Customer Lifetime Value

Team members :
Louiza AOUAOUCHE
Salim ABDOU DAOURA

Supervised by :
Mr Pascal BIANCHI

2023

Contents

1 Introduction

Customer Lifetime Value (CLV) is a key metric for companies looking to assess their long-term profitability. It can measure how much revenue a customer can generate for a company over the course of his or her lifetime as a customer. It can also synthesize churn risk of every single customer. Thus, CLV is an indispensable tool for companies that want to maximize their return on investment and develop effective marketing and loyalty strategies.

The concept of CLV emerged in the 1980s in response to a shift in marketing towards a customer relationship management (CRM) approach rather than simply acquiring new customers. Thus, companies began to realize the importance of maintaining and retaining their existing customer base rather than constantly trying to attract new customers.

In this project, we will mainly take a look at the CLV from a statistical point of view. We will also discuss the methods for calculating CLV such as Kaplan Meier estimator.

2 Parametric Estimators

2.1 Context

We observe n clients. Each client i survives for a duration of $T_i \geq 1$ month. Let a random sample of n iid random variables $T_1, \ldots, T_n$ following a distribution of unknown parameter θ. From T is defined Y such that:

$$Y_i = \min(T_i, a_i)$$

Where:

- $i \in \{1, \ldots, n\}$
- a_i is the duration from the arrival of client i to time of observation
- $\mathcal{A} := \{i \in \{1, 2, ..., n\} : Y_i < a_i\}$

To observe : not censored case : $Y_i = T_i$

2.2 Geometric distribution - not censored

2.2.1 Probability function

We assume that the n iid random variables $T_1, \ldots, T_n$ (see 2.1) follow a geometric distribution of unknown parameter θ. The probability function of the geometric distribution is given by:

$$f(k) = P(Y_i = k) = \theta(1-\theta)^{k-1} \quad \text{for } k = 1, 2, 3, \ldots$$

Where:

- θ the probability of success (customer churn)
- k the number of trials until the first success

2.2.2 Maximum Likelihood Estimator

The goal is to estimate the value of θ based on the observed data $y_1, y_2, \ldots, y_n$. To do this, we use the method of maximum likelihood. The likelihood function $L_\theta(y_1, \ldots, y_n)$ is given by:

$$L_\theta(y_1, \ldots, y_n) = \prod_{i=1}^{n} P(Y_i = y_i) = \prod_{i=1}^{n} \theta(1-\theta)^{y_i - 1} = \theta^n (1-\theta)^{\sum_{i=1}^{n}(y_i - 1)}$$

$$
\begin{aligned}
\ln(L_\theta(y_1,\dots,y_n)) &= \ln(\theta^n(1-\theta)^{\sum_{i=1}^n (y_i-1)}) \\
&= n\ln(\theta) + \ln(1-\theta)\sum_{i=1}^n (y_i-1) \\
\frac{\partial \ln(L_\theta(y_1,\dots,y_n))}{\partial\theta} &= \frac{n}{\theta} - \frac{\sum_{i=1}^n y_i - n}{1-\theta} \\
\frac{\partial \ln(L_\theta(y_1,\dots,y_n))}{\partial\theta} &= 0 \\
\frac{n}{\hat\theta} - \frac{\sum_{i=1}^n y_i - n}{1-\hat\theta} &= 0 \\
\hat\theta\left(\frac{\sum_{i=1}^n y_i - n + n}{\sum_{i=1}^n y_i - n}\right) &= \frac{n}{\sum_{i=1}^n y_i - n} \\
\hat\theta &= \frac{n}{\sum_{i=1}^n y_i} \\
\boxed{\frac{1}{\hat\theta} = \frac{\sum_{i=1}^n y_i}{n}}
\end{aligned}
$$

2.2.3 Expected value

Let's study the bias of the estimator $C\hat{L}V := \frac{1}{\hat\theta}$:

$$
\begin{aligned}
E\left[\frac{1}{\hat\theta}\right] &= E\left(\frac{\sum_{i=1}^n y_i}{n}\right) \\
&= \frac{\sum_{i=1}^n E(y_i)}{n} \\
&= \frac{\frac{n}{\theta}}{n} \\
&= \frac{1}{\theta}
\end{aligned}
$$

2.3 Geometric distribution - with censored data

2.3.1 Context

We assume that the n iid random variables $T_1,\dots,T_n$ of section 2.1) follow a geometric distribution of unknown parameter θ and we note $T \sim G(\theta)$ where $\theta \in [0,1]$ with uniform censored data. Here, we are in the case where:

$$Y_i = \min(T_i, a_i)$$

2.3.2 Probability function

Let's compute the probability function of Y. For $i \in \{1,\dots,n\}$:

$$
P_\theta(Y_i = y_i) = \begin{cases} (1-\theta)^{y_i-1}\theta & \text{if } y_i \le a_i \\ P_\theta(T_i > a_i) & \text{else} \end{cases}
$$

