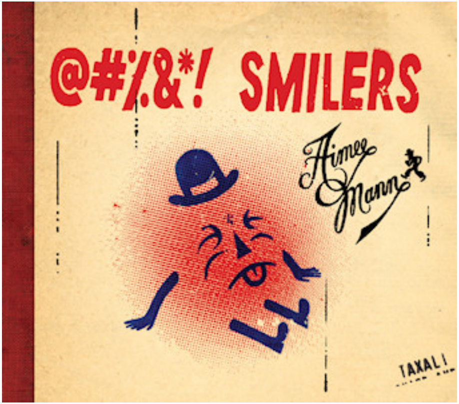
&#%!@?

Censoring and the Analysis of Partially Known Data

Phil Yates

Department of Mathematics
Saint Michael's College

October 10, 2014



Aimee Mann's 2008 album, *@#%&*! Smilers*

Censoring

Censoring

Examples of censoring:

Censoring

Examples of censoring:

- Suppose a student has five minutes to complete a task in the classroom. They fail to complete the task. How long would it **actually** take for the student to complete the task?

Censoring

Examples of censoring:

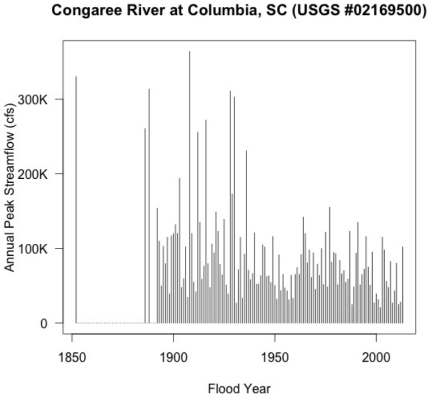
- Suppose a student has five minutes to complete a task in the classroom. They fail to complete the task. How long would it **actually** take for the student to complete the task?
- Suppose we are interested at which age children are able to count from 1–10 at school. Keep in mind that some children are already able to count before joining school. At what age should we say they were able to count?

Censoring

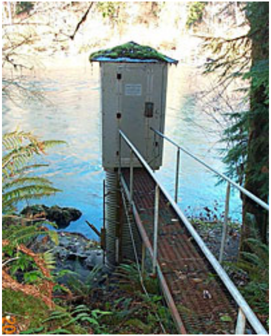
Examples of censoring:

- Suppose a student has five minutes to complete a task in the classroom. They fail to complete the task. How long would it **actually** take for the student to complete the task?
- Suppose we are interested at which age children are able to count from 1–10 at school. Keep in mind that some children are already able to count before joining school. At what age should we say they were able to count?
- Saint Michael's College students are asked in a survey the age at which they first tried marijuana. What do we do with students who answer "never"? What do we do with students who report using marijuana but forget when they first tried it?

Congaree River: Columbia, SC



Historic Floods: 1852, Possible historic floods: 1886, 1888
Flood Record: 1892 to 2013



Gage #12041200: Hoh River at U.S. Highway 101 near Forks, WA

Congaree River: Columbia, SC

Gervais Street Bridge, Columbia, SC
First version: 1827 to 1865
Second version: 1870 to 1928
Third version: 1928 to current
Flood gage about 0.2 miles south of current Gervais Street Bridge

1865: Columbia, SC

- Ruins, as seen from the State House, 1865.
- General Sherman's Union troops were slowed entering Columbia by a major flood on the Congaree River.

Flood Frequency Data

Annual Peak Flows

The maximum momentary peak discharge in each year of record

- The year of record begins the previous October and ends in September of the current year
- Peak flows are measured in cubic feet per second (cfs)

Flood Frequency Data

Annual Peak Flows

The maximum momentary peak discharge in each year of record

- The year of record begins the previous October and ends in September of the current year
- Peak flows are measured in cubic feet per second (cfs)

Historical Flood Data

Any observation of flood stage or conditions made before actual flood data were collected systematically

- Data collected from old newspapers, diaries, museums, libraries, etc.

Modeling Flood Frequency Data

Annual peak flows are assumed to be independent and identically distributed.

