Incomplete Data

Incomplete Data

- Incomplete Data in insurance is a common problem and prevalent in basically all data sets.
- Can be incomplete from missing data or lack of filed claims

Types of Incomplete Data

- Data can be incomplete from many different reasons. From simply missing data, censored data or truncated data.
  - Incomplete data from missing data is caused by data sets simply missing values.
  - Incomplete data is considered censored when the number of values in a set are known, but the values themselves are unknown.
  - Incomplete data is said to be truncated when there are values in a set that are excluded.

Truncation

- Two main types of truncation
  - Data is said to be truncated from below when the set of missing data is all the values below a specific value in the set.
  - Data is said to be truncated from above when the set of missing data is all the values above a specific value in the set.

Insurance Truncation

- Insurance data sets are often truncated due to multiple reasons
  - Deductibles
  - Total loss limits

Models for Incomplete Data

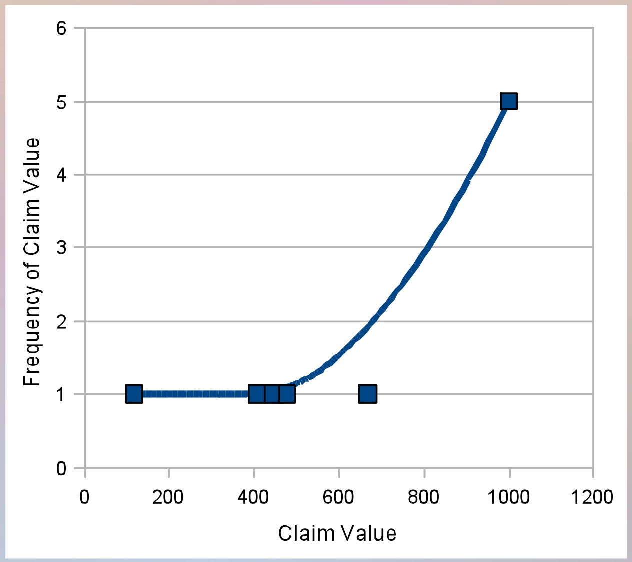
- Many different models are used to estimate distributions containing incomplete data. A few are:
  - Shifted Models
  - Maximum Likelihood Estimations
  - The Expectation Maximization Algorithm

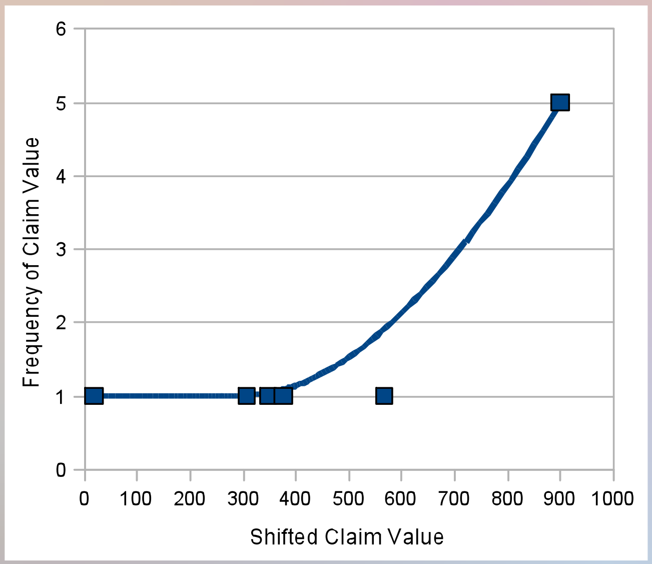
Shifted Models

- Simply shift the data set to estimate the missing values
  - This is a very inaccurate way to estimate incomplete data
  - Only takes into account current data and does not look at what values the missing data could actually be

Shifted Models

- Using data set: {117, 407, 446, 476, 667, 1,000, 1,000, 1,000, 1,000, 1,000} as our set of claims
- Shifted data set could be: {17, 307, 346, 376, 567, 900, 900, 900, 900, 900}





MLE

- To look at the next two models we have to consider maximum likelihoods
- MLE is the Maximum Likelihood Estimate

MLE

Let $X=\{X_1,\ldots\ldots.,X_n\}$ be a random vector and

$\{f_X(x|\theta):\theta \in \Theta\}$

this model is parameterized by $\theta=\{\theta_1,\ldots\ldots.,\theta_n)$, which is the parameter vector in the parameter space Θ.