Where:

$$P_\theta(T_i > a_i) = \sum_{t=a_i+1}^{\infty} P_\theta(T_i = t) = \sum_{t=a_i+1}^{\infty} (1-\theta)^{t-1}\theta$$
$$= \frac{(1-\theta)^{a_i}}{1-(1-\theta)}\theta = (1-\theta)^{a_i}$$

The final expression of the probability function of Y in terms of θ is then given by:

$$P_\theta(Y_i = y_i) = \begin{cases} (1-\theta)^{y_i-1}\theta & \text{if } y_i \leq a_i \\ (1-\theta)^{a_i} & \text{if } y_i > a_i \end{cases}$$

$$\boxed{P_\theta(Y_i = y_i) = \left[(1-\theta)^{y_i-1}\theta\right]^{\mathbf{1}_{y_i \leq a_i}} \times \left[(1-\theta)^{a_i}\right]^{\mathbf{1}_{y_i > a_i}}}$$

2.3.3 Maximum Likelihood Estimator

To estimate the value of θ based on the observed data $\{y_1, \ldots, y_n\} \subset \mathbf{N}$, we use the method of maximum likelihood. The likelihood function $L_\theta(y_1, \ldots, y_n)$ is given by:

$$\begin{aligned} L_\theta(y_1, \ldots, y_n) &= P_\theta(Y_1 = y_1, \ldots, Y_n = y_n) \\ &\underset{\text{(iid)}}{=} \prod_{i=1}^{n} P(Y_i = y_i) \\ &= \prod_{i=1}^{n} \left[(1-\theta)^{y_i-1}\theta\right]^{\mathbf{1}_{y_i \leq a_i}} \times \left[(1-\theta)^{a_i}\right]^{\mathbf{1}_{y_i > a_i}} \\ \ln L_\theta(y_1, \ldots, y_n) &= \sum_{i=1}^{n} 1_{y_i \leq a_i} \ln((1-\theta)^{y_i-1}\theta) + \sum_{i=1}^{n} 1_{y_i > a_i} \ln((1-\theta)^{a_i}) \\ &= \sum_{i \in \mathcal{A}} \ln((1-\theta)^{y_i-1}\theta) + \sum_{i \notin \mathcal{A}} \ln((1-\theta)^{a_i}) \end{aligned}$$

Where:

- "Available data" set : $\mathcal{A} = \{i = 1, \ldots, n; y_i \leq a_i\}$

$$\begin{aligned} \ln L_\theta(y_1, \ldots, y_n) &= \sum_{i \in \mathcal{A}} \left((y_i - 1)\ln(1-\theta) + \ln(\theta)\right) + \sum_{i \in \mathcal{N}} a_i \ln(1-\theta) \\ &= |\mathcal{A}| \ln(\theta) + \ln(1-\theta)\left(\sum_{i \in \mathcal{A}} (y_i - 1) + \sum_{i \in \mathcal{N}} a_i\right) \end{aligned}$$

$$\frac{\partial \ln L_\theta(y_1, \ldots, y_n)}{\partial \theta} = \frac{|\mathcal{A}|}{\theta} - \frac{1}{1-\theta}\left(\sum_{i \in \mathcal{A}} (y_i - 1) + \sum_{i \in \mathcal{N}} a_i\right) = 0$$
$$(1-\theta)|\mathcal{A}| = \theta\left(\sum_{i \in \mathcal{A}} y_i - |\mathcal{A}| + \sum_{i \in \mathcal{N}} a_i\right)$$
$$|\mathcal{A}| = \theta\left(|\mathcal{A}| + \sum_{i \in \mathcal{A}} y_i - |\mathcal{A}| + \sum_{i \in \mathcal{N}} a_i\right)$$
$$\hat{\theta} = \frac{|\mathcal{A}|}{\sum_{i \in \mathcal{A}} y_i + \sum_{i \in \mathcal{N}} a_i}$$

$$\boxed{\frac{1}{\hat{\theta}} = \frac{\sum_{i\in\mathcal{A}} y_i + \sum_{i\in\mathcal{N}} a_i}{|\mathcal{A}|}}$$

2.4 Exponential distribution - not censored

In this case, the n iid random variables $T_1, \ldots, T_n$ (of section 2.1) follow an exponential distribution of unknown parameter θ without censored data.

2.4.1 Probability function

Let's compute the probability function of Y. For $i \in \{1, \ldots, n\}$, we have:

$$P_\theta(y_i) = \theta e^{-\theta y_i}$$

2.4.2 Maximum Likelihood Estimator

$$\begin{aligned}
\ln L_\theta(y_1, \ldots, y_n) &= \sum_{i=1}^{n} \ln\left(\theta e^{-\theta y_i}\right) \\
&= \sum_{i=1}^{n} \ln(\theta) - -\theta y_i) \\
&= n\ln(\theta) - \theta \sum_{i=1}^{n} y_i \\
\frac{\partial \ln L_\theta(y_1, \ldots, y_n)}{\partial\theta} &= \frac{n}{\hat{\theta}} - \sum_{i=1}^{n} y_i = 0
\end{aligned}$$

$$C\hat{L}V := \frac{1}{\hat{\theta}_n} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Observation: When the data follows a geometric or exponential distribution in the non-censored case, we obtain the same estimator.

