Is it a Baby or a Bathtub? & How Many Fish? 
Two Studies in Applied Computing

Stanford University Department of 
Electrical Engineering
Computer Systems Colloquium (EE380)

February 8, 2006

Rose Ray, Ph.D. 
Principal Scientist
Exponent, Inc.
Outline

• Case 1: Is it a Baby or a Bathtub?
  – Problem Description
  – Available Data
  – Statistical Failure Analysis
  – Hazard Functions
  – Calculation Methods
    • Parametric
    • Life table
  – Results

• Case 2: How Many Fish?
  – Problem Description
  – Available Data
  – Statistical Method
  – Monte Carlo Simulation
  – Results

• Questions
Case 1: Problem Description

• A new design of battery powered small appliance is introduced in the fall.
  – Initial sales are primarily for the holiday gift market
  – In January the manufacturer begins to receive complaints of overheating batteries
  – Incident dates are as early as Dec 25
  – Failures may pose a safety hazard

• Should the product be recalled?
Available Data

• For Each Reported Failure
  – Age at failure
  – Description of the failure mode (based upon analysis of returned product)
  – Sales and Production

• Product sales by Production Lot
  – Number Sold
  – Current age of Product
### Available Data: Failure Data

<table>
<thead>
<tr>
<th>Obs</th>
<th>Sales</th>
<th>Age at Incident</th>
<th>Age at Withdrawal</th>
<th>Incident Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2973</td>
<td>1</td>
<td>30</td>
<td>burn/hot</td>
</tr>
<tr>
<td>2</td>
<td>600</td>
<td>2</td>
<td>45</td>
<td>Leak</td>
</tr>
<tr>
<td>3</td>
<td>802</td>
<td>1</td>
<td>60</td>
<td>Leak</td>
</tr>
<tr>
<td>4</td>
<td>2560</td>
<td>5</td>
<td>75</td>
<td>Leak</td>
</tr>
<tr>
<td>5</td>
<td>2970</td>
<td>7</td>
<td>90</td>
<td>Hot</td>
</tr>
<tr>
<td>6</td>
<td>1518</td>
<td>25</td>
<td>105</td>
<td>burn/hot</td>
</tr>
<tr>
<td>8</td>
<td>2559</td>
<td>15</td>
<td>30</td>
<td>burn/hot</td>
</tr>
<tr>
<td>9</td>
<td>500</td>
<td>10</td>
<td>45</td>
<td>burn/hot</td>
</tr>
<tr>
<td>10</td>
<td>1519</td>
<td>14</td>
<td>60</td>
<td>burn/hot</td>
</tr>
<tr>
<td>11</td>
<td>2000</td>
<td>1</td>
<td>75</td>
<td>burn/hot</td>
</tr>
<tr>
<td>12</td>
<td>500</td>
<td>2</td>
<td>90</td>
<td>burn/hot</td>
</tr>
<tr>
<td>13</td>
<td>600</td>
<td>3</td>
<td>105</td>
<td>burn/hot</td>
</tr>
<tr>
<td>14</td>
<td>1000</td>
<td>2</td>
<td>90</td>
<td>burn/hot</td>
</tr>
</tbody>
</table>
### Available Data: Sales Data (Exposure Data)

<table>
<thead>
<tr>
<th>Obs</th>
<th>Sales</th>
<th>Current Age (Age at Withdrawal)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20,101</td>
<td>Total</td>
</tr>
<tr>
<td>2</td>
<td>5,532</td>
<td>Production Lot 6</td>
</tr>
<tr>
<td>3</td>
<td>1,100</td>
<td>Production Lot 5</td>
</tr>
<tr>
<td>4</td>
<td>2,321</td>
<td>Production Lot 4</td>
</tr>
<tr>
<td>5</td>
<td>4,560</td>
<td>Production Lot 3</td>
</tr>
<tr>
<td>6</td>
<td>4,470</td>
<td>Production Lot 2</td>
</tr>
<tr>
<td>7</td>
<td>2,118</td>
<td>Production Lot 1</td>
</tr>
</tbody>
</table>
Statistical Failure Analysis

• Probability distribution of time to failure is key information for determining appropriate course of action.

