Incomplete Data **30M**

In previous examples, we were given complete data, i.e. the exact values of the full dataset. However, in real life, we often have to work with incomplete information, i.e. not knowing the exact values of the observations or only having a partial dataset. Examples of these include truncation, censoring and grouping.

Truncated Data

Truncated data is an example of an incomplete dataset.

A dataset that is *left-truncated* or *truncated from below* at d means the dataset does not include data **at or below** d. Thus, we condition the likelihood of every observation in this dataset on being above d.

$$\Pr(X = x \mid X > d) = \frac{p(x)}{\Pr(X > d)} \qquad \text{(discrete)}$$

$$f(x \mid X > d) = \frac{f(x)}{\Pr(X > d)} \qquad \text{(continuous)}$$

A dataset that is *right-truncated* or *truncated from above* at u means that the dataset does not include data **at or above** u. Thus, we condition the likelihood of every observation in this dataset on being below u.

$$\Pr(X = x \mid X < u) = \frac{p(x)}{\Pr(X < u)} \qquad \text{(discrete)}$$

$$f(x \mid X < u) = \frac{f(x)}{\Pr(X < u)} \qquad \text{(continuous)}$$

The **left-truncated** case is commonly seen in insurance in the form of a **deductible**. When the loss amount is below the deductible, the policyholder won't receive any reimbursement for it, so the loss likely won't be reported. Thus, an insurer will only record a loss when it exceeds the deductible.

Example S3.1.2.1

You are given a sample of losses:

$$6 \quad 7 \quad 9 \quad 10$$

No information is available for losses of 5 or less.

Assume losses follow an exponential distribution with mean θ.

Determine the maximum likelihood estimate of θ.

Solution

The sample is left-truncated at 5. Thus, the likelihood of each loss is

$$\frac{f(x)}{S(5)}$$

Construct the likelihood function.

$$\begin{aligned} L(\theta) &= \frac{f(6)f(7)f(9)f(10)}{S(5)^4} \\ &= \frac{\frac{1}{\theta^4}e^{-(6+7+9+10)/\theta}}{\left(e^{-5/\theta}\right)^4} \\ &= \frac{1}{\theta^4}e^{-12/\theta} \end{aligned}$$

Then, calculate the log-likelihood function.

$$l(\theta) = -4\ln\theta - \frac{12}{\theta}$$

Take the first derivative of the log-likelihood function.

$$l'(\theta) = -\frac{4}{\theta} + \frac{12}{\theta^2}$$

Set the derivative equal to zero and solve for θ.

$$\begin{aligned} -\frac{4}{\theta} + \frac{12}{\theta^2} &= 0 \\ \hat{\theta} &= \frac{12}{4} \\ &= \mathbf{3} \end{aligned}$$

Censored Data

Censored data is an example of observations for which the exact values are not known.

We say an observation is *left-censored* or *censored from below* at d when the value is known to be **at most** d but is only recorded as d. The likelihood of the observation is

$$\Pr(X \leq d)$$

We say an observation is *right-censored* or *censored from above* at u when the value is known to be **at least** u but is only recorded as u. The likelihood of the observation is

$$\Pr(X \geq u)$$

The **right-censored** case is commonly seen in insurance whenever there is a **policy limit**. In this case, a policyholder can claim at most the policy limit, even when the loss actually exceeds the maximum covered loss (the point at which the policy limit is reached). Thus, for losses that exceed the maximum covered loss, records will likely only document them as the maximum covered loss. So while the exact amounts of the losses are not known, it is known that those losses are at least the maximum covered loss.

Example S3.1.2.2

For insurance coverage with a policy limit of 10:

- You observed the following claim payments:

$$2 \quad 3 \quad 5 \quad 8 \quad 10 \quad 10$$

- Claim sizes, X, follow a distribution with the following survival function:

$$S(x) = \frac{1}{x^{\alpha}}, \quad 1 < x < \infty, \ \alpha > 0$$

Calculate the maximum likelihood estimate of $\text{VaR}_{0.90}(X)$.

Solution

Note that there are two payments at the policy limit, 10. These payments can be the result of losses at or above 10. Thus, the likelihood for each of the payments should be

$$\Pr(X \geq 10)$$

Claim sizes range from 1 to infinity. The only distribution on the exam table with an infinite domain that does not begin at 0 is the single-parameter Pareto. Compare the survival function to conclude that claim sizes follow a single-parameter Pareto distribution with parameters α and $\theta = 1$.

$$X \sim \text{S-P Pareto}\,(\alpha, 1)$$

One property of maximum likelihood estimation is that we can substitute the MLE of a parameter into any function of that parameter to compute the MLE of that function.

$$\hat{g}(\theta) = g\left(\hat{\theta}\right)$$

Thus, from the exam table, the MLE of the 90th percentile is

$$\widehat{\text{VaR}}_{0.90}(X) = \theta(1-0.90)^{-1/\hat{\alpha}} = 0.1^{-1/\hat{\alpha}}$$

where $\hat{\alpha}$ is the MLE of α.