2.5 Exponential distribution - censored

2.5.1 Context

To the previous case, we add uniformly distributed censored data and we recall the definition of random variable Y (of section 2.1) in such case :

$$Y_i = \min(T_i, a_i)$$

2.5.2 Probability function

Let's compute the probability function of Y. First, for $i \in \{1, \ldots, n\}$, we have:

$$P_\theta(y_i) = \begin{cases} \theta e^{-\theta y_i} & \text{if } y_i < a_i \\ P_\theta(Y_i = a_i) = e^{-\theta a_i} & \text{if } y_i = a_i \end{cases}$$

2.5.3 Maximum Likelihood Estimator

We define : $\mathcal{A} := \{i \in \{1, 2, ..., n\} : Y_i < a_i\}$

$$\begin{aligned}
\ln L_\theta(y_1, \ldots, y_n) &= \sum_{i \in \mathcal{A}} \ln\left(\theta e^{-\theta y_i}\right) + \sum_{i \notin \mathcal{A}} \ln\left(e^{-\theta y_i}\right) \\
&= \sum_{i \in \mathcal{A}} \ln(\theta) - \sum_{i=1}^{n} \theta y_i \\
&= |\mathcal{A}| \ln(\theta) - \theta \sum_{i=1}^{n} y_i \\
\frac{\partial \ln L_\theta(y_1, \ldots, y_n)}{\partial \theta} &= \frac{|\mathcal{A}|}{\hat{\theta}} - \sum_{i=1}^{n} y_i = 0
\end{aligned}$$

$$C\hat{L}V := \frac{1}{\hat{\theta}_n} = \frac{\sum_{i=1}^{n} y_i}{|\mathcal{A}|}$$

Observation: *Here also, with some transformations we can retrieve the geometric estimator.*

$$C\hat{L}V_{exp} := \frac{1}{\hat{\theta}_n} = \frac{\sum_{i=1}^{n} y_i \mathbf{1}_{Y_i<a_i} + y_i \mathbf{1}_{Y_i>=a_i}}{|\mathcal{A}|} = \frac{1}{\hat{\theta}} = \frac{\sum_{i \in \mathcal{A}} y_i + \sum_{i \in \mathcal{N}} a_i}{|\mathcal{A}|} = C\hat{L}V_{geom}$$

We pursue assuming the equivalence between both of the estimators.

Let's simplify the expression of $C\hat{L}V$:

- Denominator term:

$$\begin{aligned}
\frac{1}{n}|\mathcal{A}| &= \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{T_i<a_i} \\
&= \frac{1}{n} \sum_{i=1}^{n} \mathbf{E}(\mathbf{1}_{T_i<a_i}) + (\mathbf{1}_{T_i<a_i} - \mathbf{E}(1_{T_i<a_i})) \\
&= \frac{1}{n} \sum_{i=1}^{n} \mathbf{P}(T_i < a_i) + o_n(1) \\
&= \frac{1}{n} \sum_{i=1}^{n} 1 - e^{-\theta a_i} + o_n(1)
\end{aligned}$$

To show that $\frac{1}{n}\sum_{i=1}^{n}(\mathbf{1}_{T_i<a_i} - \mathbf{E}(1_{T_i<a_i}))$ is $o_n(1)$, we need to show that it converges to zero in probability, as $n \to \infty$. Using the Law of Large Numbers (LLN), we have that $\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{T_i<a_i}$ converges in probability to $\mathbf{E}(\mathbf{1}_{T_i<a_i})$, and likewise, $\frac{1}{n}\sum_{i=1}^{n} \mathbf{E}(1_{T_i<a_i})$ converges in probability to $\mathbf{E}(\mathbf{E}(1_{T_i<a_i}))$.

$$\frac{1}{n} \sum_{i=1}^{n} (\mathbf{1}_{T_i<a_i} - \mathbf{E}(1_{T_i<a_i})) \underset{n \longrightarrow +\infty}{\longrightarrow} 0$$

- Numerator term:

$$\frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}(Y_i) + \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{E}(Y_i))$$

In one side:

$$\begin{aligned}
\mathbf{E}(Y_i) &= \mathbf{E}(\min(T_i, a_i)) \\
&= \mathbf{E}(T_i \mathbf{1}_{T_i <= a_i}) + \mathbf{E}(a_i \mathbf{1}_{T_i > a_i}) \\
&= \int_0^{a_i} t\theta e^{-\theta t} dt + a_i \mathbf{P}_\theta(T_i > a_i) \\
&= \int_0^{a_i} t\theta e^{-\theta t} dt + a_i \mathbf{P}_\theta(T_i > a_i) \\
&= \left[-te^{-\theta t}\right]_0^{a_i} + \int_0^{a_i} e^{-\theta t} dt + a_i e^{-\theta a_i} \\
&= \frac{1}{\theta}(1 - e^{-\theta a_i})
\end{aligned}$$