• Typical failure time distributions
  – Weibull
  – Exponential
  – Log Normal

• \( F(t) = \text{probability of failure at or before time } t \)
Hazard Function: $H(t)$

- $H(t) = \text{probability of failure at time } t \text{ conditional on survival to time } t$.
- $H(t) = \frac{dF(t)}{1-F(t)}$
- $H(t) = \text{failure rate at time } t$

- **Typical hazard patterns**
  - Infant mortality (burn in)
  - Constant (memory less)
  - Wear out
  - Bathtub
Hazard Patterns

Hazard Function

Age

Infant Mortality
Hazard Patterns

- Hazard Function
- Constant

Age:
- 0
- 1
- 5
- 3
- 0
- 4
- 5
- 6
- 0
- 7
- 5
- 75
Hazard Patterns

- Hazard Function
- Infant Mortality
- Wear Out

Graph showing two hazard patterns: one decreasing with age (Infant Mortality) and another increasing with age (Wear Out).
Hazard Patterns

- Infant Mortality
- Bathtub
- Wear Out
Calculation Methods: Parametric

• Example: Weibull Distribution
  – Family of Failure time distributions
  – Can model both infant mortality or wear out.
  – Single Weibull cannot model both

• Hazard Function
  – $H(t) = \left( \frac{\alpha}{\beta} \right) \left( \frac{t}{\beta} \right)^{(\alpha-1)}$
    - $\alpha < 1$  ➤ decreasing failure rate
    - $\alpha = 1$  ➤ constant failure rate
    - $\alpha > 1$  ➤ increasing failure rate
## Estimated Weibull Parameters for Case Example

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>95% Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weibull Scale (β)</td>
<td>1</td>
<td>5.54E+10</td>
<td>3.15E+11</td>
<td>793717.7</td>
</tr>
<tr>
<td>Weibull Shape (α)</td>
<td>1</td>
<td>0.3563</td>
<td>0.0976</td>
<td>0.2083</td>
</tr>
</tbody>
</table>

- Indicates decreasing failure rate
- Upper Bound <1 indicates shape is statistically significantly <1
Is it a Baby or a Bathtub?

• If truly an infant mortality failure mode, all or most all failures may have occurred by the time complaints have been received.
  – No recall action is needed.

• If failure rate is constant or increasing with time, more failures are expected
  – Recall may be necessary
Is it a Baby or a Bathtub?

- Weibull estimate of decreasing failure rate depends very strongly on the assumption of Weibull form of the hazard function.

- At early age Weibull estimates may indicate a decreasing failure rate and miss the start of late age failures.

- Lifetable method can be used to verify.
Non Parametric Method: Lifetable

- Calculate hazard function directly with no assumption about shape of time to failure distribution
  - Partition data into short age intervals
  - Calculate number
    - At risk at start of interval
    - Failed during interval
    - Withdrawn during interval
  - Hazard for interval =
    Failures during interval / (At risk at start - .5 withdrawn)

- Compare estimated hazard at intervals toward the end of experience to determine whether infant mortality is the only failure pattern
### Case Example: Lifetable using SAS