We need to estimate α. Start by constructing the likelihood function. The likelihoods of the first four data points are found using the PDF since the exact values are known. For the last two data points, we only know that they are at least 10. Thus, use the survival function evaluated at 10 for their likelihoods.

$$\begin{aligned} L(\alpha) &= f(2) \cdot f(3) \cdot f(5) \cdot f(8) \cdot S(10)^2 \\ &= \frac{\alpha}{2^{\alpha+1}} \cdot \frac{\alpha}{3^{\alpha+1}} \cdot \frac{\alpha}{5^{\alpha+1}} \cdot \frac{\alpha}{8^{\alpha+1}} \cdot \left(\frac{1}{10^{\alpha}}\right)^2 \\ &= \frac{\alpha^4}{240^{\alpha+1}100^{\alpha}} \end{aligned}$$

The log-likelihood function is

$$l(\alpha) = 4\ln\alpha - (\alpha+1)\ln 240 - \alpha \ln 100$$

Take the derivative of the log-likelihood function with respect to α.

$$l'(\alpha) = \frac{4}{\alpha} - \ln 240 - \ln 100$$

Then, set it equal to 0 to solve for α.

$$\begin{aligned} \frac{4}{\alpha} - \ln 240 - \ln 100 &= 0 \\ \hat{\alpha} &= \frac{4}{\ln 240 + \ln 100} \\ &= 0.3966 \end{aligned}$$

Finally, calculate the final answer.

$$\widehat{\text{VaR}}_{0.90}(X) = 0.1^{-1/0.3966} = \mathbf{332.2403}$$

Now, let's do a couple of examples that combine truncation and censoring.

Example S3.1.2.3

All losses follow a Weibull distribution with parameters $\theta = 1$ and τ.

An insurance company offers three types of policies. You are given:

- Policy A has no deductible and no policy limit.
- Policy B has a deductible of a and no policy limit.

- Policy C has a deductible of b and a policy limit of c.
- Two claims of sizes x_1 and x_2 are observed from Policy A.
- One claim of size x_3 is observed from Policy B.
- Three claims of sizes x_4, x_5, and $b + c$ are observed from Policy C.

You are to estimate the value of τ using the maximum likelihood method.

Determine the log-likelihood function, $l(\tau)$.

Solution

The likelihoods of the two claims from Policy A are found using only the PDF.

Policy B has a deductible of a. Thus, the claims observed are left-truncated at a. The likelihoods need to be conditioned on being greater than a.

$$f(x \mid X > a) = \frac{f(x)}{S(a)}$$

Policy C has a deductible of b. Thus, the claims observed are left-truncated at b. The likelihoods need to be conditioned on being greater than b.

$$f(x \mid X > b) = \frac{f(x)}{S(b)}$$

Additionally, there is one claim observed at the maximum covered loss. The likelihood is the survival function at the maximum covered loss, conditioned on being greater than b.

$$\Pr(X > b + c \mid X > b) = \frac{S(b+c)}{S(b)}$$

The likelihood function is

$$\begin{aligned} L(\tau) &= [f(x_1) \cdot f(x_2)] \cdot \left[\frac{f(x_3)}{S(a)}\right] \cdot \left[\frac{f(x_4) \cdot f(x_5) \cdot S(b+c)}{S(b)^3}\right] \\ &= \left(\prod_{i=1}^{5} \tau\, x_i^{\tau-1} e^{-x_i^\tau}\right) \cdot e^{-(b+c)^\tau} \cdot \left(e^{-a^\tau}\right)^{-1} \cdot \left(e^{-b^\tau}\right)^{-3} \\ &= \tau^5 \left(\prod_{i=1}^{5} x_i\right)^{\tau-1} e^{-\left(\sum_{i=1}^{5} x_i^\tau\right) - (b+c)^\tau + a^\tau + 3b^\tau} \end{aligned}$$

The log-likelihood function is

$$l(\tau) = 5 \ln \tau + (\tau - 1)\left(\sum_{i=1}^{5} \ln x_i\right) - \left(\sum_{i=1}^{5} x_i^\tau\right) - (b+c)^\tau + a^\tau \cdot$$

Example S3.1.2.4

You are given the following information about a group of policies:

Claim Payment	Deductible	Policy Limit
30	-	80
50	10	100
80	10	100
120	20	150
150	30	150

Assume payments at the policy limit resulted from losses above the maximum covered loss.