In another side, we want to prove that $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{E}(Y_i)) \underset{n \longrightarrow +\infty}{\longrightarrow} 0$. It's equivalent to prove the convergence in probability:

$$\forall \epsilon > 0, \mathbf{P}(S_n > \epsilon) \underset{n \longrightarrow +\infty}{\longrightarrow} 0$$

For this, we use the Tchebychev inequality to deduce the convergence from the convergence of $\mathbf{E}(S_n^2)$:

$$\mathbf{P}(S_n > \epsilon) \leq \frac{\mathbf{E}(S_n^2)}{\epsilon^2}$$

Here, we have:

. $S_n = \sum_{i=1}^{n}(Y_i - \mathbf{E}(Y_i))$

$$\begin{aligned}
\mathbf{E}(S_n^2) = \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{E}(Y_i))^2\right) &= \frac{1}{n^2}\mathbf{V}\left(\sum_{i=1}^{n}(Y_i - \mathbf{E}(Y_i))\right) \\
&= \frac{1}{n^2}\mathbf{E}\left(\sum_{i=1}^{n}(Y_i - \mathbf{E}(Y_i))^2\right) \\
&\leq \frac{1}{n^2}\sum_{i=1}^{n} 4a_i^2 = \frac{4}{n}\left(\frac{1}{n}\sum_{i=1}^{n} a_i^2\right)
\end{aligned}$$

- We can, now, conclude:

$$\mathbf{E}(S_n^2) \underset{n \longrightarrow +\infty}{\longrightarrow} 0$$
$$\mathbf{P}(S_n > \epsilon) \underset{n \longrightarrow +\infty}{\longrightarrow} 0$$

So:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{E}(Y_i)) \underset{n \longrightarrow +\infty}{\longrightarrow} 0$$

We finally deduce that:

$$\frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}(Y_i) + o_n(1)$$

$$C\hat{L}V = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i}{\frac{1}{n}|\mathcal{A}|} = \frac{1}{\theta}\frac{\frac{1}{n}\sum_{i=1}^{n}(1 - e^{-\theta a_i}) + o_n(1)}{\frac{1}{n}\sum_{i=1}^{n}(1 - e^{-\theta a_i}) + o_n(1)} \underset{n \longrightarrow +\infty}{\longrightarrow} \frac{1}{\theta}$$

2.5.4 Confidence interval

We established that $\frac{\theta\sum_{i=1}^{n}Y_i-|\mathcal{A}|}{\sqrt{|\mathcal{A}|}} \sim \mathcal{N}(0,1)$

$$P(-q_{1-\frac{\alpha}{2}} \leq \frac{\theta\sum_{i=1}^{n}Y_i-|\mathcal{A}|}{\sqrt{|\mathcal{A}|}} \leq q_{1-\frac{\alpha}{2}}) = 1-\alpha$$

$$P(-\sqrt{|\mathcal{A}|}q_{1-\frac{\alpha}{2}} \leq \theta\sum_{i=1}^{n}Y_i-|\mathcal{A}| \leq \sqrt{|\mathcal{A}|}q_{1-\frac{\alpha}{2}}) = 1-\alpha$$

$$P(|\mathcal{A}|-\sqrt{|\mathcal{A}|}q_{1-\frac{\alpha}{2}} \leq \theta\sum_{i=1}^{n}Y_i \leq |\mathcal{A}|+\sqrt{|\mathcal{A}|}q_{1-\frac{\alpha}{2}}) = 1-\alpha$$

$$P(\frac{|\mathcal{A}|-\sqrt{|\mathcal{A}|}q_{1-\frac{\alpha}{2}}}{\sum_{i=1}^{n}Y_i} \leq \theta \leq \frac{|\mathcal{A}|+\sqrt{|\mathcal{A}|}q_{1-\frac{\alpha}{2}}}{\sum_{i=1}^{n}Y_i}) = 1-\alpha$$

$$P(\frac{\sum_{i=1}^{n}Y_i}{|\mathcal{A}|+\sqrt{|\mathcal{A}|}q_{1-\frac{\alpha}{2}}} \leq \frac{1}{\theta} \leq \frac{\sum_{i=1}^{n}Y_i}{|\mathcal{A}|-\sqrt{|\mathcal{A}|}q_{1-\frac{\alpha}{2}}}) = 1-\alpha$$

$$C\hat{L}V \in \frac{\sum_{i=1}^{n}Y_i}{|\mathcal{A}| \pm q_{1-\frac{\alpha}{2}}\sqrt{|\mathcal{A}|}}$$

The 95 % confidence interval, is given by $q_{1-\frac{\alpha}{2}} = 1.95$. Thus

$$IC_{95\%} = \left[\frac{\sum_{i=1}^{n}Y_i}{|\mathcal{A}|+1.96\sqrt{|\mathcal{A}|}}, \frac{\sum_{i=1}^{n}Y_i}{|\mathcal{A}|-1.96\sqrt{|\mathcal{A}|}}\right]$$

2.6 Pareto distribution - non censored data

2.6.1 Context

We observe n clients. Each client i survives for a duration of $T_i \geq 1$ month. Let a random sample of n iid random variables $T_1, \ldots, T_n$ following a pareto distribution of unknown parameter $\theta > 1$.