Let $Y = \log_{10} X$, where X is the annual peak flow. Typically it is assumed that these $\log_{10}$ annual peak flows have the following probability density function (p.d.f.):

$$f_Y(y) = \frac{(y-\gamma)^{\alpha-1}\exp[-(y-\gamma)/\beta]}{\beta^\alpha \Gamma(\alpha)},$$

$\alpha > 0, \beta > 0, y > \gamma$, where:

α — the shape parameter
β — the scale parameter
γ — the shift or location parameter

Bulletin #17B (1981)

- FEMA uses it as the trade standard

Bulletin #17B (1981)

- FEMA uses it as the trade standard
- Method of moments used to estimate parameters of log-Pearson type III distribution – this distribution is the method of choice for hydrologists in the United States

Bulletin #17B (1981)

- FEMA uses it as the trade standard
- Method of moments used to estimate parameters of log-Pearson type III distribution – this distribution is the method of choice for hydrologists in the United States
- When dealing with mixed populations, splits data into number of groups and fits a separate curve for each group

Bulletin #17B (1981)

- FEMA uses it as the trade standard
- Method of moments used to estimate parameters of log-Pearson type III distribution – this distribution is the method of choice for hydrologists in the United States
- When dealing with mixed populations, splits data into number of groups and fits a separate curve for each group
- When collecting annual peak flow data, estimates the 99th percentile of the log-Pearson type III distribution; this is used as an estimate for the 1 percent chance — FEMA uses this in regulatory policy for floodplains

Other Works

American River flood frequency analyses (1999)

- Used both systematic annual peak flows and one historical flood observation from 1862

Other Works

American River flood frequency analyses (1999)

- Used both systematic annual peak flows and one historical flood observation from 1862
- Used EMA, the Expected Moments Algorithm (Cohn, et al., 1997) to fit the data to a log-Pearson type III distribution

Other Works

American River flood frequency analyses (1999)

- Used both systematic annual peak flows and one historical flood observation from 1862
- Used EMA, the Expected Moments Algorithm (Cohn, et al., 1997) to fit the data to a log-Pearson type III distribution
- Recommended the independent identically distributed approach to flood estimation

Other Works

American River flood frequency analyses (1999)

- Used both systematic annual peak flows and one historical flood observation from 1862
- Used EMA, the Expected Moments Algorithm (Cohn, et al., 1997) to fit the data to a log-Pearson type III distribution
- Recommended the independent identically distributed approach to flood estimation
- Tentative use of mixed models as remedy for certain types of non-stationarity in the flood data

Other Works (Mixtures)

Singh (1987a,b) used a mixture of two normal distributions and two lognormal distributions in order to consider an observed flood that is composed of two component distributions

Other Works (Mixtures)

Singh (1987a,b) used a mixture of two normal distributions and two lognormal distributions in order to consider an observed flood that is composed of two component distributions

Hirschboeck (1987) & Diehl and Potter (1987) discussed various reasons to believe that there may be two or more components to the observed flood's distribution

Other Works (Mixtures)

Singh (1987a,b) used a mixture of two normal distributions and two lognormal distributions in order to consider an observed flood that is composed of two component distributions

Hirschboeck (1987) & Diehl and Potter (1987) discussed various reasons to believe that there may be two or more components to the observed flood's distribution

- in the southeastern region of US, tropical storms

Other Works (Mixtures)

Singh (1987a,b) used a mixture of two normal distributions and two lognormal distributions in order to consider an observed flood that is composed of two component distributions

Hirschboeck (1987) & Diehl and Potter (1987) discussed various reasons to believe that there may be two or more components to the observed flood's distribution

- in the southeastern region of US, tropical storms
- in the western, and even midwestern and northeastern, regions of US, snowmelt

Finite Mixture Model

Suppose Y is a random variable or vector that takes values from sample space $\mathcal{Y}$.

$$p(y) = \pi_1 f_1(y) + \ldots + \pi_k f_k(y) \quad (y \in \mathcal{Y}),$$

where

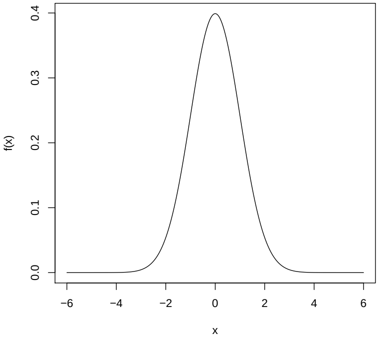
$$\pi_j > 0, \quad j = 1, \ldots, k; \quad \pi_1 + \ldots + \pi_k = 1$$

and

$$f_j(\cdot) \geq 0, \quad \int_{\mathcal{Y}} f_j(y)\,\mathrm{d}x = 1, \quad j = 1, \ldots, k,$$