So the Likelihood Function is a map $L:\Theta \longrightarrow R$ given by

$L(\theta|x)=f_X(x|\theta)$

MLE

The parameter vector θ such that

L(θ)≥L(θ) for all θЄΘ

is called a Maximum Likelihood of θ

- Now using this Maximum Likelihood Estimate we can gain a model of our data from the previous example

MLE Example

- We will use the shifted values we created from the shifted model: {17, 307, 346, 376, 567, 900, 900, 900, 900, 900}
- Random variables X and Y will be used
  - X is the amount of loss or the ground-up loss variable
  - Y is the amount paid per claim

MLE Example Continued

undefined, $X \leq 100$

Y={ $X - 100$, $100 < X \leq 900$

900, $X > 900$

- Now using the MLE we can get distribution and density functions to use later in our model

MLE Example Continued

- Distribution Function for Y is:

  0, $y = 0$

- $F_Y(y) = \{F_x(y + 100) - F_x(100),\ 0 < y < 900$

  1, $y \geq 900$

- Density Funtion for Y is:

$\dfrac{f_X(y + 100)}{1 - F_X(100)}$, $0 \leq y < 900$

- $f_Y(y) = \{ \dfrac{1 - F_X(1{,}000)}{1 - F_X(100)}$, $y = 900$

  0, $y > 900$

MLE Example Continued

- Now the MLE is used to estimate the values of the truncated data
- The Weibull Probability Distribution Function is used to estimate the parameters of the function
- The Weibull Distribution is defined by:

$f(x) = \Gamma(x/\theta)^{\Gamma} e^{-(x/\theta)^{\Gamma}}$, and

x

$F(x) = 1 - e^{-(x/\theta)^{\Gamma}}$

Weibull Parameter Estimates

- The part of the likelihood function given by the 5 values at the upper limit of our set is:

$$f_Y(900) = \frac{1 - F_X(1{,}000)}{1 - F_X(100)} = \frac{e^{[-(1{,}000/\theta)^{\Gamma}]}}{e^{[-(100/\theta)^{\Gamma}]}}$$

- The part of the likelihood function from the values below the limit is given by:

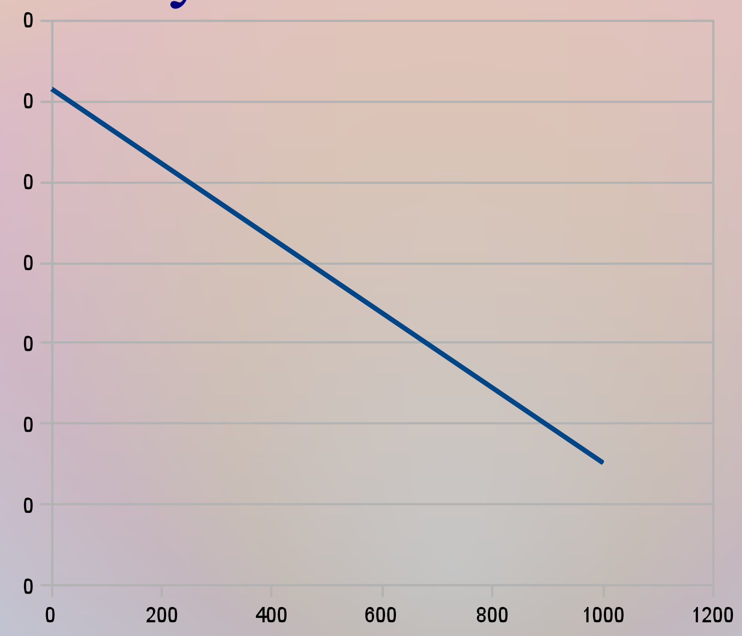
$$f_Y(x) = f_X(x+100) = \frac{\dfrac{\Gamma(x+100)^{\Gamma - 1}}{\theta^{\Gamma}} e^{[-((x+100)/\theta)^{\Gamma}]}}{e^{[-(100/\theta)^{\Gamma}]}}$$

