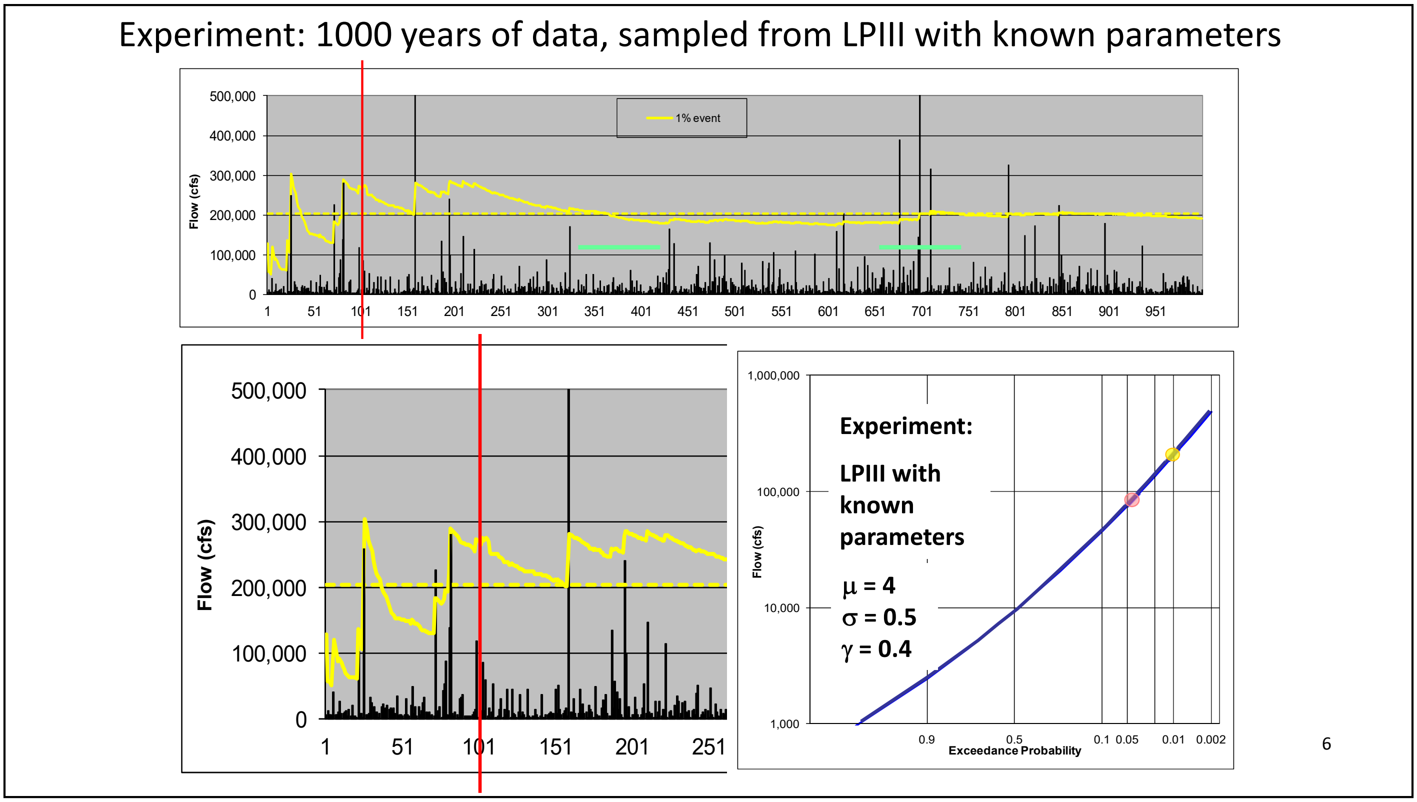
This is a statistical experiment in which we start with a KNOWN probability distribution, and randomly sample many "years" of data from it. We can perform parameter estimation on that sampled data, and see how well we can predict the true values that in this case are known.

This example samples 1000 "years" of peak annual flows from an LP3 distribution with mean=4, standard dev=0.5, and skew=0.4. Parameters are re-estimated each year, with each additional value. Mean and standard deviation are estimated well with 50 to 100 years of data. However, skew does not stabilize until 300 to 400 years of data are available. The 1% (100-yr) and 4% (25-yr) events are also shown, with the 1% event being much more sensitive to the erratic skew value.

Note how the occurrence of a large flood event impacts the skew estimate. Note how a span of years without large events affects the skew estimate.

Note the occurrence of spans of time that are not representative of the full data set (225 to 425 has mostly very low events, and 650 to 700 has very high events.)

This example is meant to display our uncertainty in our frequency curve estimate, even with 100 years of data available to us. The example actually underestimates our real uncertainty because it represents an ideal case in which the distribution truly is a LogPearson III. In reality, the choice of LP3 is a simplification that introduces additional error.



Simplification of the previous slide, which removes all but the 1% event line for a clearer view.

NOTES FROM PREVIOUS SLIDE:

This is a statistical experiment in which we start with a KNOWN probability distribution, and randomly sample many “years” of data from it. We can perform parameter estimation on that sampled data, and see how well we can predict the true values that in this case are known.

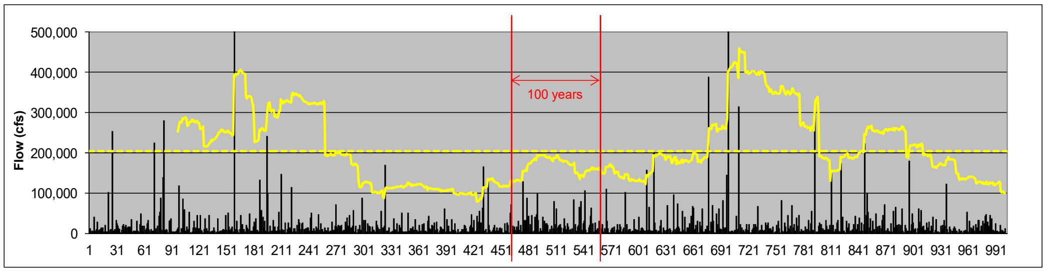
This example samples 1000 “years” of peak annual flows from an LP3 distribution with mean=4, standard dev=0.5, and skew=0.4. Parameters are re-estimated each year, with each additional value, and the 1% value of the LP3 computed. Note how the occurrence of a large flood event impacts the skew estimate as so the 1% value. Note how a span of years without large events affects the skew estimate.

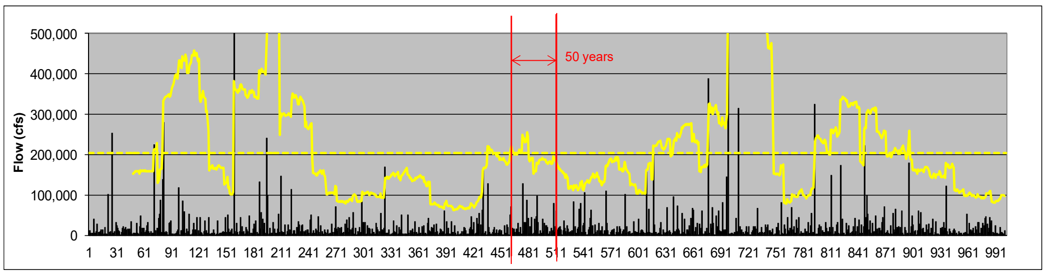
Note the occurrence of spans of time that are not representative of the full data set (225 to 425 has mostly very low events, and 650 to 700 has very high events.)

This example is meant to display our uncertainty in our frequency curve estimate, even with 100 years of data available to us. The example actually underestimates our real uncertainty because it represents an ideal case in which the distribution truly is a LogPearson III. In reality, the choice of LP3 is a simplification that introduces additional error

Sample Size?

Same data, sliding 100 year window and 50 year window



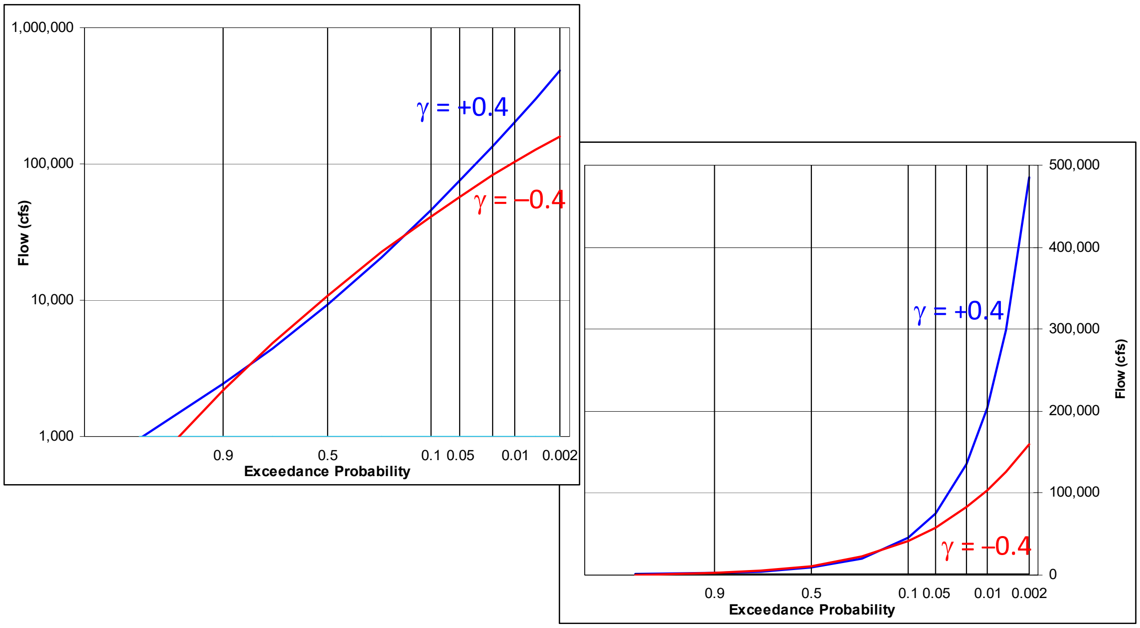


7

In the previous graphs, the effect of the increasing sample size obscures the effect of individual events on the parameter estimates. In these two graphics, a single sample size is maintained, and the "window" is moved across the data set. The plotted value is at the end of the window it represents, so the relevant window is to the left of the value.

Note the greater variability in the estimates using 50 years of record, as compared to using 100 years. Note the still large amount of variability in the 100 year estimates. The impact of individual large events is now clearer, as is the effect of periods without large events. Note the decrease in the skew estimate (and 1% event) between the previously noted low years of 225 to 425, and the rise between 650 and 700, as many large events are included.

What about a more negative skew?



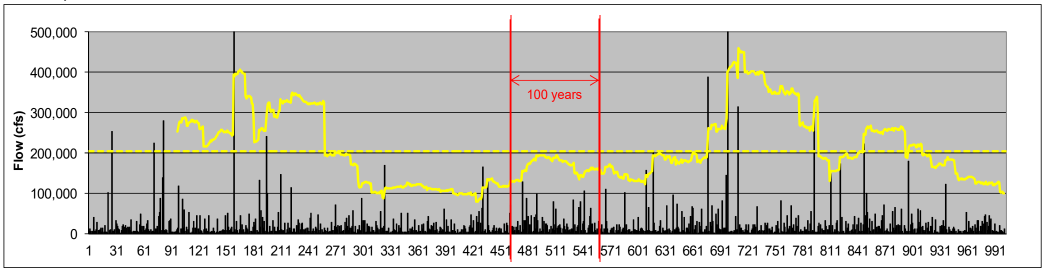
8

How does the skew of the data set affect the uncertainty in the estimate? The previous samples are from a distribution with a positive skew of 0.4. Next we compare to samples from a distribution with a negative skew of -0.4.

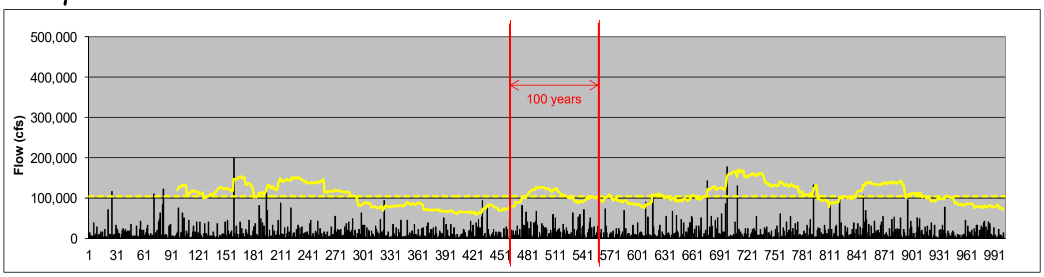
Note that these are the skews of the logarithms of the flows. On a linear scale, the skew is positive in both cases, as seen in the image on the right.

What about a more negative skew?

$\gamma = +0.4$



$\gamma = -0.4$



9

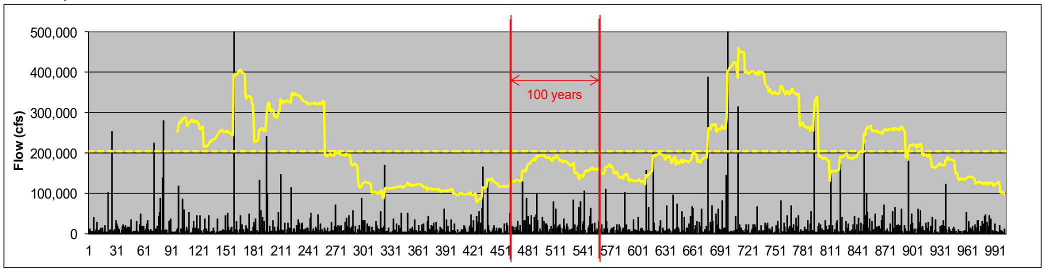
When skew is more negative, change has less influence on the 100-yr event

Although the parameter traces can no longer be seen on the slide, estimates of mean and standard deviation are unaffected by the change in the skew coefficient, and are reasonable with 100 years of record.   The estimate of skew coefficient is equally as variable with the negative as the positive value of skew coefficient.