2.6.2 Probability function

The probability function of T. First, for $i \in \{1, \ldots, n\}$, we have:

$$1 - F_\theta(t) = P_\theta(T_i > t) = \frac{1}{t^\theta}$$

$$f_\theta(t) = \frac{\theta}{t^{\theta+1}}$$

$$CLV = \mathbf{E}(T_i) = \frac{\theta}{\theta-1}$$

$$\mathbf{V}(T_i) = \frac{\theta}{\theta-2} - \frac{\theta^2}{(\theta-1)^2}$$

2.6.3 Maximum Likelihood Estimator

The maximum likelihood resolution [2] method gives:

$$\begin{aligned} L_\theta(y_1, \ldots, y_n) &= f_\theta(Y_1 = y_1, \ldots, Y_n = y_n) \\ &\underset{\text{(iid)}}{=} \prod_{i=1}^{n} f_\theta(y_i) \\ \ln L_\theta(y_1, \ldots, y_n) &= \sum_{i=1}^{n} \ln\left(\frac{\theta}{y_i^{\theta+1}}\right) \\ &= \sum_{i=1}^{n} \ln(\theta) - (\theta+1)\ln(y_i) \\ \frac{\partial \ln L_\theta(y_1, \ldots, y_n)}{\partial \theta} &= \sum_{i=1}^{n} \frac{1}{\theta} - \ln(y_i) = \frac{n}{\theta} - \sum_{i=1}^{n} \ln(y_i) = 0 \\ \hat{\theta} &= \frac{n}{\sum_{i=1}^{n} \ln(y_i)} \end{aligned}$$

$$C\hat{L}V = \frac{\hat{\theta}}{\hat{\theta} - 1} = \frac{1}{1 - \frac{\sum_{i=1}^{n} \ln(y_i)}{n}}$$

2.7 Pareto distribution - with censored data

2.7.1 Context

We observe n clients. Each client i survives for a duration of $T_i \geq 1$ month. Let a random sample of n iid random variables $T_1, \ldots, T_n$ following a pareto distribution of unknown parameter $\theta > 1$. From T is defined Y such that:

$$Y_i = \min(T_i, a_i)$$

Where:

- $i \in \{1, \ldots, n\}$
- a_i is time from the moment of arrival of client i until time of observation

2.7.2 Probability function

Let's compute the probability function of Y. First, for $i \in \{1, \ldots, n\}$, we have:

$$P_\theta(T_i > t) = \frac{1}{t^\theta}$$

$$CLV = \mathbf{E}(T_i) = \frac{\theta}{\theta - 1}$$

2.7.3 Maximum Likelihood Estimator

The maximum likelihood resolution method gives:

$$C\hat{L}V = \frac{1}{1 - \frac{\sum_{i=1}^{n} \ln(y_i)}{|\mathcal{A}|}}$$

2.7.4 Confidence interval

We established that for large value of n, according to central limit theorem, $\frac{\theta \sum_{i=1}^{n} ln(Y_i) - |\mathcal{A}|}{\sqrt{|\mathcal{A}|}} \sim \mathcal{N}(0,1)$ This leads us to the formula of confidence interval at 95%:

$$IC_{95\%} = \left[\frac{1}{1 - \frac{\sum_{i=1}^{n} ln(Y_i)}{|\mathcal{A}| - 1.96\sqrt{|\mathcal{A}|}}}, \frac{1}{1 - \frac{\sum_{i=1}^{n} ln(Y_i)}{|\mathcal{A}| + 1.96\sqrt{|\mathcal{A}|}}} \right]$$

2.8 Simulation

To simulate an environment and test the consistency of the estimators, here are the steps that we followed:

- Create a generator that provides with censored and not censored data following a given distribution
- Monte Carlo simulation : create multiple samples and compute the $C\hat{L}V$ values according to the estimator that suits the most the generated data.
- Compare the computed $C\hat{L}V$ values with the theoritical CLV and plot the distribution of errors. Study the convergence of computed values' average to the theoretical CLV value and conclude about the consistency of the estimator.
- Compute the confidence intervals : Depending on whether or not we have the theoretical formula.
  - Theoretical formula available :
    - Validate the formula:
      - Calculating the the formula of confidence interval (95%) for each of the Monte Carlo samples.
      - Find the average confidence interval and check how many CLV values fall in it, we define it as confidence level.
      - If the confidence level is around 95% we validate the theoretical formula.
    - Practical use :
      - Generate a dataset following the wanted distribution and compute the CI.
      - Generate n bootstrap samples, compute the CLV values and check how many of them fall in the previous CI.
      - If the confidence level is around 95% we validate the bootstrap method as a good way to approach the theoretical CI when we have only one dataset and we don't necessarily know the CI formula.
  - Theoretical formula not available : we take 95% of Monte Carlo distribution as confidence interval.
- Test the estimator on real data.