Weibull Parameter Estimates

- Now the Simplex Method is used to get maximum values for the parameters of the Weibull Distribution
  - The Simplex method is an iterative method used in maximum likelihood estimations to select the value that will give the largest change toward the minimum or maximum solutions
- Using the Simplex Method the estimates for our parameters are:
  - $\theta = 1{,}199.09$
  - $\Gamma = 0.700744$

MLE Example

- With the parameter estimates we now have we can plug in any value for x to estimate the probability of that x value for any value below the upper limit



MLE Example

- The probability at the upper limit of 1,000 is calculated using the likelihood function for that limit.
  - P = 49.40377%

Expectation Maximization Algorithm

- A much simpler model to use to get more accurate results for larger sets of data
- A two step process to iteratively calculate the probabilities of all possible values in the total set, including missing values from our given set
- Uses a log likelihood function instead of simply an MLE

Two Steps

- The two steps of the Expectation Maximization Algorithm are the E-step and the M-step
  - The E-step is the expectation step and estimates the missing data given the observed data and the current estimate of the model parameters using the conditional expectation
  - The M-step maximizes the likelihood function assuming the estimates from the E-step

EM Algorithm

- The EM Algorithm is derived from the fact that for any probability distribution Q(z):

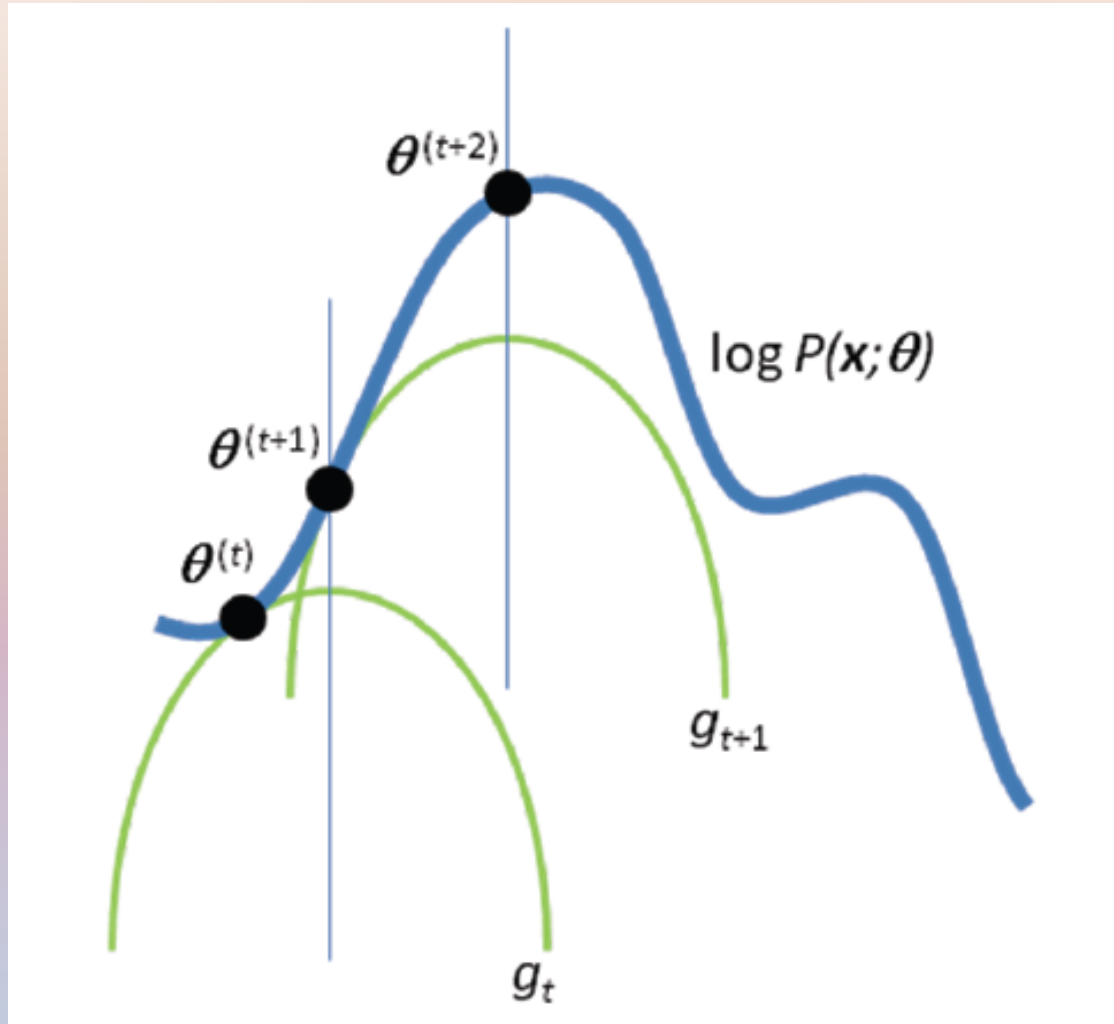
$$\log\left(\sum_z P(x,z;\theta)\right) = \log\left(\sum_z Q(z)\cdot\frac{P(x,z;\theta)}{Q(z)}\right) \geq \sum_z Q(z)\log\left(\frac{P(x,z;\theta)}{Q(z)}\right).$$