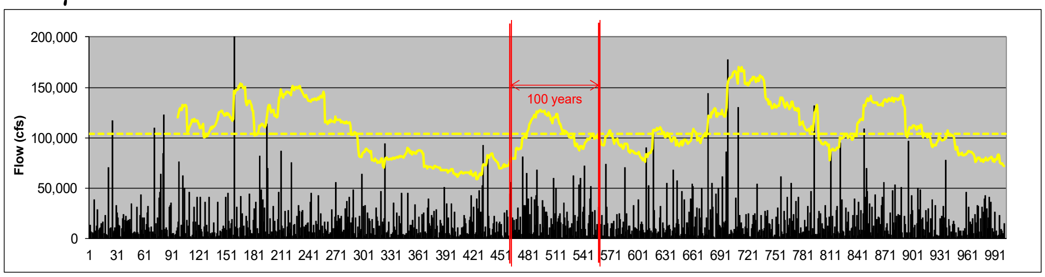
However, the 1% event (at the upper end of the distribution) is much less variable with a negative skew than a positive skew, being less sensitive to the skew coefficient. The result is that a frequency curve with negative skew is less uncertain for the less frequent events (the upper end) than a curve with positive skew. The negatively skewed curve has more uncertainty at the low end, which we don’t see here.

What about a more negative skew?

$\gamma = +0.4$



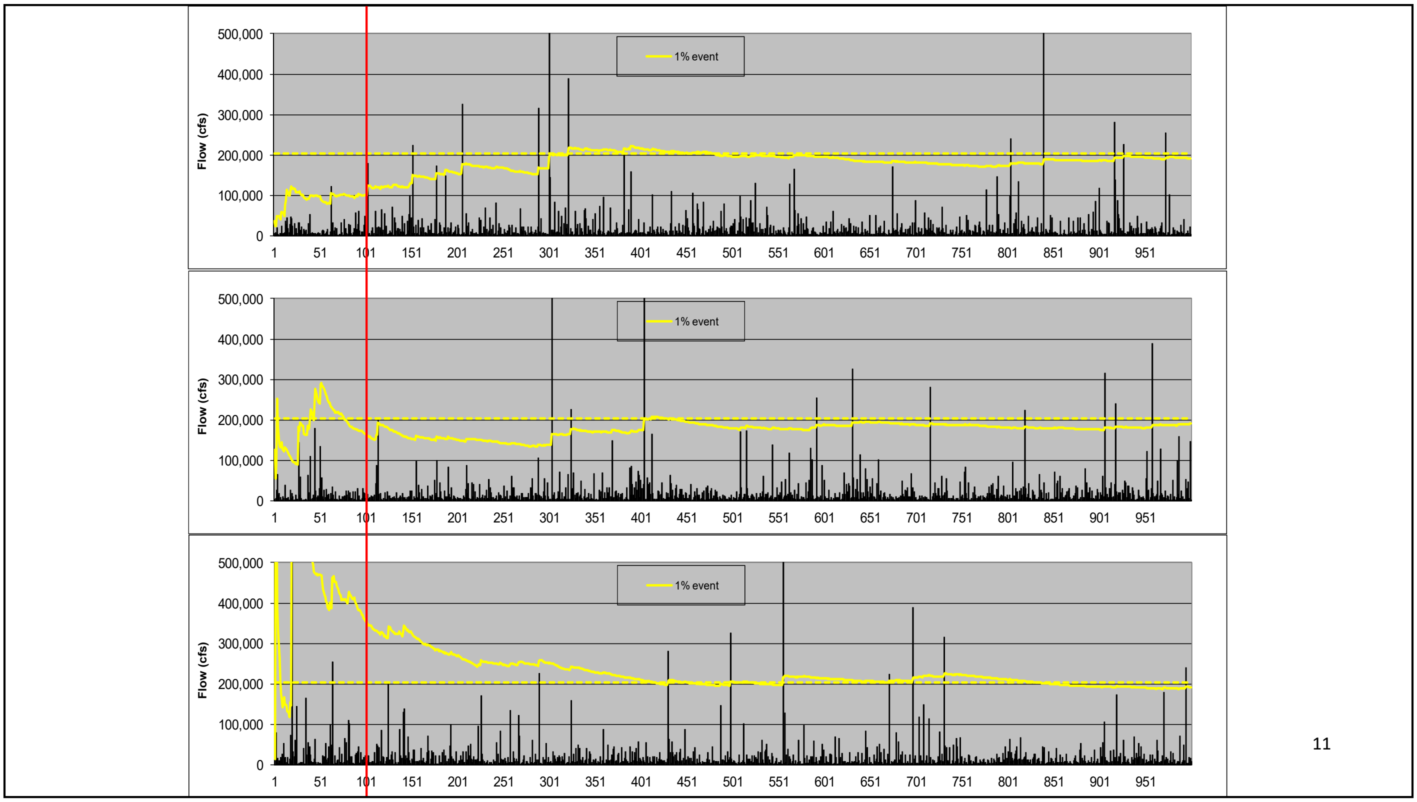
$\gamma = -0.4$



10

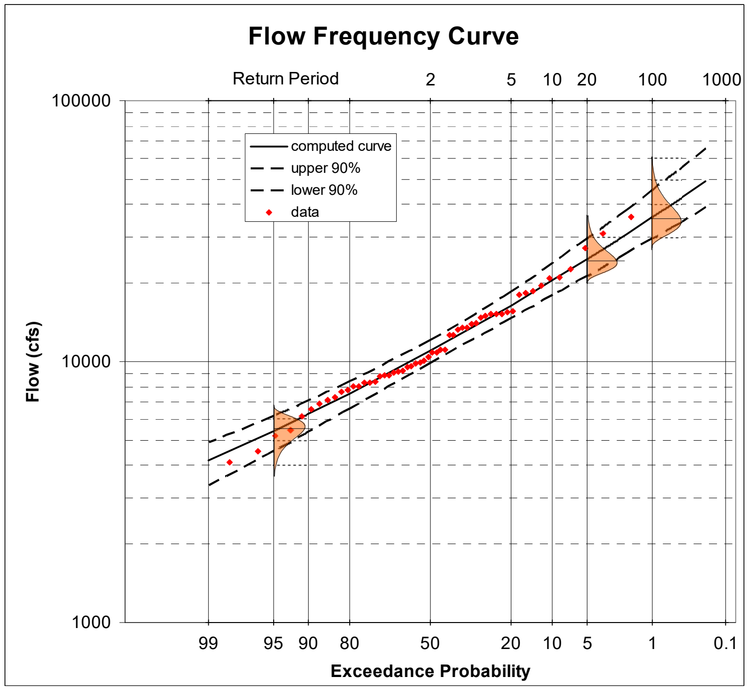
When skew is more negative, change has less influence on the 100-yr event

Just changed the vertical axis on the lower plot from 0 to 500,000 down to 0 to 200,000 cfs. This shows that on the scale of interest, the 1% estimate seems as variable.



The order in which events occur has an effect on the how the parameter estimates develop and change over time, although for any individual estimate (for example, at year 150) the order of the years included is irrelevant.

These graphics show the same 1000 years of artificial data, randomly shuffled to create a new time series of the same events. Note the difference in early estimates (within the first 100 years) in the first graph, in which no large events occur, and then third graph, when large events occur early.



12

When fitting a flood frequency curve to data, the picture we're accustomed to seeing is the fitted curve and the confidence interval. This lecture explored the uncertainty in this process, both what defines the confidence interval and what exists but does not inform the interval.

Uncertainty in Frequency Estimates

Definition:

Uncertainty is the degree to which we are unsure of an estimate, quantified by an estimated error

The estimated error is often stated in terms of an interval, but is more completely defined by a sampling distribution

The uncertainty in a frequency curve (or, probability distribution) is strongly driven by the question of **whether the sample is representative of the population**

13

Goals

- To understand the causes of uncertainty in estimating a frequency curve, or other probability distribution
- To become familiar with the uncertainty in each aspect of estimation
- To understand the confidence interval and uncertainty distribution around our frequency estimates

14

Topics

Motivation

Contributing Factors

Quantifying Sampling Error

- Sampling Distributions
- Confidence Intervals

Uncertainty in Frequency Estimates

- Analytical
  - Expected Probability
- Graphical

15

Factors Contributing to Uncertainty

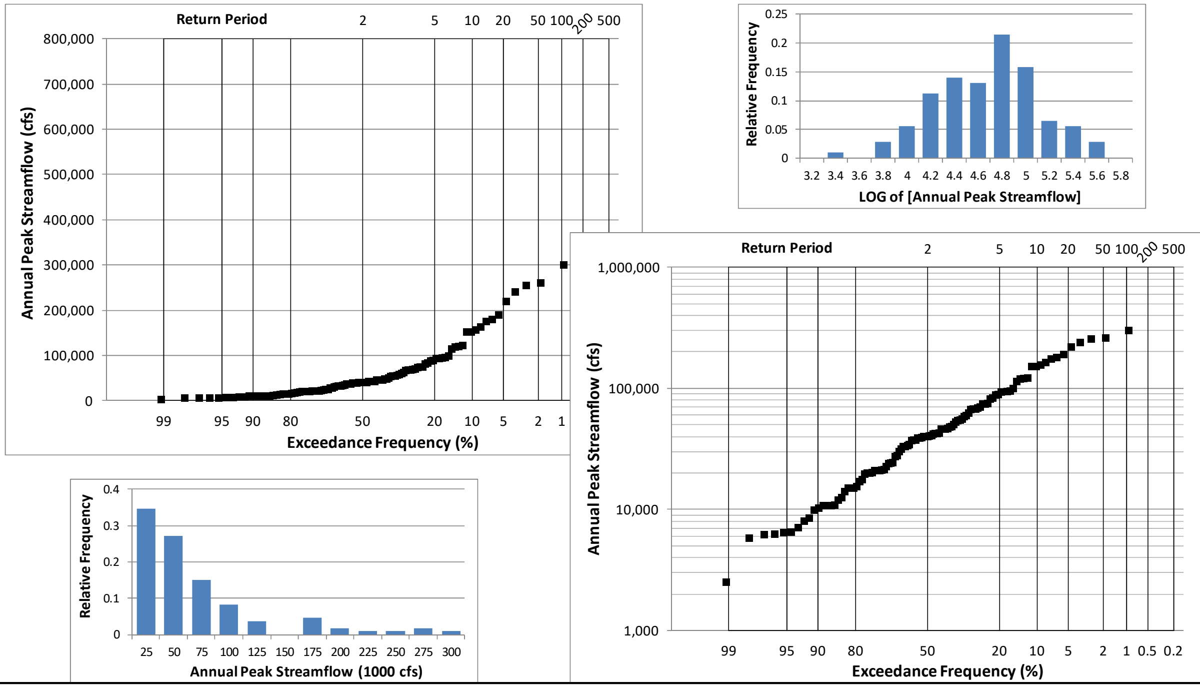
1. **Measurement Error**
   – How well are large flows measured?
2. **Model Error**
   – Does the log-Pearson type III distribution really describe annual peak flow frequency? Some other distribution?
   – What method is used to estimate parameters?

16

The distribution we choose to model our data has an impact on the resulting estimates of extreme probabilities, and we do not know the actual parent population distribution, if one of the common analytical choices is even a reasonable fit.

There are also various methods of estimating the parameters of the chosen distribution, such as Method of Moments with product moments, Method of Moments with L-moments, Method of Maximum Likelihood, etc that will sometimes provide different estimates. Which is the best choice?

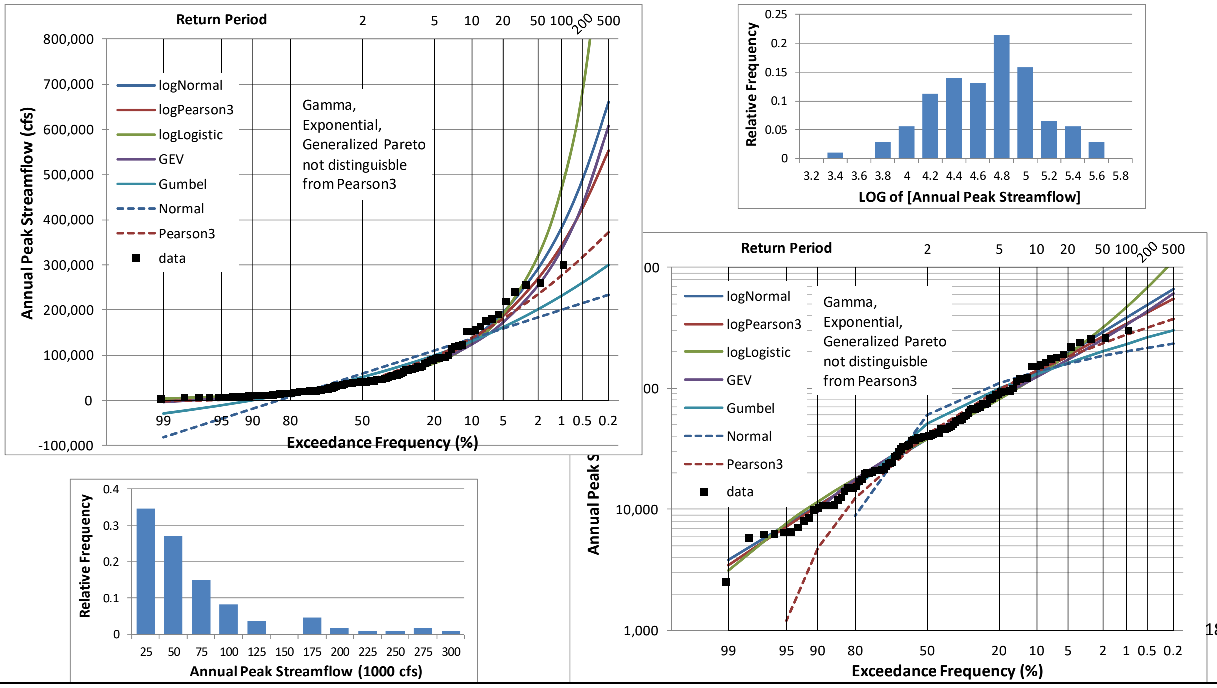
How much difference from distribution choice?