2.8.1 Example of simulation : exponential distribution

- Not censored
  - With Monte Carlo simulation, we generate $n = 10000$ samples having an exponential distribution with an expected value $\mathbf{CLV}_{\text{theory}} = 10$. An example of such a sample is given in Figure 1.

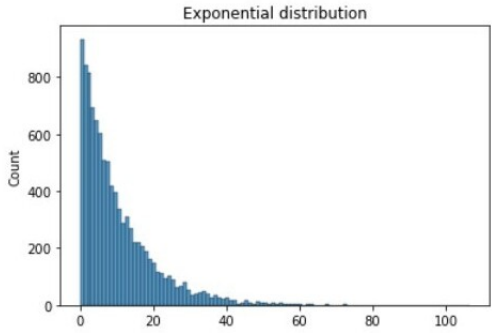

Figure 1: Distribution of one generated sample

  - We estimate the unknown parameter $\hat{\theta}$ of each sample using the estimator 2.4.2, and then deduce the values of $C\hat{L}V$.
  - For each sample $i \leq n$, we estimate the error $e_i = C\hat{L}V_i - CLV_{\text{theory}}$. The distribution of the error is observed in Figure 2.

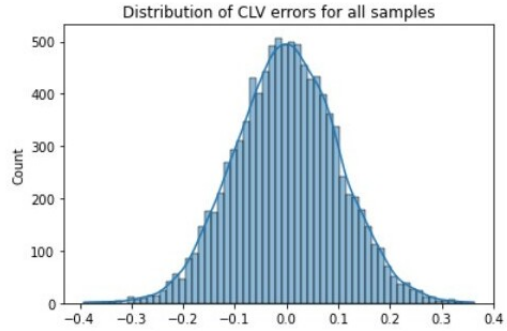

Figure 2: Distribution of estimation error (n samples)

The plot illustrates that the prediction errors follow a Gaussian distribution with a mean of zero. This indicates that, as per the Law of Large Numbers, the estimator tends to converge towards the theoretical value, which can be visualized in Figure 3.

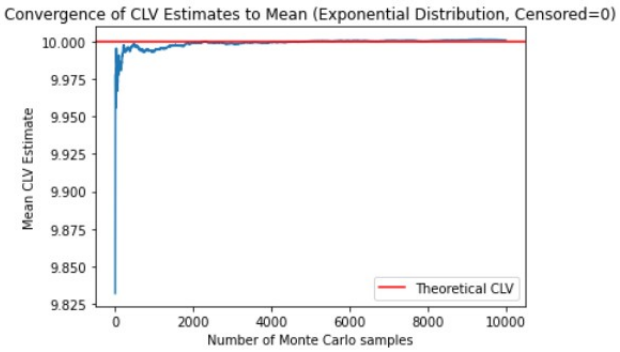

Figure 3: Estimator's convergence

In fact, after a certain number of samples (3000 in this case), we reach a steady mean of CLV values around the theoretical one.

– For the confidence interval, since we don't have the theoretical formula, we use the Monte Carlo method, which is supposed to converge to the theoretical values. It consists of taking 95% of the resulting Gaussian distribution (Figure 4). We obtain $IC_{95\%} = [9.80, 10.20]$.

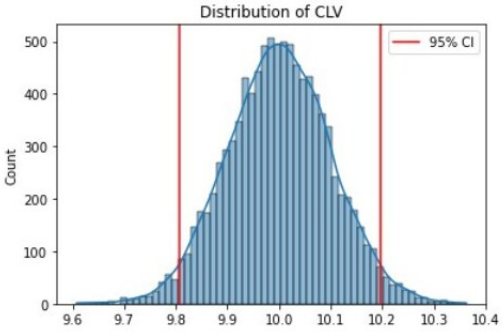

Figure 4: Confidence interval

- Censored

  – The same reasoning as for non-censored data is used for the censored case (Figures 5, 6, 7).