- Now to update our estimate of θ we get:

$$\hat{\theta}^{(t+1)} = \arg\max_{\theta} g_t(\theta)$$

- Where:

$$g_t(\theta) = \sum_z P\left(z \middle| x;\hat{\theta}^{(t)}\right)\log\left(\frac{P(x,z;\theta)}{P\left(z \middle| x;\hat{\theta}^{(t)}\right)}\right).$$

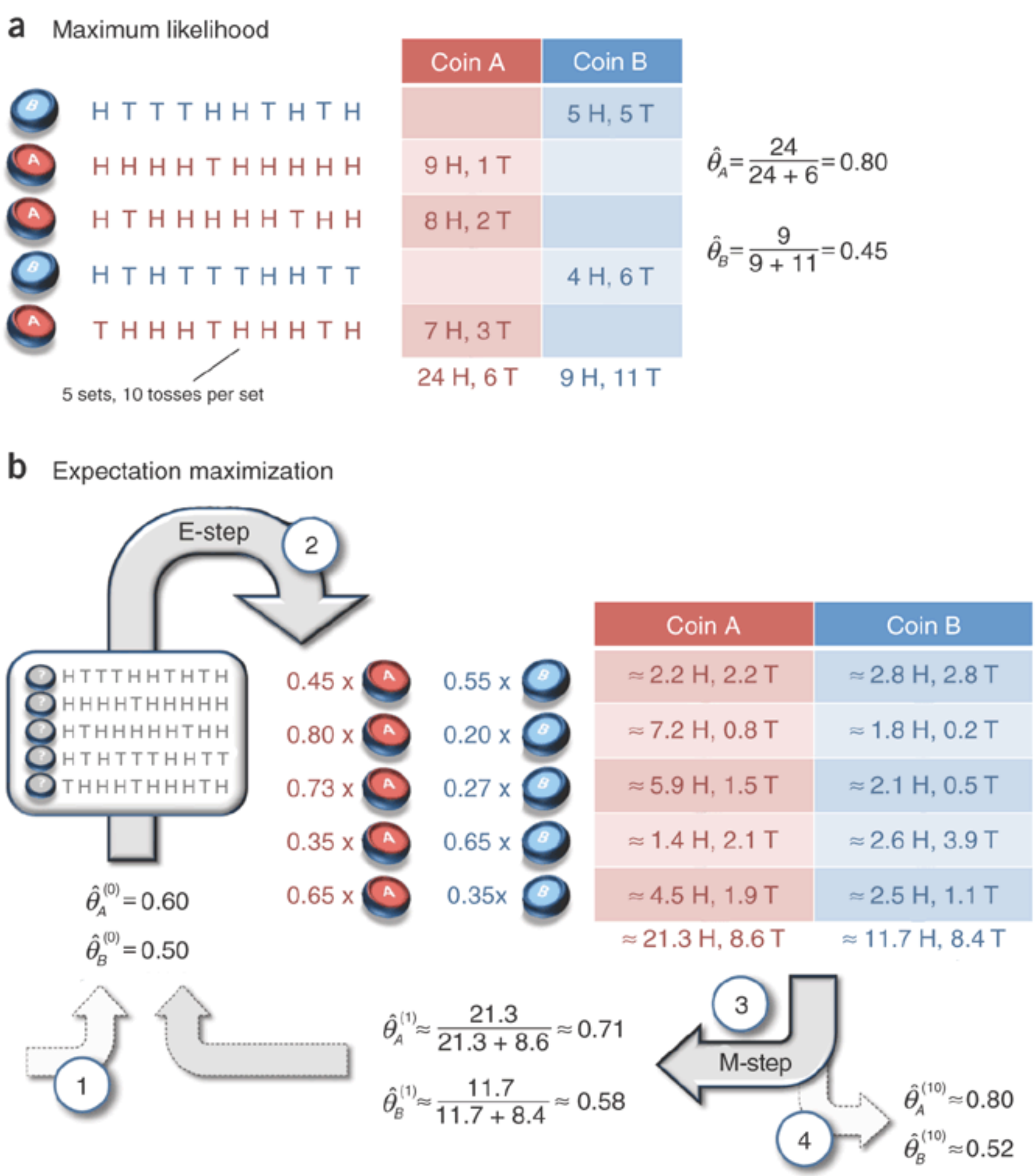


$\theta^{(t+2)}$
$\theta^{(t+1)}$
$\theta^{(t)}$
$\log P(\mathbf{x};\theta)$
g_{t+1}
g_t

EM Algorithm

- Essentially this means that for each new maximum value of θ we have:

$$\hat{\theta}^{(t+1)} = \arg\max_{\theta} \sum_{z} P\left(z \middle| x; \hat{\theta}^{(t)}\right) \log P(x, z; \theta).$$



a Maximum likelihood
Coin A
Coin B
H T T T H H T H T H
H H H H T H H H H H
H T H H H H H T H H
H T H T T T H H T T
T H H H T H H H T H
5 H, 5 T
9 H, 1 T
8 H, 2 T
4 H, 6 T
7 H, 3 T
24 H, 6 T
9 H, 11 T
5 sets, 10 tosses per set
$\hat{\theta}_A = \frac{24}{24+6} = 0.80$
$\hat{\theta}_B = \frac{9}{9+11} = 0.45$
b Expectation maximization
E-step
2