<table>
<thead>
<tr>
<th>Interval</th>
<th>Failed</th>
<th>Number</th>
<th>Number Withdrawn During Interval</th>
<th>Effective Sample Size</th>
<th>Conditional Probability of Failure</th>
<th>Conditional Probability Standard Error</th>
<th>Survival</th>
<th>Failure</th>
<th>Survival Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0, 10)</td>
<td>9</td>
<td>10</td>
<td>0</td>
<td>20,114</td>
<td>0.045%</td>
<td>0.015%</td>
<td>100.000%</td>
<td>0.000%</td>
<td>0.000%</td>
</tr>
<tr>
<td>[10, 20)</td>
<td>3</td>
<td>20</td>
<td>0</td>
<td>20,105</td>
<td>0.015%</td>
<td>0.009%</td>
<td>99.960%</td>
<td>0.045%</td>
<td>0.015%</td>
</tr>
<tr>
<td>[20, 30)</td>
<td>1</td>
<td>30</td>
<td>0</td>
<td>20,102</td>
<td>0.005%</td>
<td>0.005%</td>
<td>99.940%</td>
<td>0.060%</td>
<td>0.017%</td>
</tr>
<tr>
<td>[30, 40)</td>
<td>0</td>
<td>40</td>
<td>0</td>
<td>5,532</td>
<td>0.000%</td>
<td>0.000%</td>
<td>99.940%</td>
<td>0.065%</td>
<td>0.018%</td>
</tr>
<tr>
<td>[40, 50)</td>
<td>0</td>
<td>50</td>
<td>0</td>
<td>1,100</td>
<td>0.000%</td>
<td>0.000%</td>
<td>99.940%</td>
<td>0.065%</td>
<td>0.018%</td>
</tr>
<tr>
<td>[50, 60)</td>
<td>0</td>
<td>60</td>
<td>0</td>
<td>13,469</td>
<td>0.000%</td>
<td>0.000%</td>
<td>99.940%</td>
<td>0.065%</td>
<td>0.018%</td>
</tr>
<tr>
<td>[60, 70)</td>
<td>0</td>
<td>70</td>
<td>0</td>
<td>2,321</td>
<td>0.000%</td>
<td>0.000%</td>
<td>99.940%</td>
<td>0.065%</td>
<td>0.018%</td>
</tr>
<tr>
<td>[70, 80)</td>
<td>0</td>
<td>80</td>
<td>0</td>
<td>4,560</td>
<td>0.000%</td>
<td>0.000%</td>
<td>99.940%</td>
<td>0.065%</td>
<td>0.018%</td>
</tr>
<tr>
<td>[80, 90)</td>
<td>0</td>
<td>90</td>
<td>0</td>
<td>6,588</td>
<td>0.000%</td>
<td>0.000%</td>
<td>99.940%</td>
<td>0.065%</td>
<td>0.018%</td>
</tr>
<tr>
<td>[90, 100]</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>4,470</td>
<td>0.000%</td>
<td>0.000%</td>
<td>99.940%</td>
<td>0.065%</td>
<td>0.018%</td>
</tr>
<tr>
<td>[100, ]</td>
<td>0</td>
<td>.</td>
<td>0</td>
<td>2,118</td>
<td>0.000%</td>
<td>0.000%</td>
<td>99.940%</td>
<td>0.065%</td>
<td>0.018%</td>
</tr>
</tbody>
</table>
Case Study: Hazard Function

Baby or Bathtub
Lifetable Analysis

Hazard Function

age
Case Study 1: Results

• Very high rate of failure at first usage or very early age results in Weibull estimates indicating decreasing rate
• No apparent failures observed after 30 days of use is strong indicator of infant mortality failure.
  – No recall is required
  – Changes to manufacturing and Quality Control process implemented to remove infant mortality failures prior to shipment
Case Study 2: How many fish?

• Problem Description:
  – Samples of fish captured in an urban river indicate levels of PCBs, metals and pesticide making them generally unsafe for human consumption
  – Despite posted warnings not to eat the fish, recreational angling occurs.
  – What is the human health risk associated with this activity?
    • How many people eat the fish
    • How much is eaten per year
    • Is it enough to warrant dredging of the river?

• The area is sparsely used and fishing permits are not required.
  – There is no reasonable way to design a random sample of the angler population
Urban Fishing Site
Data Available

• A typical creel-angler survey is conducted by selecting an interview day, then moving up and down the fishing area and interviewing anglers as they are encountered by the survey team.
  – A single interview day may not adequately capture seasonal fishing activity.
  – In a sparsely used area, it is likely that no anglers will be observed on the selected interview day.

• Consequently, a multiple-interview, longitudinal survey of anglers was conducted.
  – The survey is based upon a random sample of days across a 12 month period.
  – The survey is stratified by month and weekends vs. weekdays
A New Statistical Procedure

- A statistical procedure was designed to estimate angler activities based on the survey data and to calculate exposure factors necessary for illustrating the fish consumption pathway of recreational anglers in a human health risk assessment for the river.

- The estimates would be suitable for EPA-type risk assessments.
The method is based upon transforming the random sample of days to a probability sample of anglers by estimating the probability an angler would be interviewed at some time during the interview process.