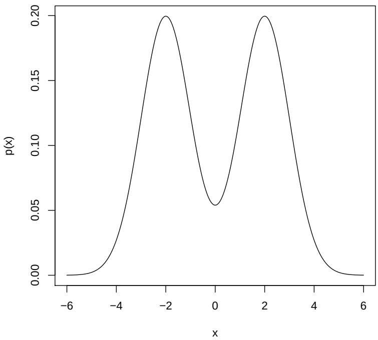
then Y has a finite mixture distribution

Finite Mixture Model



Standard Normal p.d.f.: $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$

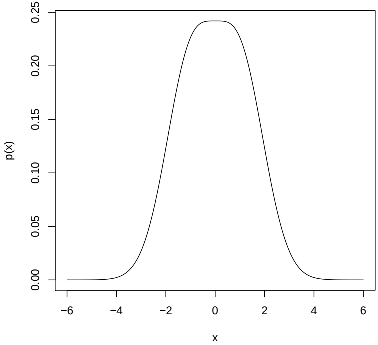
Finite Mixture Model



Finite Mixture Distribution:

$$p(x) = 0.5 * \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x+2)^2} + 0.5 * \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-2)^2}$$

Finite Mixture Model



Finite Mixture Distribution:

$$p(x) = 0.5 * \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x+1)^2} + 0.5 * \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-1)^2}$$

Finite Mixture Model: Congaree River

A finite mixture model with two normal components can be fit to the Congaree River's $\log_{10}$ annual peak flows (1892–2013):

Finite Mixture Model: Congaree River

A finite mixture model with two normal components can be fit to the Congaree River's $\log_{10}$ annual peak flows (1892–2013):

$$p(y) = (1-\tau)\frac{1}{\sqrt{2\pi}\sigma_0}e^{-\frac{1}{2\sigma_0^2}(y-\mu_0)^2} + \tau\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{1}{2\sigma_1^2}(y-\mu_1)^2}$$

Finite Mixture Model: Congaree River

A finite mixture model with two normal components can be fit to the Congaree River's $\log_{10}$ annual peak flows (1892–2013):

$$p(y) = (1-\tau)\frac{1}{\sqrt{2\pi}\sigma_0}e^{-\frac{1}{2\sigma_0^2}(y-\mu_0)^2} + \tau\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{1}{2\sigma_1^2}(y-\mu_1)^2}$$

An ECM (Expectation/Conditional Maximization) algorithm (Meng and Rubin, 1993) produced the following estimates for the finite mixture model:

Finite Mixture Model: Congaree River

A finite mixture model with two normal components can be fit to the Congaree River's $\log_{10}$ annual peak flows (1892–2013):

$$p(y) = (1-\tau)\frac{1}{\sqrt{2\pi}\sigma_0}e^{-\frac{1}{2\sigma_0^2}(y-\mu_0)^2} + \tau\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{1}{2\sigma_1^2}(y-\mu_1)^2}$$

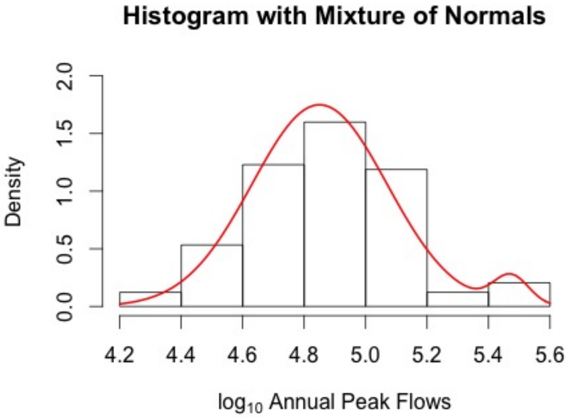
An ECM (Expectation/Conditional Maximization) algorithm (Meng and Rubin, 1993) produced the following estimates for the finite mixture model:

$$\hat{\mu}_0 = 4.8500 \quad \hat{\mu}_1 = 5.4732$$

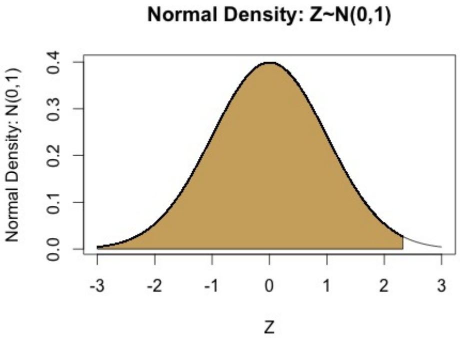
$$\hat{\sigma}_0 = 0.2198 \quad \hat{\sigma}_1 = 0.05757$$