HTTTHHTHTH
HHHHTHHHHH
HTHHHHHTHH
HTHTTTHHTT
THHHTHHHTH
$\hat{\theta}_A^{(0)} = 0.60$
$\hat{\theta}_B^{(0)} = 0.50$
0.45 x
0.55 x
0.80 x
0.20 x
0.73 x
0.27 x
0.35 x
0.65 x
0.65 x
0.35x
Coin A
Coin B
≈ 2.2 H, 2.2 T
≈ 2.8 H, 2.8 T
≈ 7.2 H, 0.8 T
≈ 1.8 H, 0.2 T
≈ 5.9 H, 1.5 T
≈ 2.1 H, 0.5 T
≈ 1.4 H, 2.1 T
≈ 2.6 H, 3.9 T
≈ 4.5 H, 1.9 T
≈ 2.5 H, 1.1 T
≈ 21.3 H, 8.6 T
≈ 11.7 H, 8.4 T
1
$\hat{\theta}_A^{(1)} \approx \frac{21.3}{21.3+8.6} \approx 0.71$
$\hat{\theta}_B^{(1)} \approx \frac{11.7}{11.7+8.4} \approx 0.58$
3
M-step
4
$\hat{\theta}_A^{(10)} \approx 0.80$
$\hat{\theta}_B^{(10)} \approx 0.52$

Summary

- Shifted Models
  - Primitive and very inaccurate
- MLE's
  - Still relatively basic and primarily only takes into account observed data and ends with an estimate for all data including unobserved data
- EM Algorithm
  - Most productive two step model to estimate pdf's using an estimate of missing data and maximizing probabilities of all data in range