17

This is a data set of annual maximum flow shown with linear and with log axis, and histogram of both flow and log flow.

How much difference from distribution choice?



18

This is a data set of annual maximum flow shown with linear and with log axis, and histogram of both flow and log flow.

This figure contains several different distributions fit to this data set. Some, such as Normal and Gumbel, are clearly unreasonable because they suggest negative values. The others are reasonable in the middle of the curve, but can differ quite a lot at the extreme tails. We are often most interested in the extreme upper tail of maximum data, so these differences are notable.

Factors Contributing to Uncertainty

1. **Measurement Error**
   - How well are large flows measured?
2. **Model Error**
   - Does log-Pearson III distribution really describe flow frequency? Some other distribution?
   - What method is used to estimate parameters?
3. **Sampling error**
   - Error in estimates of distribution parameters due to limited sample size, causing possibly unrepresentative sample
   - this is the primary factor we quantify

19

We also have significant error in our estimates due to the lack of adequate data from which to estimate probability distributions. Since our sample size increases with observations made over time, this error is referred to as time sampling error -- the error from a short record. Even what seems to be a long record is still much shorter than we would prefer.

This lecture is primarily concerned with time sampling error.

Topics

Motivation

Contributing Factors

Quantifying Sampling Error

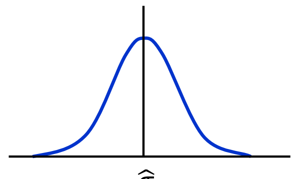
- Sampling Distributions
- Confidence Intervals

Uncertainty in Frequency Estimates

- Analytical
  - Expected Probability
- Graphical

20

Sampling Distributions



Definition

A sampling distribution gives us a description of the uncertainty in an **estimate** of a population value from a sample of size N *often summarized by standard error*

The estimate may be of a

- distribution parameter, or
- moment (e.g., the mean), or
- a probability (e.g., likelihood of exceeding 1000 cfs)
- a quantile (e.g., the 1%-chance event)

21

A sampling distribution give us a description of the uncertainty in the estimate of a distribution parameter or quantile.

The sampling distribution just describes how wrong we might be when estimating a parameter from a limited sample. We're going to study it with Monte Carlo simulation – this time generated samples of a particular size, again and again, and seeing how well they estimate the parameter. That will help us learn about the error in estimating the parameter.

Properties of Estimators

- Choosing an estimator is the subject of statistical estimation theory
- Estimation theory provides criteria for the selection of an estimator of a population value
- We'll focus on **consistency** and **unbiasedness**
- We'll use estimation of the MEAN as an example to work through

22

Estimator Consistency and Unbiasedness

Consistency

As the number of observations becomes very large, the estimated value ***approaches*** the population value

- the estimate gets better as the sample size increases

I repeated this phrase a lot on Monday!

Unbiasedness – has no bias

The expected value *(or the average of many instances)* of the estimator ***equals*** the population value

- even with small N, the estimator has no expected error
- "aiming at the right target"

23

Estimating the Mean

Consistent estimators of the mean → gets better as sample gets bigger

(1)

$$\hat{\mu} = \overline{X} = \sum_{i=1}^{i=N} \frac{X_i}{N}$$

(2)

$$\hat{\mu} = \sum_{i=1}^{i=N} \frac{X_i}{N-9}$$

However, equation (2) is biased

An example of an ***inconsistent*** estimator of the mean is the average of the largest and smallest values

24

Both of these estimators are consistent, in that they improve with sample size N, in spite of the fact that the N-9 in the denominator of the second estimator makes it biased, always overestimating the mean. While the bias is there, the -9 becomes less relevant as N gets large.

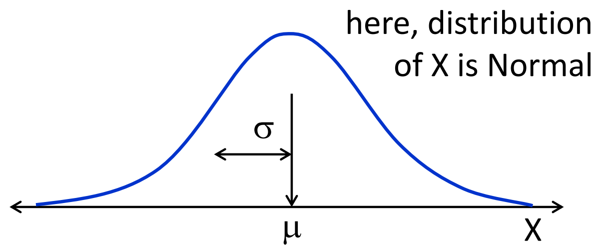
Creating a Sampling Distribution

Consider the sample mean, an <u>unbiased</u> estimator of the pop. mean

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{i=N} X_i$$

Let's examine properties of estimating the mean from 10 data points

- There exists an actual "population" probability distribution of X
- The distribution of X has a mean μ and standard deviation σ
- As an experiment, we'll generate 10-member random samples of flow X from the distribution of X above, 100 times



25

We're going to construct the sampling distribution – this is the Monte Carlo simulation. We're going to generate random samples, compute and estimate of the mean, and see how wrong they are.

The idea of the sampling distribution will be developed by looking at an estimate of the mean from 10 sample members. We can do a statistical experiment with a know distribution to examine the formation and properties of the sampling distribution of this estimator.

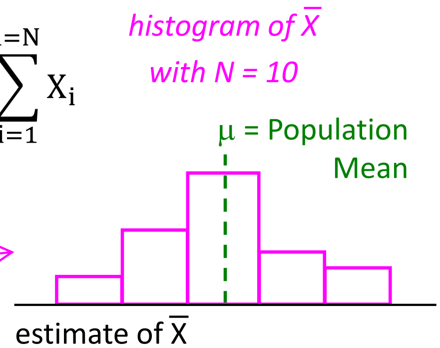
The experiment will start with a known distribution, and create random samples of 10 members each.

Creating a Sampling Distribution

- Generate 100 different 10-member samples of flow X
- Compute an estimate of the mean, $\overline{X}$, from each sample of 10

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{i=N} X_i$$

- A frequency analysis of the 100 estimates of $\overline{X}$ *(one from each 10-member sample)* would result in this histogram
- the average of 100 $\overline{X}$ would be close to the population value of the mean, μ
- By CLT, the histogram of $\overline{X}$ could be approximated by a Normal distribution



26

The experiment will start with a known distribution, and create random samples of 10 members each, and estimates the mean (using the sample average) of each sample. Done 100 times, this produces 100 estimates of the mean, each from a sample of size 10. This pink histogram is made up of the 100 estimates of the mean. It will be centered at the population values, because the estimator is unbiased.

Central Limit Theorem

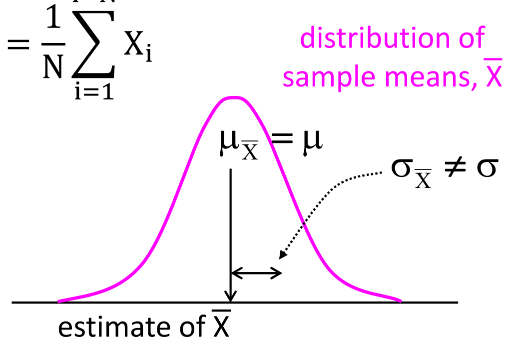
- The central limit theorem states that
  if we have a large number of
  independent, identically distributed (IID) random variables,
  the distribution of the sum is *approximately* Normal,
  regardless of the underlying distribution.
- The larger the sample size, N, the closer to Normal
  (i.e., the better the approximation)

27

Sampling Distribution for the Mean

- CENTRAL LIMIT THEOREM: Since the estimator adds identical RVs, as N becomes very large, the distribution of $\overline{X}$ is Normal, with mean equal to the population value.
- The pink distribution to the right is the sampling distribution for the sample mean, $\overline{X}$

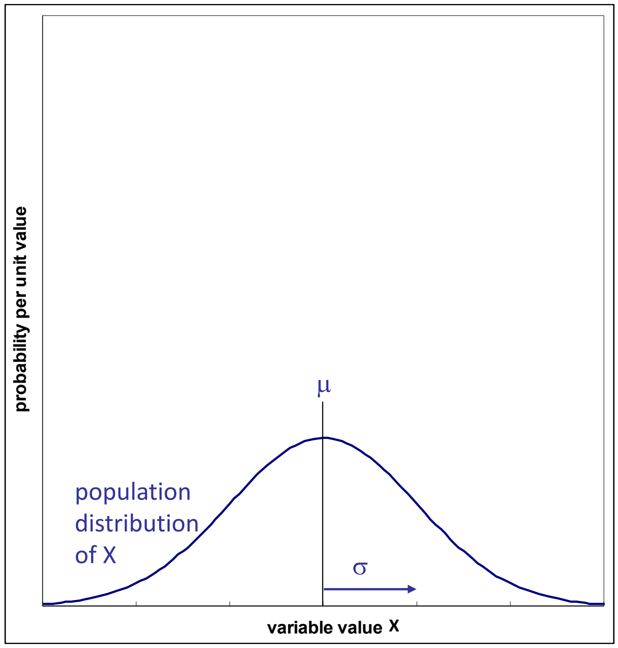
$$\overline{X} = \frac{1}{N}\sum_{i=1}^{i=N} X_i$$



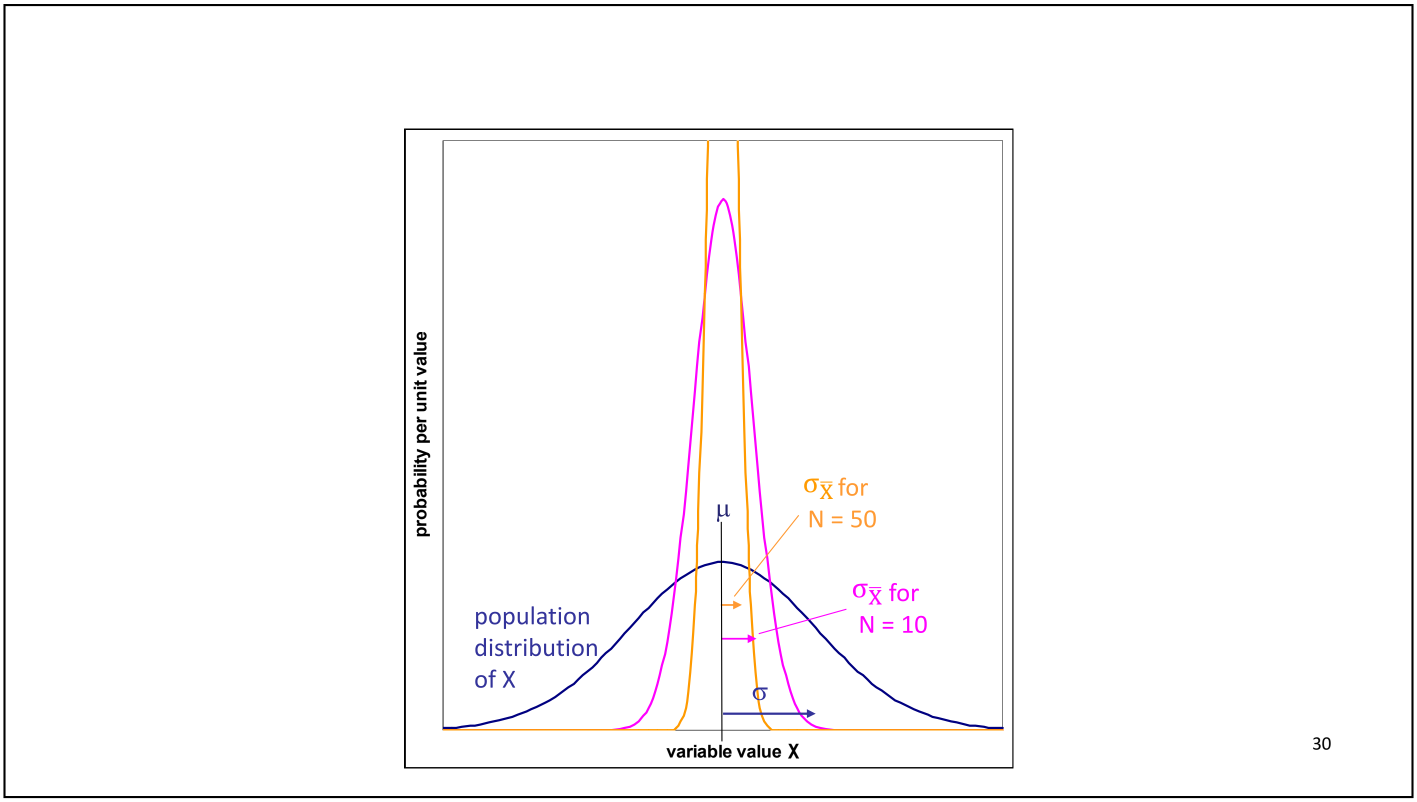
The standard deviation of the distribution-of-sample-means is equal to $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{N}}$

28

Because the Central Limit Theorem tells us that our sample of sample means is asymptotically Normal, we switch the histogram to a Normal distribution. It is centered on the known mean of the original distribution. Its variance is equal to sigma-squared/N.



The PDF of the parent population.



The population distribution is shown as the darker PDF. The sampling distribution of the estimate of the mean, X-bar, with sample size N=10 is shown as the lighter PDF (similar to the histogram on the last slide). If sigma is the standard deviation of the population distribution, then sigma divided by the square root of N is the standard deviation of the sampling distribution of X-bar.