$$\hat{\tau} = 0.03626$$

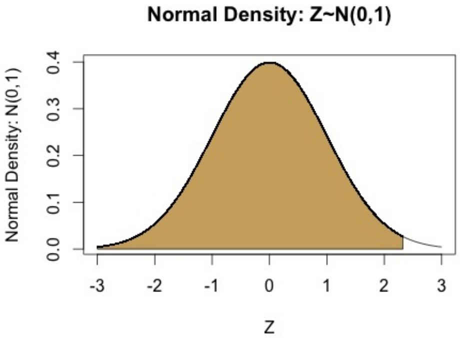
Finite Mixture Model: Congaree River



1% Chance Flood Estimation



1% Chance Flood Estimation



The 99th percentile of the standard normal distribution is $\Phi^{-1}(0.99) = 2.3263$.

1% Chance Flood Estimation

To find the 1 percent chance flood when the distribution is from a finite mixture model, one needs to find the 99th percentile of the finite mixture distribution. For example, if the distribution was a finite mixture model of normal densities, then the 100-year flood, Q, is

$$\int_{-\infty}^{\log_{10} Q} (1-\pi) f_0(y|\mu_0, \sigma_0^2) + \pi f_1(y|\mu_1, \sigma_1^2) dy = 0.99,$$

where $f_j(y) \sim N(\mu_j, \sigma_j^2)$ and $\log_{10} Q$ is the logged value of the 1 percent chance flood.

To find $\log_{10} \hat{Q}$, first obtain the 99th percentile from each component of the mixing distribution.

1% Chance Flood Estimation

These percentiles from each mixing component are used as end points to search for the root of

$$\int_{-\infty}^{\log_{10} \hat{Q}} (1-\pi) f_0(y|\mu_0, \sigma_0^2) + \pi f_1(y|\mu_1, \sigma_1^2) dy - 0.99 = 0.$$

In order to obtain a $100(1-\alpha)\%$ confidence interval for $\log_{10} Q$ in this example, a delta method argument can be made to find the standard error of $\log_{10} Q$ (Grego & Yates, 2010)

1% Chance Flood Estimation

These percentiles from each mixing component are used as end points to search for the root of

$$\int_{-\infty}^{\log_{10}\hat{Q}} (1-\pi) f_0(y|\mu_0, \sigma_0^2) + \pi f_1(y|\mu_1, \sigma_1^2) dy - 0.99 = 0.$$

In order to obtain a $100(1-\alpha)\%$ confidence interval for $\log_{10} Q$ in this example, a delta method argument can be made to find the standard error of $\log_{10} Q$ (Grego & Yates, 2010)

95% CI for Q: 277,944 cfs to 382,873 cfs

Weren't We Talking About Censoring?

Yes, we were! There are three ways to incorporate historic floods into a flood frequency analysis.

Weren't We Talking About Censoring?

Yes, we were! There are three ways to incorporate historic floods into a flood frequency analysis.

Type I Censoring: We determine a threshold for historic floods. All we know is how many floods, say r, are larger than the threshold during an historic period of M flood years.

Weren't We Talking About Censoring?

Yes, we were! There are three ways to incorporate historic floods into a flood frequency analysis.

Type I Censoring: We determine a threshold for historic floods. All we know is how many floods, say r, are larger than the threshold during an historic period of M flood years.

Type II Censoring: We treat a historic flood year's unknown annual peak flow as being smaller than the smallest known historic flood. Why? If it was larger, we would probably have some historic record of it!

Weren't We Talking About Censoring?

Yes, we were! There are three ways to incorporate historic floods into a flood frequency analysis.

Type I Censoring: We determine a threshold for historic floods. All we know is how many floods, say r, are larger than the threshold during an historic period of M flood years.

Type II Censoring: We treat a historic flood year's unknown annual peak flow as being smaller than the smallest known historic flood. Why? If it was larger, we would probably have some historic record of it!

Order Statistics Approach: (Grego, Yates, & Mai, hopefully 2014) This is a blend of the two censoring types. The historic flood is the largest flood of the preceding $m_1 - 1$ flood years. All that is known about the annual peak flows of the next m_2 flood years is that they are smaller than the historic flood.