Information collected during the interviews is used to estimate the number of days fished in each month during the year and combined with the number of days that interviews were conducted to estimate the probability that each angler like the observed angler would be interviewed at least once.
New Statistical Procedure (continued)

• The overall sampling weight = \( \frac{1}{\text{probability of being interviewed at least once}} \)

• The overall sampling weights are used to estimate:
  – Number of anglers
  – Number of consuming anglers
  – Mean, median & 95\textsuperscript{th} percentile of fish consumption.
Monte Carlo Simulation

- Computer simulation of stochastic process
- Can be used to:
  - Estimate population parameters when analytic methods are not available
  - Estimate probabilities from complicated processes
  - Validate estimation techniques
  - Calculate the sampling error of estimates
Monte Carlo Simulation

• Monte Carlo simulation was developed to test the new statistical method for the angler survey.

• This allowed researchers to generate simulated angler populations of varying sizes and fishing characteristics and to test the estimation procedure against the known populations produced by the simulation.

• The simulation results could then be used to demonstrate the validity of the estimation procedure and to provide confidence bound estimates of fish consumption.
The Simulation Program

- The simulation program has a Windows interface that allows the researcher to set parameters by entering values in text boxes.

- After entering parameters, the researcher selects the desired number of simulation runs and begins the simulation with a mouse click.
Simulation Program Screen Shot
Simulation Output

• As a result of the simulation, the program outputs comma delimited files that can be used by the researcher to test the angler estimation procedure.
Two Parts to a Simulation Run

• Randomly generate an angler population and subsequent fishing results for one year.

• Obtain information from anglers in simulated “interviews” taken on randomly selected days.

• The simulated interviews obtain the same type of information from the simulated anglers as the interviews in the planned multi-day survey.
Simulation Parameters

• **Population size:** Number of anglers in the population

• **P(anglers on weekends):** Fraction of anglers who fish on weekends, with the remainder fishing on weekdays.

• **P(angler fishes on given day):** Probability that a given angler fishes on a given non-winter day, stratified by weekend/weekday.

• **Winter:** Anglers fish less during the winter, so, on winter days (December, January, February, March), the chance that an angler fishes is reduced by 50%.
**Example**

- If population size = 200, \( P(\text{anglers on weekends}) = .4 \), and \( P(\text{angler fishes on given day}) = .3 \), then on each non-winter weekend day, the program will randomly select \( 200 \times .4 \times .3 = 24 \) weekend anglers who will fish on that day.

- On each non-winter weekday, the program will randomly select \( 200 \times .6 \times .3 = 36 \) weekday anglers who will fish on that day.

- On each winter weekend day, \( P(\text{angler fishes}) = .5 \times .3 = .15 \), and so the program will select \( 200 \times .4 \times .15 = 12 \) weekend anglers who will fish on that day, etc.
More Parameters

• Skill Level:
  – On each fishing day, an angler catches a random number of fish according to the Poisson probability distribution.
  – Before the simulation is run, 50% of anglers are randomly selected to be “good” anglers and 50% to be “poor” anglers. The researcher selects the Poisson parameter $\lambda$ for good anglers, which determines the average number of fish per day caught by good anglers.
  – The average number of fish per day caught by poor anglers = $0.5\lambda$. 

More Parameters

- **P(angler eats fish):** Fraction of the angler population who eat the fish they catch.
  
  - For example, if population size = 500, and P(angler eats fish) = .2, then, before the simulation is run, the program will randomly select 500x.2 = 100 fish-eating anglers.
Interview Days

• **P(weekend interview day):** Fraction of weekend days per month on which simulated angler interviews are conducted.
  – We assign 8 weekend days per month, so, for example, if $P(\text{weekend interview day}) = 0.5$, the program will randomly select $8 \times 0.5 = 4$ weekend interview days each month for conducting interviews.

• **P(weekday interview day):** Fraction of weekdays per month on which simulated angler interviews are conducted.
  – There are 23 weekdays per 31 day month, so, for example, if $P(\text{weekday interview day}) = 0.4$, the program will randomly select $23 \times 0.4 = 9$ weekday interview days each month.
Running the Simulation

• After parameters are set, the researcher selects the number of simulation runs, n, and clicks the “simulate n years” command.