The lightest PDF is the sampling distribution of the mean with sample size N=50. Note the uncertainty in the estimate is much smaller with a larger sample size. The standard deviation of the population distribution, sigma, is now divided by the square root of 50, rather than the square root of 10. A smaller standard deviation of the sampling distribution means that the error in the estimate is smaller, and therefore the estimate is better.

Demo of Sampling Distributions:

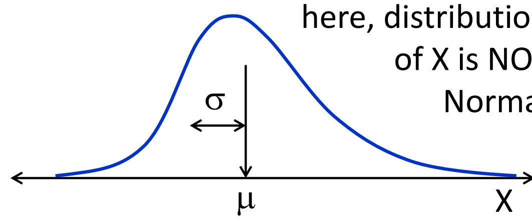
–estimating the mean from sample size N

Sampling Distribution

= Distribution of Uncertainty

When distribution of X *isn't* Normal

- There exists an actual "population" probability distribution of X
- This time, the distribution of X is **NOT** Normal
  - The distribution of X still has a mean μ and standard deviation σ
  - With the same experiment, assume that we have 100 different 10-member random samples of flow X from the distribution of X



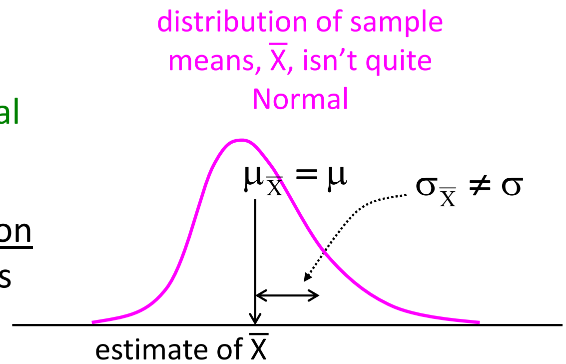
32

This is the same as the previous development of the sampling distribution of the mean, except the starting distribution is not Normal.

Sampling Distribution for the Mean

Because the estimator is unbiased, as the number of 10-member samples becomes very large, the mean of the sample is equal to the population value.

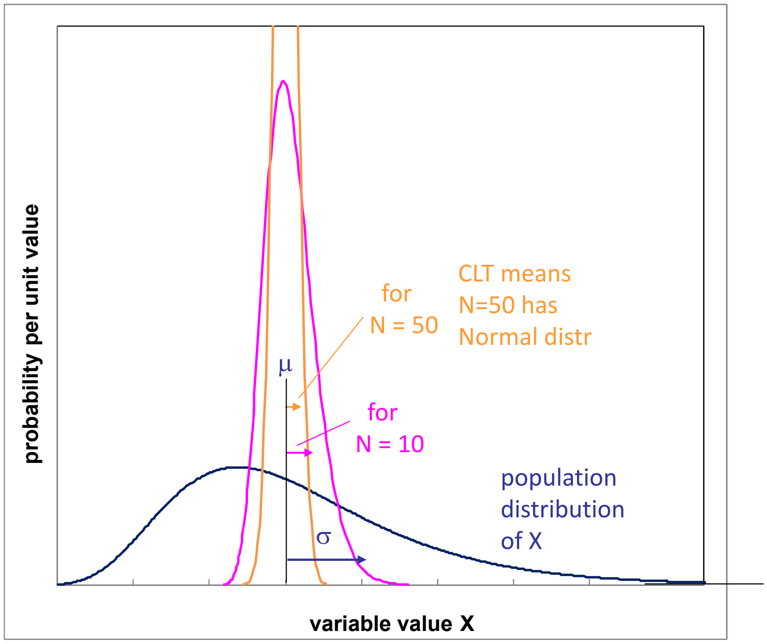
For small N (< 30), the sampling distribution for the mean is not quite Normal (CLT says Normal approx not good enough)



The standard deviation of the distribution-of-sample-means is equal to $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{N}}$

33

When the sample size is small enough, such as 10 members, the sampling distribution is not quite Normal. Sample size must be larger than about 30 to see that the sampling distribution of the mean will be Normal, even when the original distribution is not.



34

For small N of 10, the sampling distribution of the mean is not quite a Normal distribution with the underlying PDF is not Normal. But for N as large as 50, the sampling distribution of the mean has reached a Normal distribution, and would regardless of the population distribution.

Sampling Distributions

Sampling distribution for the **Sample Mean**

The **standard error σ** of sampling distribution of the sample mean, $\overline{X}$, is given by

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{N}}$$

Where σ is the population standard deviation of the random variable X

If X is Normally distributed, then the sampling distribution for the mean is exactly Normal, *regardless* of the sample size

If X is not Normal, sampling distribution is only Normal as $N \rightarrow \infty$

35

Sampling Distributions

Sampling distribution for the **Standard Deviation**

When X is Normally distributed, the **standard error σ** for the sample standard deviation, S_X, is given by

$$\sigma_{S_X} = \frac{\sigma}{\sqrt{2N}}$$

The sampling distribution of $(N-1)\left(\frac{S_X}{\sigma}\right)^2$ is distributed as chi-square, given that X is normally distributed

36

The spread of the sampling distribution of the estimate of standard deviation, Sx, is very similar to that of X-bar, except divided by the square root of 2N rather than N.

While the sampling distribution of X-bar is Normal, the sampling distribution of Sx is chi-square.

Sampling Distributions

Sampling distribution for the **Skew Coefficient**

The **standard error σ** for the sample skew, g or G, is given by

$$\sigma_g = \sqrt{\frac{6N(N-1)}{(N-2)(N+1)(N+3)}} \propto \frac{1}{\sqrt{N}}$$

The formula is applicable asymptotically for any random variable

37

Note that the sampling distribution of the skew estimate, G, is also proportional to one over the square root of N.

Sampling Distributions

Sampling distribution for the **Relative Frequency, Proportion**

The **standard error σ** for relative frequency as an estimate of probability is given by

$$\sigma_p = \sqrt{\frac{p(1-p)}{N}} \propto \frac{1}{\sqrt{N}}$$

The sampling distribution is approximately Normal for N*p>5 and N*(1-p)>5

38

For relative frequency, the sampling distribution becomes approximately Normal only when N*p>5 and N*(1-p)>5

Sampling Distributions

In General

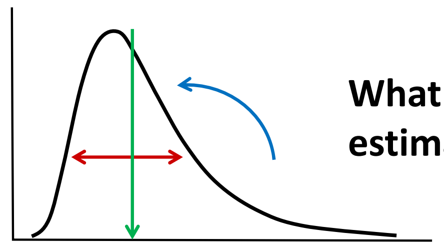
Standard error σ of the estimate for parameters, moments or probabilities is proportional to

$$\frac{1}{\sqrt{N}}$$

So, Mean Squared Error, **MSE** is proportional to 1/N

39

Estimation Matrix



What am I estimating? How?

How well am I estimating it?

sampling distribution of estimator

	mean	**standard deviation**	**skew**	→ parameter
	$\bar{X}$	S	g	→ estimator
mean (bias?)	μ	σ	biased!	
standard error (R-MSE)	$\frac{\sigma}{\sqrt{N}}$	$\frac{\sigma}{\sqrt{2N}}$	$\propto \frac{1}{\sqrt{N}}$	
skew	0	varies	varies	

40

Summary of sampling distribution parameters of distribution moments. Across the top is mean, standard deviation and skew, as parameters of the parent population. Down the side are the parameters of the sampling distributions of the estimators of the parent population parameters.

Topics

Motivation

Contributing Factors

Quantifying Sampling Error

- Sampling Distributions
- Confidence Intervals

Uncertainty in Frequency Estimates

- Analytical
  - Expected Probability
- Graphical

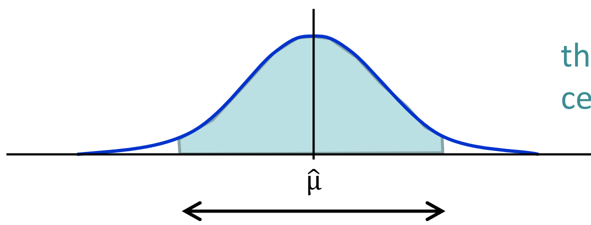
41

Confidence Intervals

Purpose

Estimate an interval that contains the population value with some probability, based on sample statistics

- based on sampling distribution



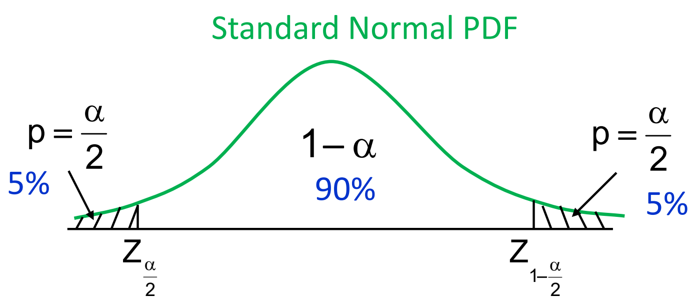
this would be a centered interval

42

Confidence Interval for the Mean

The sampling distribution of the estimate of the mean, $\overline{X}$, is Normal as $N \rightarrow \infty$, by Central Limit Theorem.

Considering Standard Normal Z ~ N[0,1], build an interval than spans desired percent of the distribution,   e.g. 90%



1-α is the desired probability within the interval, so
1-α = 90%,
α = 10%,
α/2 = 5%

43

We've established that the sampling distribution for the mean is Normal, centered on the population value.

For a Standard Normal variable, Z. We know that 90% of the Standard Normal distribution lies between -1.645 and +1.645. To generalize, in this example alpha = 10%, 1-alpha = 90%. Half of alpha is sectioned off each tail of the sampling PDF to form a range of 90% in the center.

Confidence Interval for the Mean

The sampling distribution of the estimate of the mean, $\overline{X}$, is Normal as N → ∞, by the Central Limit Theorem

standardize $\overline{X}$ by subtracting its mean and dividing by its standard deviation:

$$Z = \frac{\overline{X} - \mu}{\sigma_X / \sqrt{N}} \qquad Z \sim N(0,1)$$

90% of the standard Normal distribution exists between $Z_{.05}$ and $Z_{.95}$, which are equal to -1.645 and 1.645, so

$$P\left[\underset{Z_{.05}}{-1.645} < \frac{\overline{X} - \mu}{\sigma_X/\sqrt{N}} < \underset{Z_{.95}}{1.645}\right] = 0.90$$

α = 10%

1-α = 90%

α/2 = .05, 1-α/2 = .95

44

If standardize the Normal sampling distribution of X-bar by subtracting the mean and dividing by the standard deviation, create a Standard Normal variable, Z. Then can build the confidence interval using values of Z.

We know that 90% of the Standard Normal distribution lies between -1.645 and +1.645, so 90% of our standardized variable Z also lies in that range.

To generalize, in this example alpha = 10%, 1-alpha = 90%. Half of alpha is sectioned off each tail of the sampling PDF to form a range of 90% in the center.

Confidence Interval – Student's t

BUT, sampling distribution of X is only Normal as $N \rightarrow \infty$ For smaller N, nearly Normal, use **Student's t**

Approximate σ by S_X. Form a similar test statistic:

$$t_{N-1} = \frac{\overline{X} - \mu}{S_X / \sqrt{N}}$$

where t_{N-1} follows a **Student's t** distribution with N-1 degrees of freedom

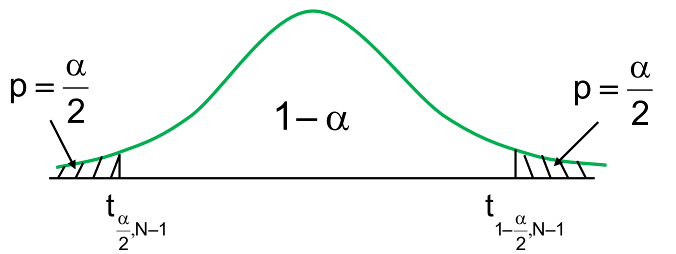
(*note: Student is a pseudonym for W.S. Gosset, writing in 1908)*

45

For small sample sizes that do not achieve exactly a Normal distribution, the Student's t distribution is used in place of the Standard Normal. The sample estimate of standard deviation, S, is used in place of the true value, sigma. The Student's t distribution is dependent on sample size, or degrees of freedom.

Confidence Interval for the Mean

PDF of "standardized" $\bar{X}$, estimate of the mean



1-α is the desired probability within the interval

A two-sided confidence interval can be formed by stating:

$$P\left[t_{\frac{\alpha}{2},N-1} < \frac{\bar{X}-\mu}{S_X/\sqrt{N}} < t_{1-\frac{\alpha}{2},N-1}\right] = 1-\alpha$$

46

Half of alpha is sectioned off from each tail of the PDF. The Student's t value at alpha/2 and at 1-alpha/2 are read from a table, and used in place of Z.

Confidence Interval for the Mean

Rewrite equation in terms of an interval on the pop. value

$$P\left[X - \frac{S_X}{\sqrt{N}} t_{\frac{\alpha}{2},N-1} < \mu < X + \frac{S_X}{\sqrt{N}} t_{1-\frac{\alpha}{2},N-1}\right] = 1 - \alpha$$

which states that the probability that the population mean is contained in the interval shown is equal to 1-α

Note

In classical statistics, the confidence interval is the random variable (and changes with the sample), not the population parameter

47

The interval is rearranged to isolate the mean, mu. The edges of the interval can be computed from sample estimates and a table of the Student's t distribution for the sample size.

In classical statistics, the correct way to describe the confidence interval is that there is a 1-alpha chance (ie, 90% chance) that the constructed interval, based on estimates from a random sample, spans the population mean. We do NOT say that there is a 90% chance that the population mean is in the interval. The difference in language is because the probability is associated with the random variables, which in this case are the sample estimates, described by their sampling distributions. The population mean is not considered a random variable, as it has a single value.