Type I Censoring, Finite Mixture Models, & Congaree River

Components of the likelihood function:

Type I Censoring, Finite Mixture Models, & Congaree River

Components of the likelihood function:

- observed annual peak flows: 1892 to 2013

Type I Censoring, Finite Mixture Models, & Congaree River

Components of the likelihood function:

- observed annual peak flows: 1892 to 2013
- threshold? $\log_{10} 300,000$

Type I Censoring, Finite Mixture Models, & Congaree River

Components of the likelihood function:

- observed annual peak flows: 1892 to 2013
- threshold? $\log_{10} 300,000$
- historic period: 1852 to 1891 ($M = 40$ flood years)

Type I Censoring, Finite Mixture Models, & Congaree River

Components of the likelihood function:

- observed annual peak flows: 1892 to 2013
- threshold? $\log_{10} 300,000$
- historic period: 1852 to 1891 ($M = 40$ flood years)
- $r = 2$ floods larger than the threshold – 1852 and 1888

Type I Censoring, Finite Mixture Models, & Congaree River

Components of the likelihood function:

- observed annual peak flows: 1892 to 2013
- threshold? $\log_{10} 300,000$
- historic period: 1852 to 1891 ($M = 40$ flood years)
- $r = 2$ floods larger than the threshold – 1852 and 1888
- Remember the two flood components!
  - both floods could be in the large flood component and the other 38 floods can be in either flood component

Type I Censoring, Finite Mixture Models, & Congaree River

Components of the likelihood function:

- observed annual peak flows: 1892 to 2013
- threshold? $\log_{10} 300,000$
- historic period: 1852 to 1891 ($M = 40$ flood years)
- $r = 2$ floods larger than the threshold – 1852 and 1888
- Remember the two flood components!
  - both floods could be in the large flood component and the other 38 floods can be in either flood component
  - one flood could be in the large flood component, one flood could be in the small flood component, and the other 38 floods can be in either flood component

Type I Censoring, Finite Mixture Models, & Congaree River

Components of the likelihood function:

- observed annual peak flows: 1892 to 2013
- threshold? $\log_{10} 300,000$
- historic period: 1852 to 1891 ($M = 40$ flood years)
- $r = 2$ floods larger than the threshold – 1852 and 1888
- Remember the two flood components!
  - both floods could be in the large flood component and the other 38 floods can be in either flood component
  - one flood could be in the large flood component, one flood could be in the small flood component, and the other 38 floods can be in either flood component
  - both floods could be in the small flood component and the other 38 floods can be in either flood component

Type I Censoring, Finite Mixture Models, & Congaree River

Components of the likelihood function:

- observed annual peak flows: 1892 to 2013
- threshold? $\log_{10} 300,000$
- historic period: 1852 to 1891 ($M = 40$ flood years)
- $r = 2$ floods larger than the threshold – 1852 and 1888
- Remember the two flood components!
  - both floods could be in the large flood component and the other 38 floods can be in either flood component
  - one flood could be in the large flood component, one flood could be in the small flood component, and the other 38 floods can be in either flood component
  - both floods could be in the small flood component and the other 38 floods can be in either flood component
- Due to the complexity of the likelihood function, the standard error estimate is produced solely from score functions (first partial derivatives)

Type II Censoring, Finite Mixture Models, & Congaree River

Components of the likelihood function:

- observed annual peak flows: 1892 to 2013

Components of the likelihood function:

- observed annual peak flows: 1892 to 2013
- historic period: 1852 to 1891

Type II Censoring, Finite Mixture Models, & Congaree River

Components of the likelihood function:

- observed annual peak flows: 1892 to 2013
- historic period: 1852 to 1891
- observed historic floods: 1852, 1886, 1888

Components of the likelihood function:

- observed annual peak flows: 1892 to 2013
- historic period: 1852 to 1891
- observed historic floods: 1852, 1886, 1888
- flood years with missing values (1853 to 1885, 1887, 1889 to 1891) are censored

Components of the likelihood function:

- observed annual peak flows: 1892 to 2013
- historic period: 1852 to 1891
- observed historic floods: 1852, 1886, 1888
- flood years with missing values (1853 to 1885, 1887, 1889 to 1891) are censored
  - these floods are assumed to be smaller than the 1886 flood (the **smallest** observed historic flood

Type II Censoring, Finite Mixture Models, & Congaree River

Components of the likelihood function:

- observed annual peak flows: 1892 to 2013
- historic period: 1852 to 1891
- observed historic floods: 1852, 1886, 1888
- flood years with missing values (1853 to 1885, 1887, 1889 to 1891) are censored
    - these floods are assumed to be smaller than the 1886 flood (the **smallest** observed historic flood
- Since the likelihood function is not as complex as in the Type I censoring situation, the standard estimate is produce form score functions and Hessian matrices (first and second partial derivatives)