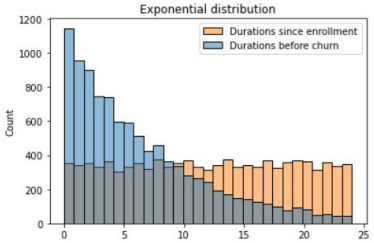

Figure 5: Distribution of one generated sample

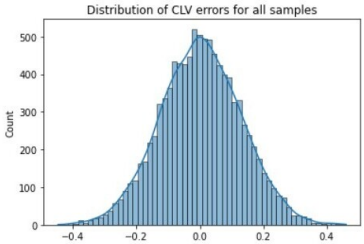

Figure 6: Distribution of estimation error

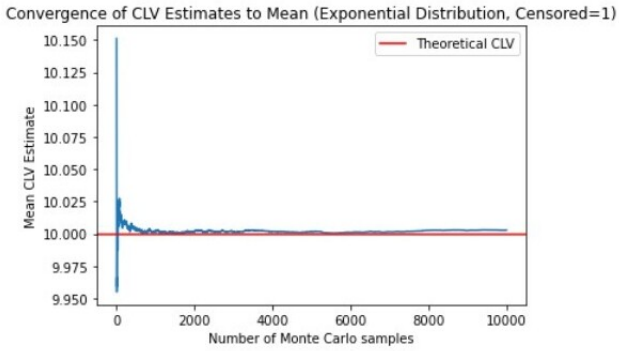

Figure 7: Convergence of the estimator

– We have observed an increase in the standard deviation of the errors distribution when transitioning from uncensored data (0.099908) to censored data (0.126826), representing a rise of 27%. The standard deviation serves as a measure of the dispersion or spread of the errors around the mean or predicted values. The elevation in the standard deviation indicates a negative impact on prediction accuracy, as it signifies a greater variability and less precision in the predicted values.

– In order to validate the theoretical formula 2.5.4 for the confidence interval, we conducted a practical analysis. We generated n confidence intervals for each Monte Carlo sample. To construct an average confidence interval, we calculated the average of the upper and lower bounds (refer to Figure 8). The resulting average confidence interval was determined to be $IC_{theory_{95\%}} = [9.76, 10.26]$.

  Next, we computed the n CLV values and assessed how many of them fell within this confidence interval. The analysis revealed that approximately 95% of the CLV values were within the calculated confidence interval. This outcome serves as evidence supporting the validity of formula 2.5.4 in estimating a 95% confidence level.

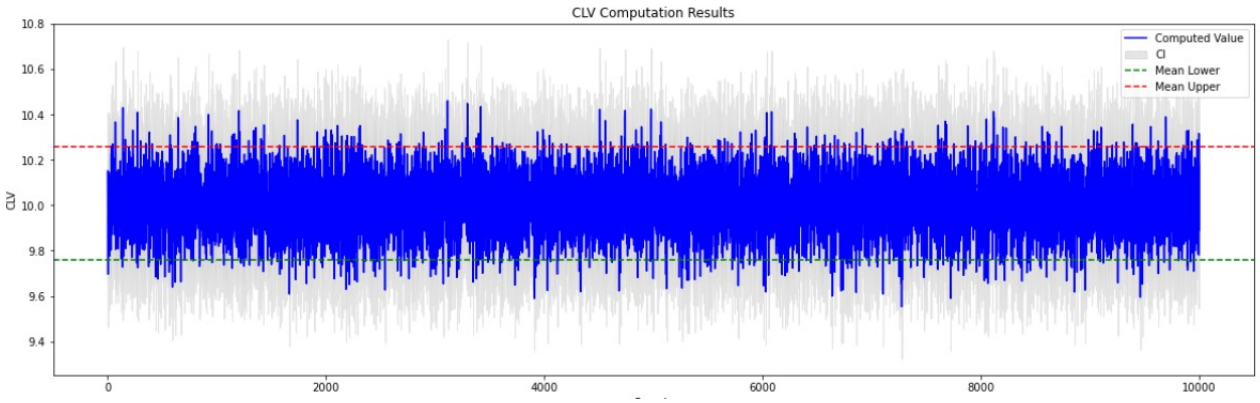

Figure 8: Confidence intervals using theoretical formula

– In this section, we aimed to validate the confidence intervals calculated using the theoretical formula by utilizing the bootstrap method. As it is challenging to precisely determine the data distribution in reality and

generate multiple datasets, the bootstrap method offers a practical solution. For the given dataset, we computed the confidence interval using the theoretical formula and obtained $IC_{theory} = [9.787332, 10.287148]$. Additionally, by employing the bootstrap method (Figure 9), we obtained $IC_{bootstrap} = [9.784276, 10.286276]$.

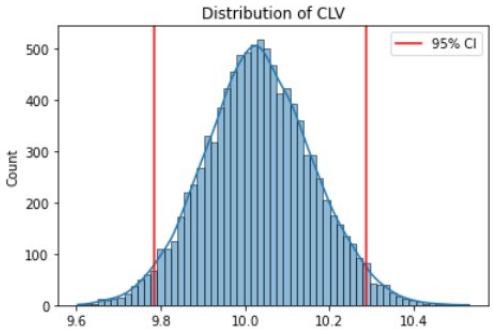

Figure 9: Bootstrapping IC 95%

We can observe that the bootstrapping method has an error of less than 0.1%. This finding allows us to draw a clear conclusion: when it is not possible to derive and compute the formula for the confidence interval (which is often the case in practice due to the unknown data distribution), the bootstrapping method serves as a reliable solution.The minimal error exhibited by the bootstrapping method indicates its effectiveness in estimating the confidence interval without relying on explicit distribution assumptions. This makes it a valuable approach when faced with situations where the exact distribution of the data is unknown.