In Baysian statistics, the population parameters are instead considered the random variables, whose distributions are improved and refined by the information in the sample.

Confidence Interval for the Mean

Example:

$\overline{X} = 1.1$, $S = 0.47$, $N = 19$, $\alpha/2 = 0.05$ (for 90% interval)

From a table of the Student t, noting that this distribution is symmetrical,

$t_{0.05,18} = t_{0.95,18} = 1.734$ for $N\text{-}1 = 18$

The confidence interval on the population mean becomes

$$P\left[1.1 - \frac{0.47}{\sqrt{19}}1.734 < \mu < 1.1 + \frac{0.47}{\sqrt{19}}1.734\right] = 0.90$$

$$P[0.91 < \mu < 1.29] = 0.90$$

48

EM 1110-2-1415
5 Mar 93

Table F-4

PERCENTAGE POINTS OF THE ONE-TAILED t-DISTRIBUTION

	Exceedance Probability													
DF	0.40	0.30	0.20	0.10	0.05	0.04	0.02	0.01	0.005	0.002	0.001	0.0005	0.0002	0.0001
1	.325	.727	1.376	3.078	6.314	7.916	15.885	31.821	63.657	159.153	318.309	636.619	1591.549	3183.099
2	.289	.617	1.061	1.886	2.920	3.320	4.849	6.965	9.925	15.764	22.327	31.599	49.985	70.700
3	.277	.584	.979	1.638	2.353	2.605	3.482	4.541	5.841	8.053	10.215	12.924	17.598	22.204
4	.271	.569	.941	1.533	2.132	2.333	2.999	3.747	4.604	5.951	7.173	8.610	10.915	13.034
5	.267	.559	.920	1.476	2.015	2.191	2.757	3.365	4.032	5.030	5.893	6.869	8.363	9.678
6	.265	.553	.906	1.440	1.943	2.104	2.612	3.143	3.708	4.525	5.208	5.959	7.074	8.025
7	.263	.549	.896	1.415	1.895	2.046	2.517	2.998	3.500	4.207	4.786	5.408	6.311	7.063
8	.262	.546	.889	1.397	1.860	2.004	2.449	2.896	3.355	3.991	4.501	5.042	5.812	6.442
9	.261	.543	.883	1.383	1.833	1.973	2.398	2.821	3.250	3.835	4.297	4.781	5.461	6.011
10	.260	.542	.879	1.372	1.812	1.948	2.359	2.764	3.169	3.716	4.144	4.587	5.202	5.694
11	.260	.540	.876	1.363	1.796	1.928	2.328	2.718	3.106	3.624	4.025	4.437	5.004	5.453
12	.259	.539	.873	1.356	1.782	1.912	2.303	2.681	3.055	3.550	3.930	4.318	4.847	5.263
13	.259	.538	.870	1.350	1.771	1.899	2.282	2.650	3.012	3.489	3.852	4.221	4.721	5.111
14	.258	.537	.868	1.345	1.761	1.887	2.264	2.624	2.977	3.438	3.787	4.140	4.616	4.985
15	.258	.536	.866	1.341	1.753	1.878	2.249	2.602	2.947	3.395	3.733	4.073	4.528	4.880
16	.258	.535	.865	1.337	1.746	1.869	2.235	2.583	2.921	3.358	3.686	4.015	4.454	4.791
17	.257	.534	.863	1.333	1.740	1.862	2.224	2.567	2.898	3.326	3.646	3.965	4.390	4.714
18	.257	.534	.862	1.330	1.734	1.855	2.214	2.552	2.878	3.298	3.610	3.922	4.334	4.648
19	.257	.533	.861	1.328	1.729	1.850	2.205	2.539	2.861	3.273	3.579	3.883	4.285	4.590
20	.257	.533	.860	1.325	1.725	1.844	2.197	2.528	2.845	3.251	3.552	3.850	4.241	4.539
21	.257	.532	.859	1.323	1.721	1.840	2.189	2.518	2.831	3.231	3.527	3.819	4.203	4.493
22	.256	.532	.858	1.321	1.717	1.835	2.183	2.508	2.819	3.214	3.505	3.792	4.168	4.452
23	.256	.532	.858	1.319	1.714	1.832	2.177	2.500	2.807	3.198	3.485	3.768	4.137	4.415
24	.256	.531	.857	1.318	1.711	1.828	2.172	2.492	2.797	3.183	3.467	3.745	4.109	4.382
25	.256	.531	.856	1.316	1.708	1.825	2.167	2.485	2.787	3.170	3.450	3.725	4.083	4.352
26	.256	.531	.856	1.315	1.706	1.822	2.162	2.479	2.779	3.158	3.435	3.707	4.060	4.324
27	.256	.531	.855	1.314	1.703	1.819	2.158	2.473	2.771	3.147	3.421	3.690	4.038	4.299
28	.256	.530	.855	1.313	1.701	1.817	2.154	2.467	2.763	3.136	3.408	3.674	4.018	4.275
29	.256	.530	.854	1.311	1.699	1.814	2.150	2.462	2.756	3.127	3.396	3.659	4.000	4.254
30	.256	.530	.854	1.310	1.697	1.812	2.147	2.457	2.750	3.118	3.385	3.646	3.983	4.234
40	.255	.529	.851	1.303	1.684	1.796	2.123	2.423	2.704	3.055	3.307	3.551	3.864	4.094
50	.255	.528	.849	1.299	1.676	1.787	2.109	2.403	2.678	3.018	3.261	3.496	3.785	4.014
60	.254	.527	.848	1.296	1.671	1.781	2.099	2.390	2.660	2.994	3.232	3.460	3.750	3.962
70	.254	.527	.847	1.294	1.667	1.776	2.093	2.381	2.648	2.977	3.211	3.435	3.719	3.926
80	.254	.526	.846	1.292	1.664	1.773	2.088	2.374	2.639	2.964	3.195	3.416	3.696	3.899
90	.254	.526	.846	1.291	1.662	1.771	2.084	2.368	2.632	2.954	3.183	3.402	3.678	3.878
100	.254	.526	.845	1.290	1.660	1.769	2.081	2.364	2.626	2.946	3.174	3.390	3.664	3.862
110	.254	.526	.845	1.289	1.659	1.767	2.078	2.361	2.621	2.940	3.166	3.381	3.652	3.848
120	.254	.526	.845	1.289	1.658	1.766	2.076	2.358	2.617	2.935	3.160	3.373	3.642	3.837
INF	.253	.524	.842	1.282	1.645	1.751	2.054	2.326	2.576	2.878	3.090	3.291	3.540	3.719

Note - Values have been generated by use of computer routines for the inverse t-distribution. A few values for exceedance probabilities of 0.005 and less may differ plus or minus 0.001 from published tables.

Table of the Student's t distribution from EM 1415.
Note that only half the distribution is listed (probabilities below 0.5) because the distribution is symmetrical.

Table of the Student's t distribution from EM 1415. Note that only half the distribution is listed (probabilities below 0.5) because the distribution is symmetrical.

Topics

Motivation

Contributing Factors

Quantifying Sampling Error

Sampling Distributions

Confidence Intervals

Uncertainty in Frequency Estimates

Analytical

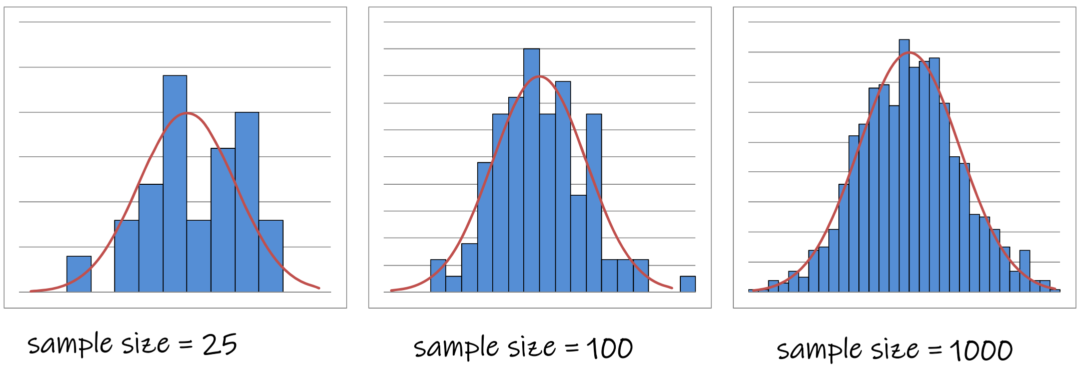
Expected Probability

Graphical

50

Uncertainty Due to Sampling Error

Confidence intervals provide a measure of uncertainty due to sampling error (limited sample size)



51

A sample has to be quite large to well represent the distribution it is drawn from. Note the histogram generated from the sample of N=25 does not closely resemble the parent PDF, while the sample of N=500 is more representative.

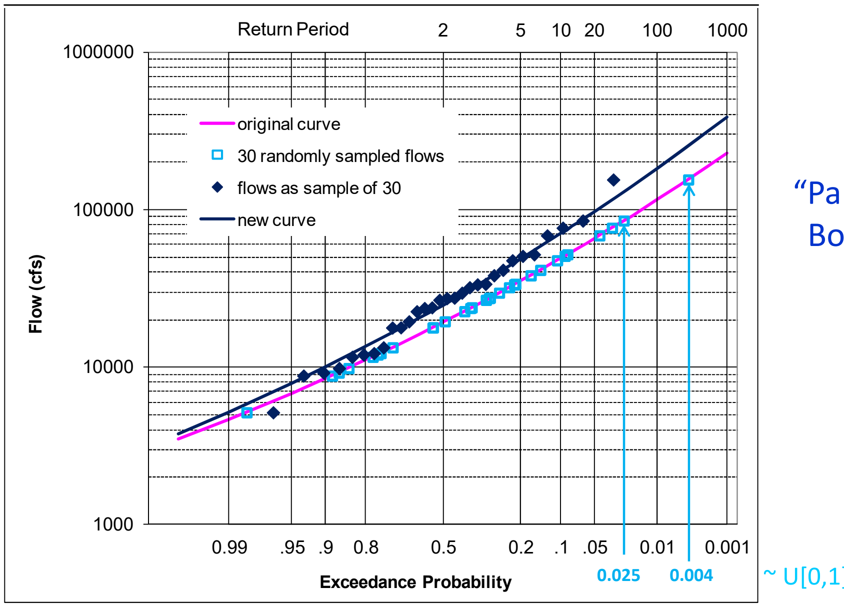
Confidence intervals on the estimates of distribution parameters and quantiles offer a description of the error caused by estimating with a small, unrepresentative sample.

Exploring Uncertainty

Creating a random sample from a known distribution

Similar to the moving window in the motivation slides, but an entire new sample each time

+ independent – don't see effect of each value



"Parametric Bootstrap"

52

This slide depicts a statistical experiment to explore the uncertainty in all quantiles of a frequency curve, for a sample of size N = 30.

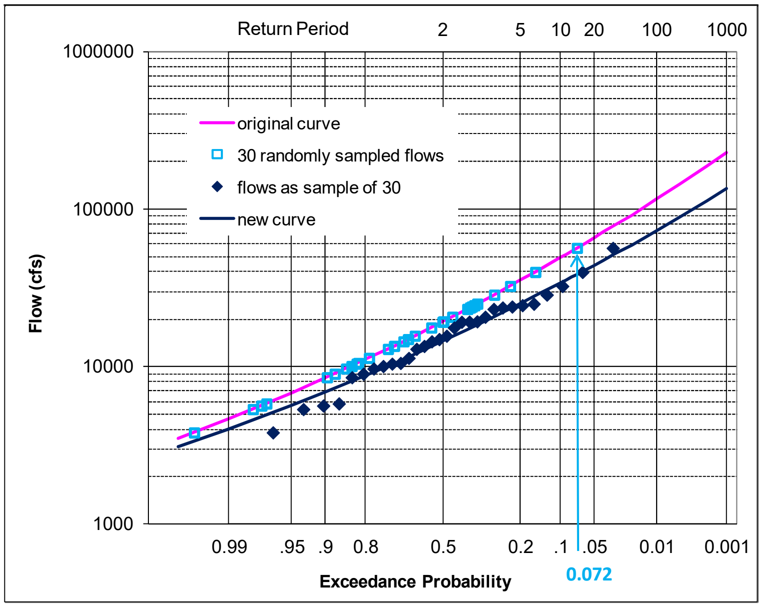
The pink frequency curve is specified as known. 30 pseudo random values U[0,1] are generated, and used as exceedance probabilities to sample 30 random annual maximum flows from the pink frequency curve, shown as light blue points. Note the smallest sampled exceedance probability is 0.004, producing a fairly large flow.

Next, the pink curve is put aside, and the 30 flows are plotted using median plotting positions in dark blue. Note that the point that was generated with exceedance probability 0.004 is plotted as the largest in 30 as 0.023. The sample values are used to estimate mean, standard deviation, skew coefficient and the resulting LP3 curve shown in dark blue.

This dark blue curve is a possible outcome of 30 years of record of a stream the had the initial pink curve as its population annual maximum flow frequency curve. This possible outcome is an overestimate of the specified "true" curve.