Components of the likelihood function:

- observed annual peak flows: 1892 to 2013

Components of the likelihood function:

- observed annual peak flows: 1892 to 2013
- historic period: 1827 to 1891

Order Statistics, Finite Mixture Models, & Congaree River

Components of the likelihood function:

- observed annual peak flows: 1892 to 2013
- historic period: 1827 to 1891
- observed historic flood: 1852

Order Statistics, Finite Mixture Models, & Congaree River

Components of the likelihood function:

- observed annual peak flows: 1892 to 2013
- historic period: 1827 to 1891
- observed historic flood: 1852
- assume that 1852 is the maximum flood in the $m_1 = 26$ flood year period (1827 to 1852) – this maximum value is the 26th order statistic in this period

Components of the likelihood function:

- observed annual peak flows: 1892 to 2013
- historic period: 1827 to 1891
- observed historic flood: 1852
- assume that 1852 is the maximum flood in the $m_1 = 26$ flood year period (1827 to 1852) – this maximum value is the 26th order statistic in this period
- assume that the 1853 to 1891 floods are all smaller than the 1852 flood

Components of the likelihood function:

- observed annual peak flows: 1892 to 2013
- historic period: 1827 to 1891
- observed historic flood: 1852
- assume that 1852 is the maximum flood in the $m_1 = 26$ flood year period (1827 to 1852) – this maximum value is the 26th order statistic in this period
- assume that the 1853 to 1891 floods are all smaller than the 1852 flood
- Due to the complexity of the likelihood function, the standard error estimate is produced solely from score functions (first partial derivatives)

Results

Type I Censoring:

$$\hat{\mu}_0 = 4.8629 \quad \hat{\mu}_1 = 5.4843$$

$$\hat{\sigma}_0 = 0.2360 \quad \hat{\sigma}_1 = 0.05162 \quad \hat{\tau} = 0.02253$$

95% CI for Q: 273,306 cfs to 382,728 cfs

Type II Censoring:

$$\hat{\mu}_0 = 4.8499 \quad \hat{\mu}_1 = 5.4718$$

$$\hat{\sigma}_0 = 0.2196 \quad \hat{\sigma}_1 = 0.05483 \quad \hat{\tau} = 0.04654$$

95% CI for Q: 310,400 cfs to 352,614 cfs

Order Statistics Approach:

$$\hat{\mu}_0 = 4.8547 \quad \hat{\mu}_1 = 5.4788$$

$$\hat{\sigma}_0 = 0.2255 \quad \hat{\sigma}_1 = 0.05049 \quad \hat{\tau} = 0.02962$$

95% CI for Q: 283,349 cfs to 368,694 cfs

Other Work and Future Work

Other Work: All of these methods have been used with a finite mixture of two Gumbel densities (Grego, Yates, & Mai, 2014?). It is a particular case of the generalized extreme value distribution, the distribution of choice for hydrologists in the United Kingdom for flood frequency analysis.

Other Work and Future Work

Other Work: All of these methods have been used with a finite mixture of two Gumbel densities (Grego, Yates, & Mai, 2014?). It is a particular case of the generalized extreme value distribution, the distribution of choice for hydrologists in the United Kingdom for flood frequency analysis.

Future Work: There are a few potential pieces of this research that can be extended.

Other Work and Future Work

Other Work: All of these methods have been used with a finite mixture of two Gumbel densities (Grego, Yates, & Mai, 2014?). It is a particular case of the generalized extreme value distribution, the distribution of choice for hydrologists in the United Kingdom for flood frequency analysis.

Future Work: There are a few potential pieces of this research that can be extended.

- The order statistics approach could be adapted to more than one observed historical flood. The problem? The joint probability density function of the observed historical floods would need to be included in the likelihood.

Other Work and Future Work

Other Work: All of these methods have been used with a finite mixture of two Gumbel densities (Grego, Yates, & Mai, 2014?). It is a particular case of the generalized extreme value distribution, the distribution of choice for hydrologists in the United Kingdom for flood frequency analysis.

Future Work: There are a few potential pieces of this research that can be extended.

- The order statistics approach could be adapted to more than one observed historical flood. The problem? The joint probability density function of the observed historical floods would need to be included in the likelihood.
- The Type I censoring and the order statistics approaches assuming an historical period of known length M flood year. Typically, this length of the period is not well-defined. Can either of these approaches be extended by treating the length of the historical record as another unknown quantity to be estimated?