• The program then generates the angler population according to the specified parameters and runs the simulation as follows:
Running the Simulation

• For each of the n simulated years:

  – Interview days are randomly selected.

  – Fishing days are randomly selected for each angler.

  – Each angler catches a random number of fish on selected fishing days.

  – Angler interviews are conducted.

  – Interview data is stored in two .csv files.
# Interview Data

A row in the interview data file has the following information:

- **Simulation #**: Identifies simulation run (simulated year).
- **Angler #**: Identifies angler
- **Month #**: Identifies month
- **Interviews**: Number of interviews with each angler that month.
- **Fish Caught**: Total number of fish caught by each angler during month on days when angler was interviewed (-1 if angler wasn’t interviewed).
Interview Data (continued)

- **Days Fished Last Month**: Total number of days fished in the previous month for each interviewed angler (-1 if angler wasn’t interviewed).

- **Eats Fish**: 1 if fish-eating angler, 0 if not.

- **Weekends**: 1 if weekend angler, 0 if not.

- **Skill Level**: 1 if good angler, 0 if poor angler.
More Interview Data

• “Months Fished” File: specifies in which months each angler fished during a simulated year, along with indicators of whether the angler was interviewed, consumer, and total fish caught. Each row specifies:

  – **Simulation#:** Identifies simulation run (year)
  – **Angler#:** Identifies angler
  – **Months Fished:** For each month: 1 if fished, 0 if not.
  – **Ever Interviewed:** 1 if angler ever interviewed during year, 0 if not
  – **Eats Fish:** 1 if fish-eating angler, 0 if not.
  – **Total Fish Caught:** Total fish caught in year by angler.
  – **Total Days Fished:** Total days fished in year by angler.
  – **Fish Caught on Random Day:** Number of fish caught on yearly random interview day.
Example Simulation Results

Population size = 385  
Good Angler Lambda = 3  
Poor Angler Lambda = 1.5  
P(angler fishes on given day) = .1  
P(angler eats fish) = .09  
P(angler is a weekend angler) = .73  
P(weekday interview day) = .2  
P(weekend interview day) = .7

<table>
<thead>
<tr>
<th></th>
<th>Average Estimate</th>
<th>Population</th>
<th>Root Mean Square Error</th>
<th>Relative Root MSE</th>
<th>Bias</th>
<th>Relative Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Anglers</td>
<td>387.7</td>
<td>385.0</td>
<td>3.2</td>
<td>0.8%</td>
<td>2.69</td>
<td>0.7%</td>
</tr>
<tr>
<td>Number of Consumers</td>
<td>38.9</td>
<td>39.0</td>
<td>0.8</td>
<td>2.0%</td>
<td>-0.06</td>
<td>-0.2%</td>
</tr>
<tr>
<td>Average Fish Consumption</td>
<td>2.8</td>
<td>2.8</td>
<td>0.3</td>
<td>11.0%</td>
<td>-0.01</td>
<td>-0.3%</td>
</tr>
<tr>
<td>Median Fish Consumption</td>
<td>0.0</td>
<td>0.0</td>
<td>0</td>
<td>0.0</td>
<td>0.00</td>
<td>0.0%</td>
</tr>
<tr>
<td>95th Percentile Fish consumption</td>
<td>22.2</td>
<td>22.5</td>
<td>3.1</td>
<td>13.8%</td>
<td>-0.26</td>
<td>-1.2%</td>
</tr>
</tbody>
</table>
Accuracy of Estimates from Multi-Interview Creel Angler Survey
Relative MSE vs Angler Avidity

Results
Additional Research

• Additional exploration of simulation parameters
  – Changes in amount of fish caught ($\lambda$)
  – Add variation in fishing frequency for anglers in the same simulation

• Make the simulation more realistic
  – Add multiple types of fish
  – Add seasonal variation to the types of fish
  – Add variation in the size of fish
  – Add variation to parts of fish consumed
Acknowledgements

• Dr. Michael Orkin, Managing Scientist, Exponent: for creation of the Monte Carlo Simulation and help with preparation of this lecture

• Sandra Thomas, Research Specialist, Exponent: for her editorial expertise
Questions?