Exploring Uncertainty

Creating a random sample from a known distribution



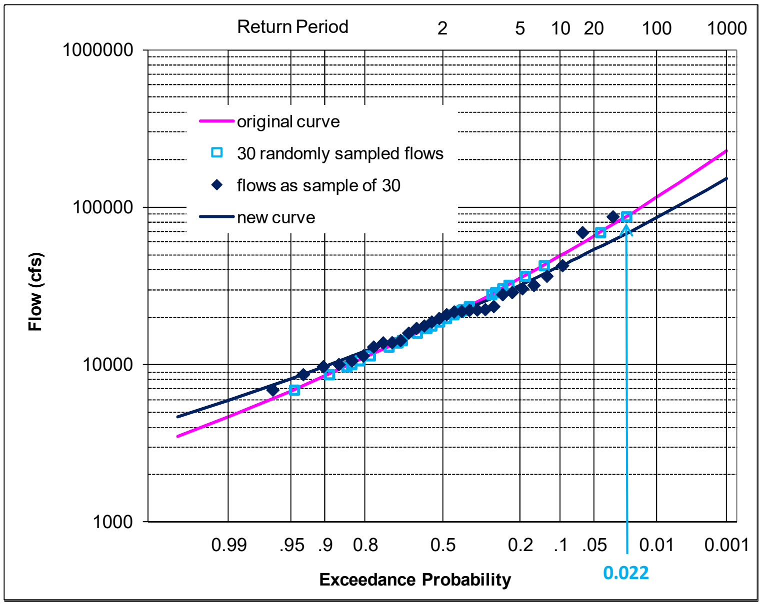
"Parametric Bootstrap"

53

This is another example of the statistical experiment in the previous slide. In this case, the smallest exceedance probability is 0.072, which provides a smaller than expected maximum flow value. As a sample, the 30 randomly sampled values produce a fitted LP3 curve that is lower than the "true" pink curve.

Exploring Uncertainty

Creating a random sample from a known distribution

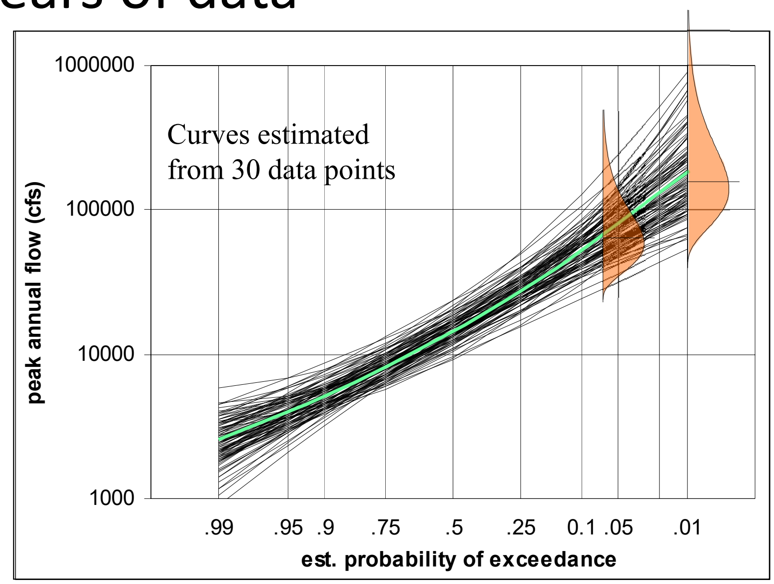


"Parametric Bootstrap"

54

This is another example of the statistical experiment in the previous slide.

Uncertainty in a frequency curve estimated from 30 years of data



"Parametric Bootstrap"

55

These estimated frequency curves are each based on a separate 30-year sample from the assumed parent probability distribution, shown as green rather than pink on this slide. Note they are all closer to the parent distribution near the median, and farther at the tails. As the assumed parent curve has a slightly positive skew, the uncertainty is greater on the upper end (less frequent events) than the lower end.

If the estimate of the 1% event from each sample is compiled, another probability distribution results from that compilation. This is a sampling distribution PDF that, if the skew coefficient is zero, is a non-central t distribution, as shown. A similar PDF exists at every quantile along the frequency curve.

Factors Affecting Frequency Curve Uncertainty

Uncertainty for a sample estimate of a statistic is proportional to: $\frac{1}{\sqrt{N}}$

ie, sample estimates of the mean, standard dev, skew, exc.prob, & quantiles

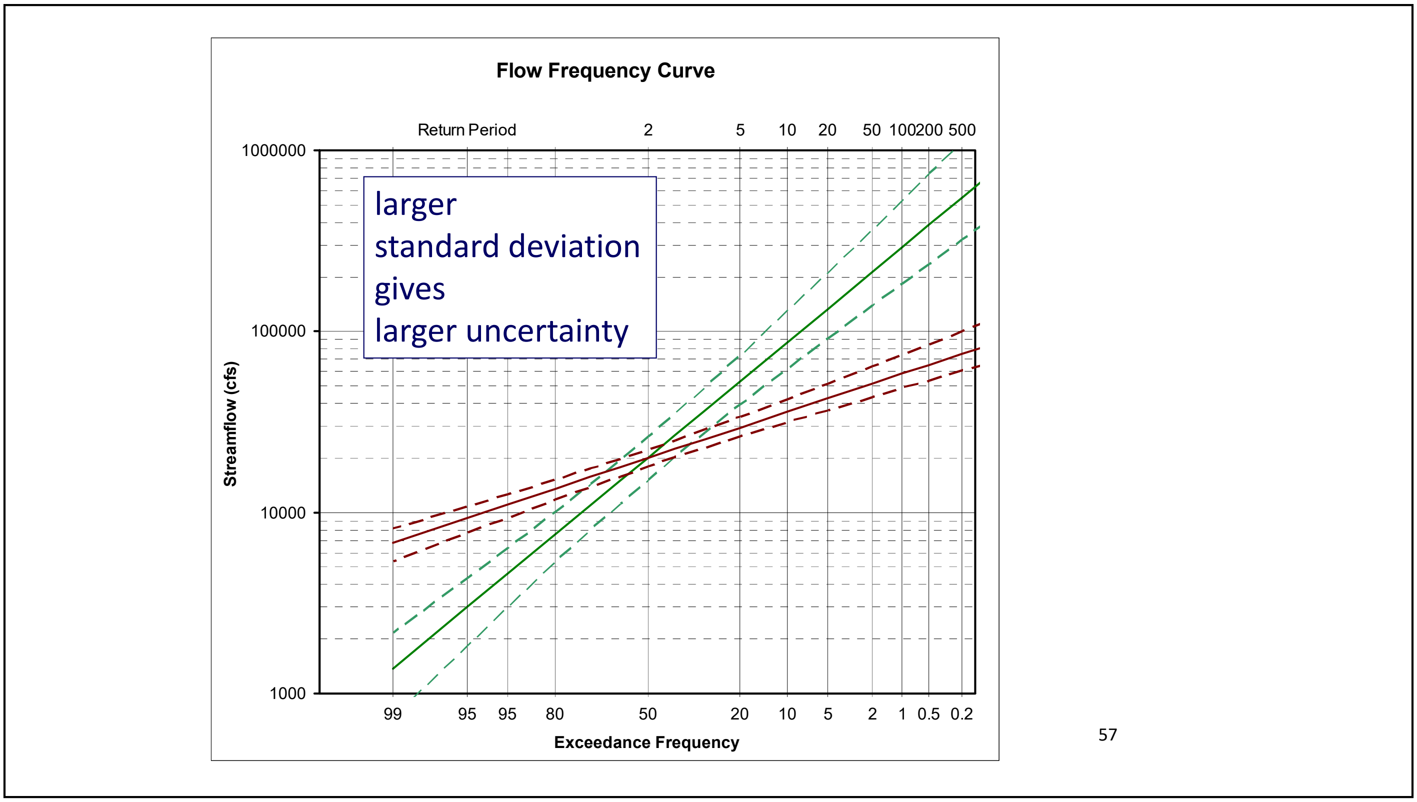
Uncertainty in estimates of exceedance probabilities or corresponding quantiles (flows) increases with:

- distance from the sample mean
- increase in variance
- increase in absolute value of skew

56

Recall from earlier, the standard deviation of the sampling distribution of each of the parameter estimates is proportional one over the square root of the sample size, N.

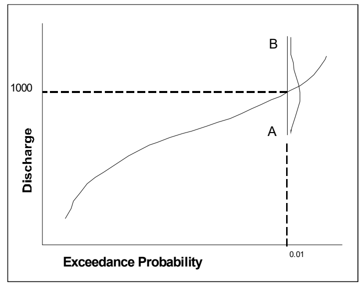
The previous slide shows the larger uncertainty at the upper and lower tail than the mean or median. The next slide shows larger uncertainty with larger variance. Uncertainty in the upper tail increases with positive skew and uncertainty in the lower tail increases with negative skew.

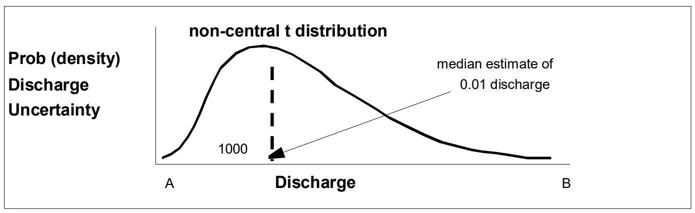


These curves have the same mean and skew coefficient, but the steeper green curve has larger variance (standard deviation) than the flatter red curve. As can be seen by the wider confidence interval, the larger variance produces greater uncertainty in the estimated frequency curve. It is intuitive that a more variable distribution will be more difficult to estimate accurately with a given sample size than a less variable distribution.

Frequency Curve Uncertainty

- The uncertainty around a flow-frequency curve is skewed (asymmetrical)
- **B17B** used the **non-central t** distribution to characterize the uncertainty in a LogPearson type3, rather than Normal or Student's t



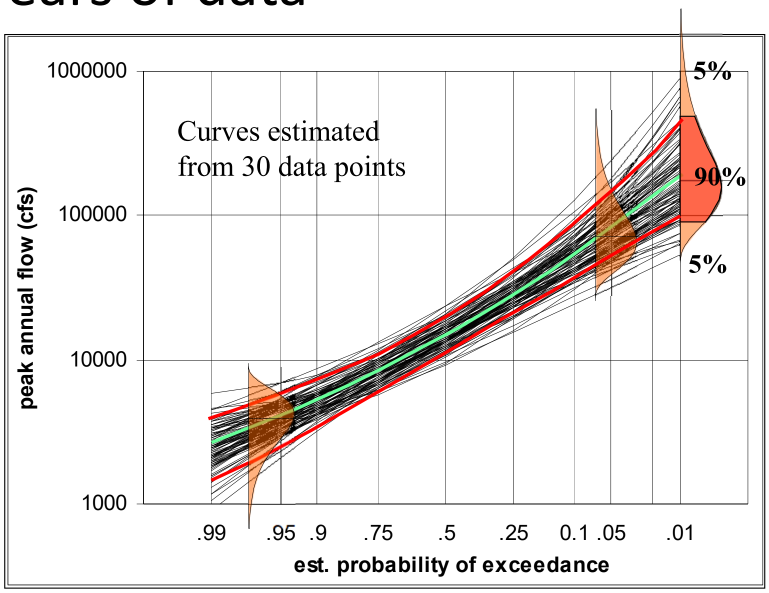


58

While the sampling distribution of the mean estimate, X-bar, is Normal, and for the standard deviation estimate, S, it is chi-square, the sampling distribution for a quantile (ie, 1% event, etc) is the non-central t distribution if the log-space skew coefficient is zero (ie, the log Normal distribution). Non-central t distribution is asymmetrical, and is dependent on the probability p and the skew g.

Note: the probability distribution around the 1% event in the upper graphic should be interpreting as a bell curve that comes out of the page. It is redrawn in the second graphic in only 2 dimensions, and so is flat on the page.

Uncertainty in a frequency curve estimated from 30 years of data



59

These estimated frequency curves are each based on a separate 30-year sample from the assumed parent probability distribution, shown as green rather than pink on this slide. Note they are all closer to the parent distribution near the median, and farther at the tails. As the assumed parent curve has a slightly positive skew, the uncertainty is greater on the upper end (less frequent events) than the lower end.

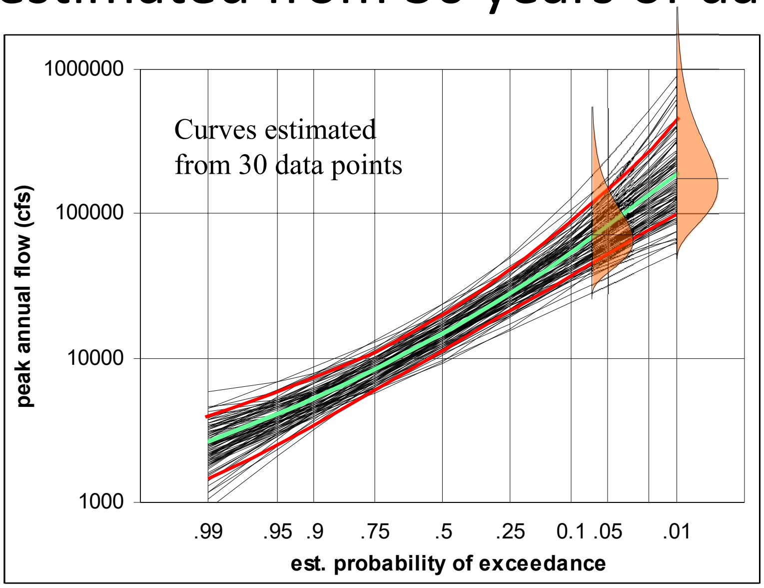
If the estimate of the 1% event from each sample is compiled, another probability distribution results from that compilation. This is a sampling distribution PDF that, if the skew coefficient is zero, is a non-central t distribution, as shown. A similar PDF exists at every quantile along the frequency curve.

The center 90% of the PDF can be delineated, generating an interval that captures 90% of these estimates. The same 90% interval can be delineated at each quantile, and connected, forming the 90% confidence interval for the entire frequency curve. This description is turned around in practice, and the 90% interval is drawn around the estimated frequency curve, defining an interval that has 90% chance of spanning the parent population curve.