You are to fit a continuous distribution to the losses using the maximum likelihood method.

Determine the likelihood function.

Solution

Notice that we are given the claim **payments** and are asked for the likelihood function for the **loss** distribution. To translate payments into losses, simply add back the deductible. A similar adjustment is required to translate policy limits into maximum covered losses.

Then, calculate the likelihoods by evaluating the PDF at each loss amount and condition it on being greater than the deductible, if any.

Loss Amount	Deductible	Maximum Covered Loss	Likelihood
30	-	80	$f(30)$
60	10	110	$\frac{f(60)}{S(10)}$
90	10	110	$\frac{f(90)}{S(10)}$
140	20	170	$\frac{f(140)}{S(20)}$
180	30	180	$\frac{S(180)}{S(30)}$

The likelihood function is the product of the likelihoods.

$$L = f(30) \cdot \frac{f(60)}{S(10)} \cdot \frac{f(90)}{S(10)} \cdot \frac{f(140)}{S(20)} \cdot \frac{S(180)}{S(30)}$$

■

Grouped Data

Similar to censored data, grouped data is an example of observations for which the exact values are not known. Grouped data is presented as the number of observations in a distinct interval. The likelihood of an observation in the interval $(a, b]$ is

$$\begin{aligned}\Pr(a < X \leq b) &= F(b) - F(a) \\ &= S(a) - S(b)\end{aligned}$$

For discrete distributions, the likelihood of grouped data can be expressed as the sum of the PMFs evaluated at every value within the range.

$$\Pr(a \leq X \leq b) = p(a) + \ldots + p(b)$$

Example S3.1.2.5

The annual number of accidents per car follows a geometric distribution with parameter β where $\beta > 0$.

A sample of 20 cars were observed in the past year:

Number of Accidents	Number of Cars
0	9
[1, 2]	6
[3, 4]	5

Calculate the maximum likelihood estimate of β.

Solution

Let N be the annual number of accidents per car.

$$N \sim \text{Geometric } (\beta)$$

The likelihood of a car having 0 accidents is $p_N(0)$.

The likelihood of a car having 1 to 2 accidents is

$$\Pr(1 \leq N \leq 2) = p_N(1) + p_N(2)$$

The likelihood of a car having 3 to 4 accidents is

$$\Pr(3 \leq N \leq 4) = p_N(3) + p_N(4)$$

Then, the likelihood function is

$$\begin{aligned} L(\beta) &= p_N(0)^9 \cdot [p_N(1) + p_N(2)]^6 \cdot [p_N(3) + p_N(4)]^5 \\ &= \left(\frac{1}{1+\beta}\right)^9 \left[\frac{\beta}{(1+\beta)^2} + \frac{\beta^2}{(1+\beta)^3}\right]^6 \left[\frac{\beta^3}{(1+\beta)^4} + \frac{\beta^4}{(1+\beta)^5}\right]^5 \\ &= \left(\frac{1}{1+\beta}\right)^9 \left\{\left[\frac{\beta}{(1+\beta)^2}\right]^6 \left[1 + \frac{\beta}{1+\beta}\right]^6\right\} \left\{\left[\frac{\beta^3}{(1+\beta)^4}\right]^5 \left[1 - \right.\right. \\ &= \frac{\beta^{6+3(5)}}{(1+\beta)^{9+2(6)+4(5)}} \left[\frac{1+2\beta}{1+\beta}\right]^{6+5} \\ &= \frac{\beta^{21}(1+2\beta)^{11}}{(1+\beta)^{52}} \end{aligned}$$

The log-likelihood function is

$$l(\beta) = 21 \ln \beta + 11 \ln (1+2\beta) - 52 \ln (1+\beta)$$

Take the derivative of the log-likelihood function with respect to β.

$$l'(\beta) = \frac{21}{\beta} + \frac{22}{1+2\beta} - \frac{52}{1+\beta}$$

Then, set it equal to 0.

$$\begin{aligned} \frac{21}{\beta} + \frac{22}{1+2\beta} - \frac{52}{1+\beta} &= 0 \\ 21(1+2\beta)(1+\beta) + 22\beta(1+\beta) - 52\beta(1+2\beta) &= 0 \\ 21 + 63\beta + 42\beta^2 + 22\beta + 22\beta^2 - 52\beta - 104\beta^2 &= 0 \\ 40\beta^2 - 33\beta - 21 &= 0 \end{aligned}$$

Solve for β using the quadratic formula.

$$\begin{aligned} \hat{\beta} &= \frac{-(-33) \pm \sqrt{(-33)^2 - 4(40)(-21)}}{2(40)} \\ &= \mathbf{1.2463} \quad \text{or} \quad -0.4213 \end{aligned}$$

Since β can't be negative, the final answer is 1.2463.