17C Update to Confidence Limits

- Bulletin 17B confidence limits are based on a LogNormal distribution, which has log-space skew = 0 *and no skew parameter*
  - With no skew parameter, it therefore does not recognize or incorporate uncertainty in skew, and so underestimates the overall quantile uncertainty
- In 17C, EMA produces confidence intervals that are more correct in this aspect
  - and also incorporate regional skew and historical info

Uncertainty in a frequency curve estimated from 30 years of data

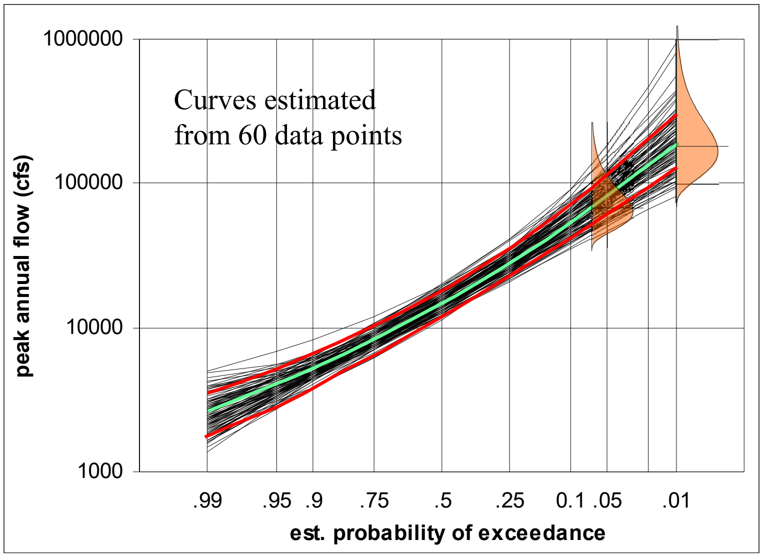


"Parametric Bootstrap"

61

30-year sample estimates of the population frequency curve, from 3 slides ago.

Uncertainty in a frequency curve estimated from 60 years of data

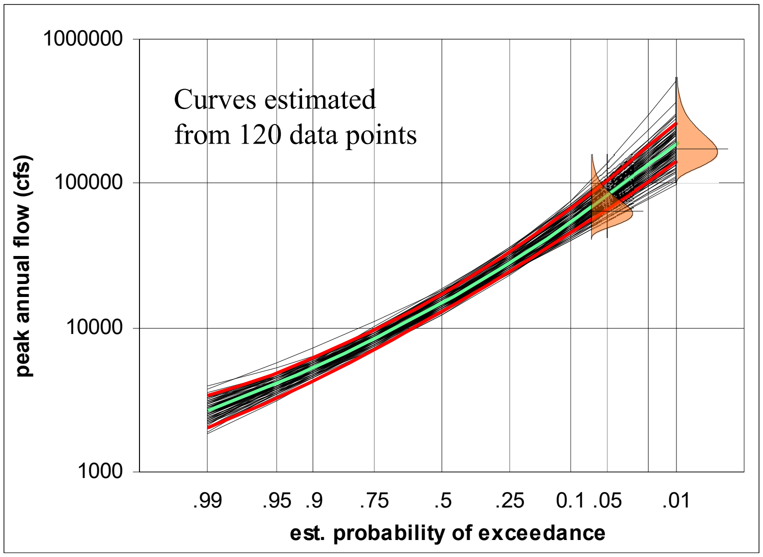


"Parametric Bootstrap"

62

60-year sample estimates of the population frequency curve. Note, estimates are closer to the assumed parent population frequency curve. This result demonstrates that estimates are better with a larger sample of the parent population, and our confidence intervals become narrower.

Uncertainty in a frequency curve estimated from 120 years of data



“Parametric Bootstrap”

63

120-year sample estimates of the population frequency curve. Note, estimates are closer yet to the assumed parent population frequency curve. This result demonstrates that estimates are better with a larger sample of the parent population, and our confidence intervals become narrower.

Topics

Motivation

Contributing Factors

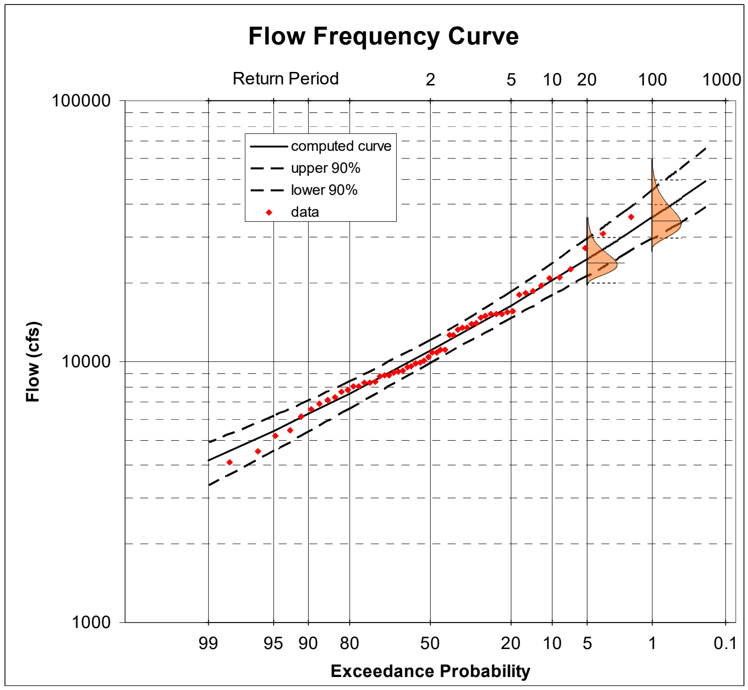
Quantifying Sampling Error

- Sampling Distributions
- Confidence Intervals

Uncertainty in Frequency Estimates

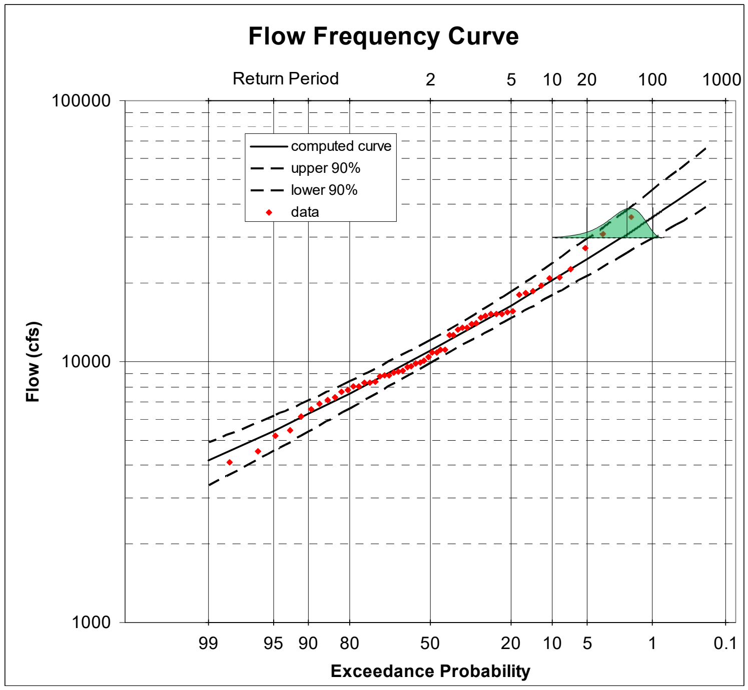
- Analytical
  - Expected Probability
- Graphical

64



We've noted that the sampling distribution of a quantile is asymmetrical. This slide shows the asymmetrical PDFs at 1% and 5%, and the resulting confidence interval for the entire curve.

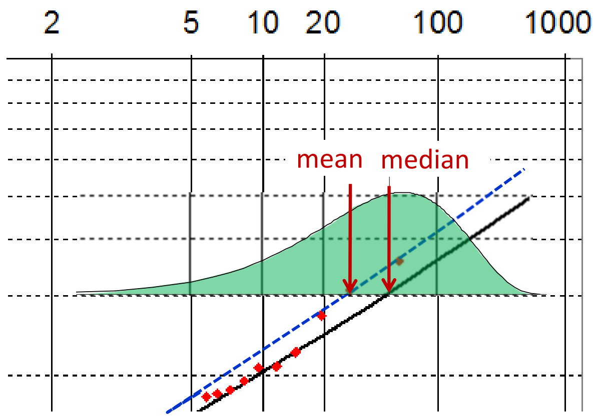
Note, for an asymmetrical distribution, the median is not equal to the mean.



The uncertainty in the frequency curve can also be considered in the exceedance probability of a particular flow, shown here at the estimate of the 1% flow but existing along the length of the curve. Like the distribution drawn as uncertainty in the flow quantile, the PDF is asymmetrical.

Expected Probability Adjustment

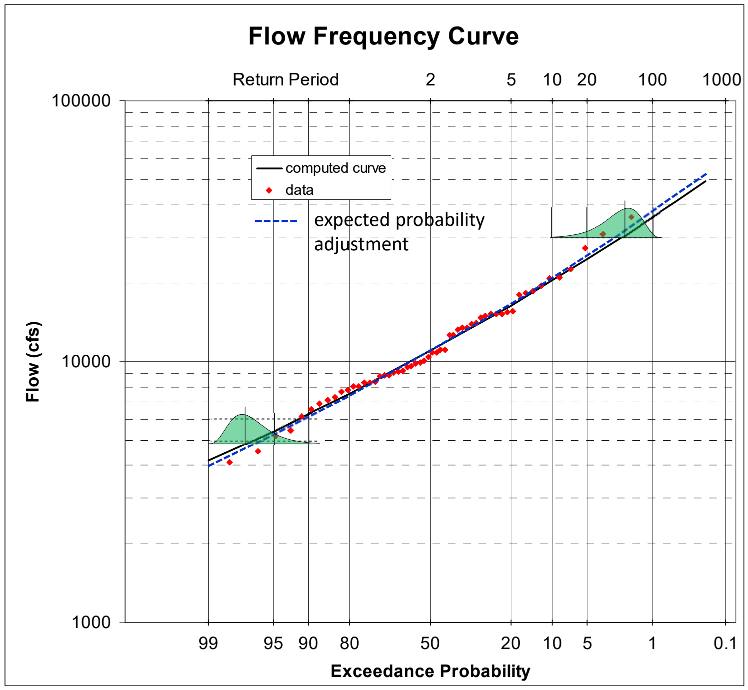
Adjusts the frequency curve from the median to the mean of the uncertainty distribution at each probability or flow



In the asymmetrical PDF of uncertainty in exceedance probably of a given flow, the mean is not equal to the median. The estimated flow frequency curve is at the median of the sampling distribution, and the mean of that distribution is a larger exceedance probability at the upper tail of the distribution.

The mean of the uncertainty distribution is the unbiased estimate of the frequency curve. When considering future outcomes from the probability distribution, and using the frequency curve to perform cost benefit analysis around the country, we prefer an unbiased estimate.

The "expected probability adjustment," shown as the dashed blue curve, moves the frequency ordinates from the median (the computed curve) to the mean of the uncertainty distribution.



68

Due to the direction of the asymmetry in the frequency curve uncertainty, the expected probability curve is higher and to the left at the top, and lower and to the right at the bottom. At the median, where the uncertainty PDF is symmetrical, there is not adjustment.

Expected Probability Adjustment

- The expected probability adjustment makes the frequency curve **unbiased**, in that it brings the estimate to the mean or expected value of probability
- This is the recommended curve when uncertainty is **NOT** carried forward into further analysis
  - Unfortunately, FEMA did not agree with its use in flood mapping
- For an analysis that does incorporate frequency curve uncertainty (HEC-FDA, HEC-WAT/FRA, RMC-RFA), the same effect is achieved, and the adjusted curve is not needed
- ***Dam safety PA does need an expected probability curve***

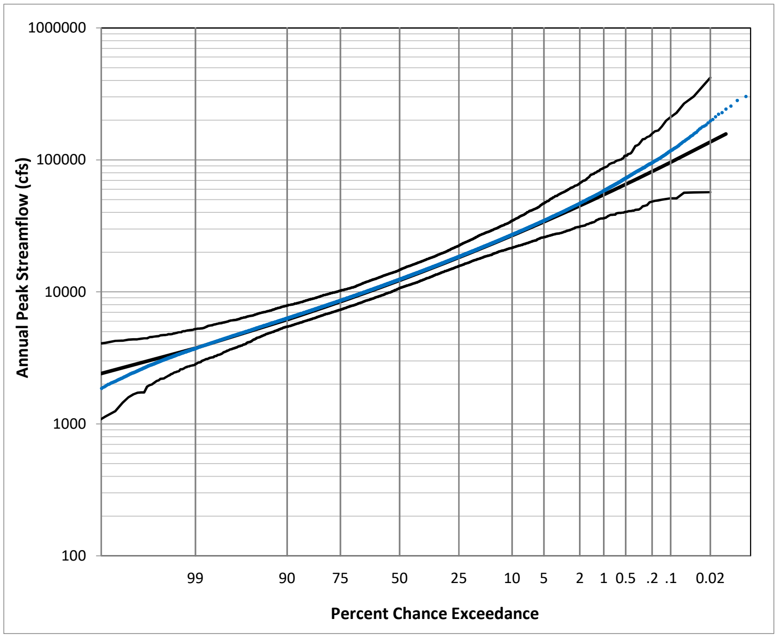
69

Expected Probability Computation

NOTE: the expected probability curve no longer has the original distribution, ie, **it is not LP3**.

How do you compute it?

1. *B17B had an adjustment, but without skew uncertainty*
2. Integrate the uncertainty CDF at each probability or flow
   – B17C output, repeated to get additional EMA confidence intervals- I^3
3. Use a similar Monte Carlo as creates confidence intervals
   – But, after refitting the LP3 frequency curve for each sample of size N, resample values from the new curve. Repeat for all curves, then pool and plot those values



71

The black frequency curve as the originally estimated LP3 curve. Monte Carlo simulation is used to resample sample size N values and fit a new potential LP3 (step 1), and then resample 1000 values from the new curve (step 2). This two step process is repeated 1000 times, and all the step 2 sampled values plotted as the light blue points. This traces out the expected probability curve.

Confidence intervals can also be estimated at each quantile of the step 1 curves by spanning 90% of the values.

Topics

Motivation

Contributing Factors

Quantifying Sampling Error

Sampling Distributions

Confidence Intervals

Uncertainty in Frequency Estimates

Analytical

Expected Probability

Graphical

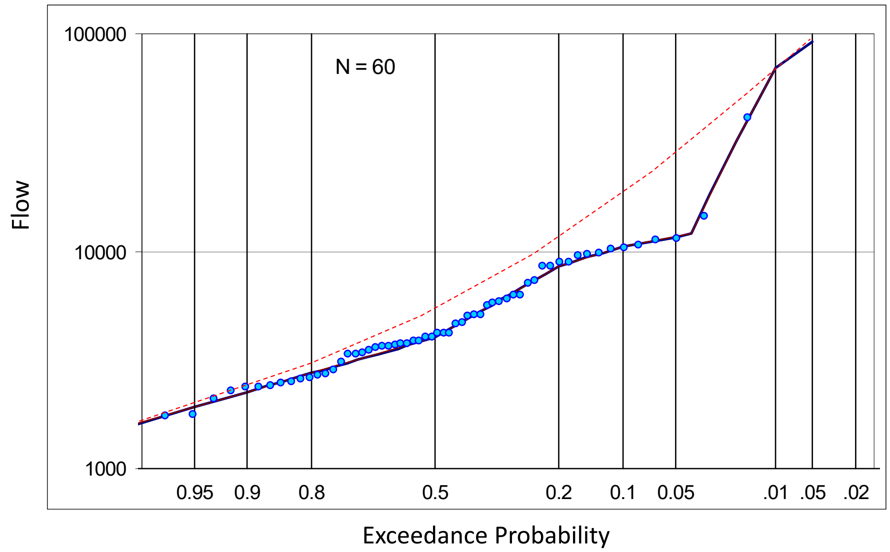
72

Confidence Intervals
Graphical Frequency Curves

- For the distribution of uncertainty around an analytical frequency curve, B17B made use of the fact that the distribution is like a logNormal distribution (so, used the non-central t distribution)
  - We also looked at a parametric bootstrap (Monte Carlo) approach
- Graphical frequency curves are purely empirical (have no distribution equation), and so we can not use a similar approach
- We need a "distribution-free" or "non-parametric" approach
  - order statistics and binomial

73

Graphical Frequency Curve



74

An example of a graphical frequency curve, including plotted points. The increase in the curve at 4% is likely due to exceeding the capacity of a flood protection reservoir.

Graphical Frequency Curve Uncertainty – Order Statistics Method

Order statistics:

- rank the observed data from smallest to largest, $Y_1, \ldots, Y_N$

Binomial distribution:

- for a given exceedance probability, p
  - for each ordered sample member, Y_j *(where j = 1, …, N),* determine the probability that it is less than the quantile Y_p
  - After computing for each j, have a CDF and PDF for Y_p
  - $P(Y_j \leq Y_p) = P(j \text{ or more observations} \leq Y_p)$
    - the probability that any one observation is less than Y_p is 1-p
  - $P(Y_j \leq Y_p) = \sum_{i=j}^{N} \binom{N}{i} (1-p)^i p^{N-i}$

note: we can approximate this sum with the incomplete beta function

75

The first step is the same as plotting the points to draw the graphical curve – rank the points from smallest to largest. This is referred to as an order statistics approach.

The uncertainty distribution is based on the binomial distribution. How likely is it that, in an N member sample, j values meet a certain criteria, given a probability 1-p of meeting that criteria? (As an example, how likely that 7 of 10 coin flips are heads, given a 50% chance of heads?)

The criteria of interest is that the sample member is below quantile Yp. The probability of interest is 1-p of being less than Yp. For sample member Yj to be below Yp, at least j members of the sample must be below Yp, because sample members are in order.

Graphical Frequency Curve Uncertainty – Order Statistics Method

Y_{60}

Y_{59}

Y_{58} — $Y_{0.04}$

this value is designated Y_{58} because 58 values are less than or equal to its value, and 2 values are greater

…

…

…

Y_3

Y_2

Y_1

For Y_{58} to be less than $Y_{.04}$, at least 58 values would need to be less than $Y_{.04}$

If only 57 values are less than $Y_{.04}$, Y_{58} would be greater than $Y_{.04}$

We don't know the value of $Y_{.04}$, but we know it's exceeded with probability 0.04

$$\Pr(\text{exactly 58 values} \le Y_p) = \binom{60}{58}(1-.04)^{58}.04^{2}$$

Pr(more than 58 values $\le Y_p$) =

$$\sum_{i=58}^{60}\binom{60}{i}(1-.04)^{60-i}.04^{i} = P(Y_{58} \le Y_{.04})$$

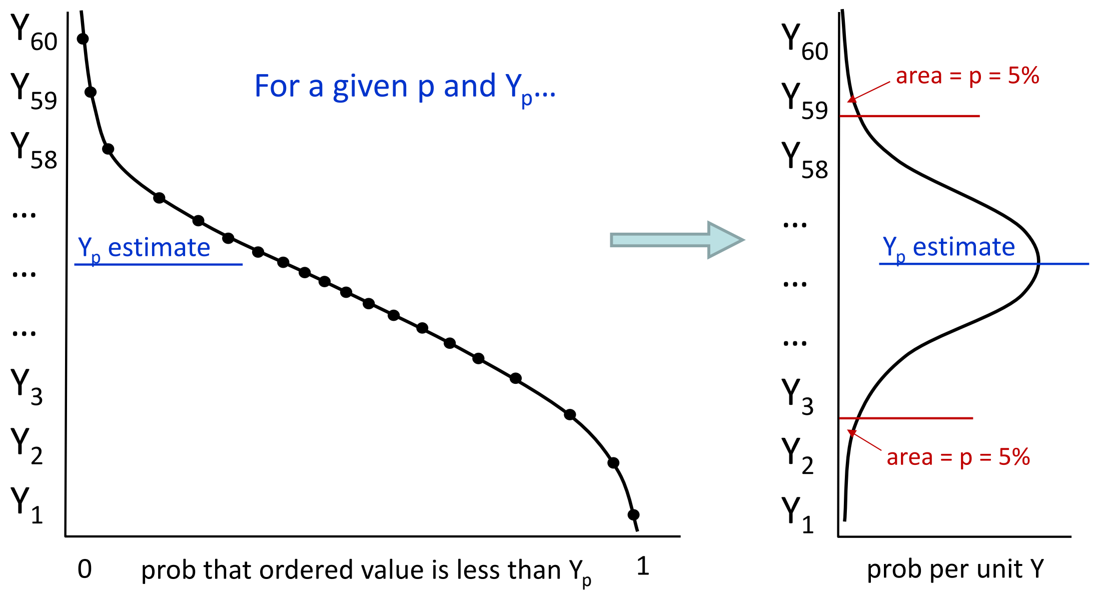
76

An example of application of the binomial distribution.

For a 60 member sample, the value Y58 is the third largest, with 2 values higher and 57 values lower. What is the probability that Y58 is less than the 4% exceedance event (25-year event), or Y.04? For Y58 to be less than Y.04, at least 58 members of the sample must be less than Y.04 – 58, 59, or 60 members must be less.

The likelihood of any sample member being less than Y.04 is (1-0.04). The equations above show first the likelihood that exactly 58 sample members are less than Y.04, and then the likelihood of 58, 59 or 60 sample members are less than Y.04, and so the probability that Y58 is less than Y.04.

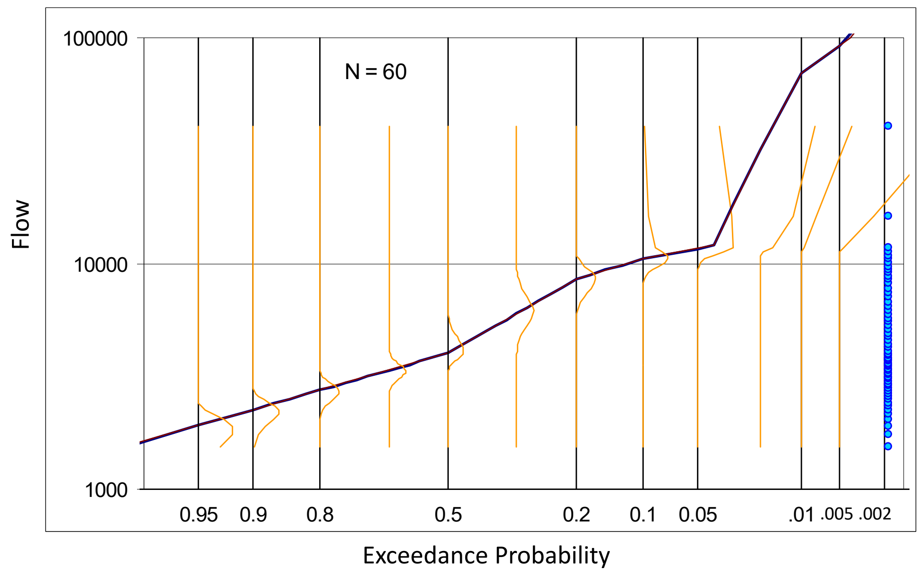
Graphical Frequency Curve Uncertainty – Order Statistics Method



77

Having computed the binomial for P(Yj < Yp) for each Yj, a cumulative distribution is formed. For each sample member Yj, we show the probability of being less than quantile Yp. The cumulative distribution (CDF) can be turned into a PDF as shown on the right, by computing probability increments. The PDF of Yp can define the uncertainty around the frequency curve, and so the 5% tails are defined, with 90% of the probability contained within.

Graphical Frequency Curve w/ uncertainty



78

When each of the constructed PDFs is placed around the quantile Yp, we see this graphic. Notice that the PDF can only be described for a given sample member, so it cannot be defined above or below the range of the data. This limit is particularly insufficent for the low frequency events.

Graphical Frequency Curve Uncertainty – Order Statistics Method

Need an uncertainty estimate for beyond the range of the data

Asymptotic approximation (large record length) for standard error of plotted quantile (flow or stage)

$$\sigma^2 = \frac{p(1-p)}{Nf_X^2}$$

where σ = standard deviation of estimate of flow or stage for exceedance probability p

N = systematic record length

f_X = estimated slope of frequency curve (probability density function)

79

Because the order statistics PDF can only be defined for the range of the sample data (or a pseudo sample generated at regular plotting positions), another method must be used to approximate the uncertainty beyond that range. The above formula is an estimate of the standard deviation of the uncertainty distribution around quantile Yp. Note, this is not a distribution, just the standard deviation. This SD estimate tends to be paired with an assumed Normal distribution of uncertainty.

Graphical Frequency Curve Uncertainty – Order Statistics Method

Confidence intervals are constructed assuming that sampling error follows a Normal distribution

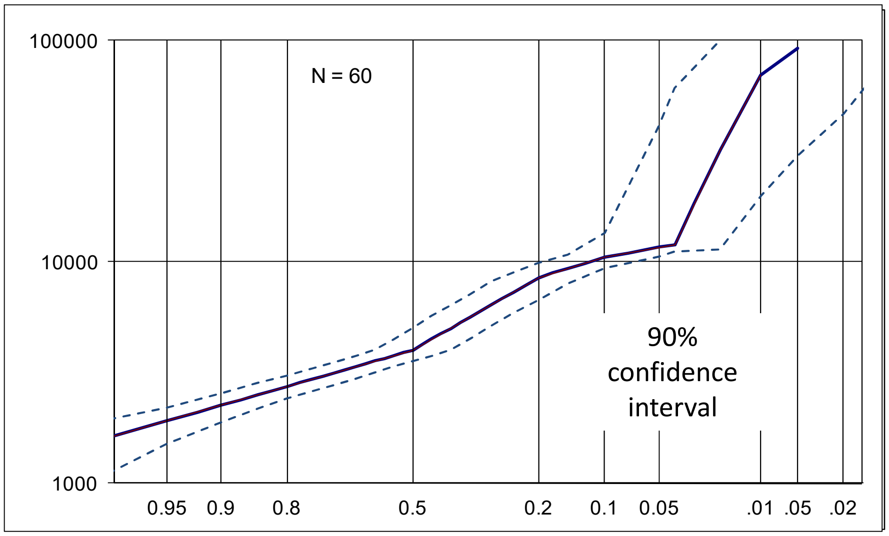
- i.e., symmetrical

Note that this is a distribution-free confidence interval and will be much wider than that obtained when the distribution is assumed (e.g., wider than for the logNormal distribution using the non-central t distribution)

80

Because the constructed PDFs are incomplete, practice is to compute the standard deviation of those PDFs, and use it to define a Normal distribution of uncertainty around the graphical frequency curve. The standard deviation from the asymptotic approximation is used the same way for the higher quantiles (lower frequencies). The result of this practice is that the uncertainty distribution is symmetrical, which is not a good assumption.

Graphical Frequency Curve w/ uncertainty



81

Graphical Uncertainty using Order Statistics

Limitations

- Not useful beyond the range of observations
  - Extrapolate beyond the range of data using an asymptotic approximation of variance, with aNormal distribution
  - The standard deviation of the Normal Distribution is set equal to that of the order statistic estimate near the observation extremes

Advantage

- Limits uncertainty when regulation moderates flows or in overbank areas *(is narrower where curve is flatter)*

Topics

Motivation

Contributing Factors

Quantifying Sampling Error

- Sampling Distributions
- Confidence Intervals

Uncertainty in Frequency Estimates

- Analytical
  - Expected Probability
- Graphical