2.9 Real data

Once we studied the consistency of our estimators on randomly generated data, we apply the result on real data. For that, we use a dataset from the package `sksurv.datasets` called `The Veterans' Administration Lung Cancer Trial` [3], which is a randomized trial of two treatment regimens for lung cancer. The data set (Kalbfleisch J. and Prentice R, (1980) The Statistical Analysis of Failure Time Data. New York: Wiley) consists of 137 patients and 8 variables, which are described below:

- **Treatment**: denotes the type of lung cancer treatment; standard and test drug.
- **Celltype**: denotes the type of cell involved; squamous, small cell, adeno, large.
- **Karnofsky_score**: is the Karnofsky score.
- **Diag**: is the time since diagnosis in months.
- **Age**: is the age in years.
- **Prior_Therapy**: denotes any prior therapy; none or yes.
- **Status**: denotes the status of the patient as dead or alive; dead or alive.

- **Survival_in_days**: is the survival time in days since the treatment.

Our interest is to study how our estimators act on this dataset. For this, we only use **Survival_in_days** and **Status** features.

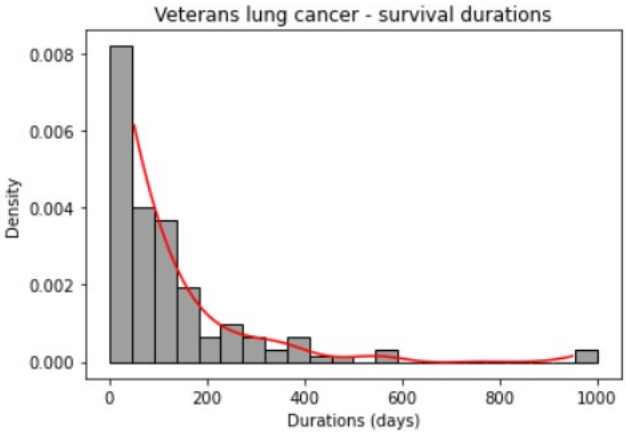

Figure 10: The Veterans' Administration Lung Cancer Trial data distribution

- Since the **Status** feature can be **True** if the patients are dead and **False** if they are alive, the data is censored. Therefore, we utilize our censored estimators (geometric, exponential and pareto) to measure the parameter of the distribution and, consequently, estimate $C\hat{L}V$.
- As observed in the estimators' validation, the geometric and exponential estimators estimate the same parameter so here we only test the exponential and pareto on real data.
- For this dataset, as observed in Figures (11, 12), the exponential estimator fits more the distribution of the data. The pareto is heavy tailed, with the probability for small values being higher than an exponential distribution. We know that if a distribution's tail is "too heavy", then its mean will not exist. This justifies the incoherence of the average CLV value found with Pareto estimator -0.29. Meanwhile for exponential distribution, we observe a good fit to the data. We compare the estimated value to the mean to measure the error and we obtain 0.07%.

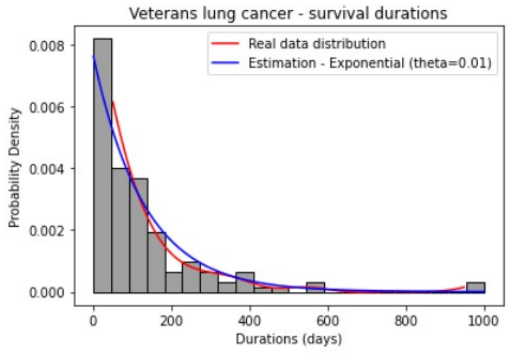

Figure 11: Exponential estimation versus real data

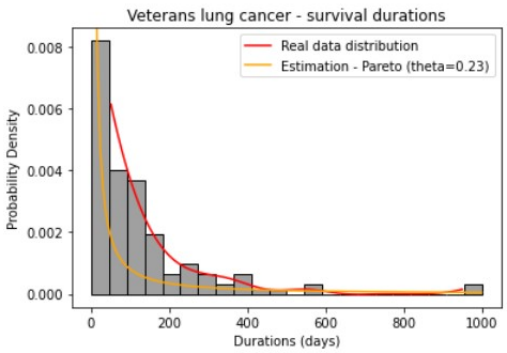

Figure 12: Pareto estimation versus real data

- The error is less than 1% so we conclude that the exponential estimator is suitable for this dataset.

- All our estimators assume that the probability of success (churning) is independent from the customer's characteristics or external effects, which may not be true in practice. Therefore, it may be necessary to use more complex models to account for changes in the probability of churn over time.

2.10 Conclusion on parametric estimators

We focused our study on three parametric estimators based on geometric, exponential, and Pareto distributions. Two scenarios were considered: one with censored data and one without. To validate our theoretical findings, we conducted simulations using real-world data. The following conclusions were drawn from our study:

- Estimating the parameter of a geometric distribution is equivalent to estimating the parameter of an exponential distribution.
- We were able to validate all the theoretical formulas through Monte Carlo simulations.
- The use of censored data resulted in an increase in distribution errors standard deviation (by 27% in the case of censored exponential distribution) and wider confidence intervals (by 25% in the case of the censored exponential distribution). However, even in the presence of censored data, the estimators remained acceptable.
- In cases where only a limited dataset is available (as is often the case in practice), bootstrapping is a reliable method for estimating confidence intervals.

3 Non Parametric Estimators

The study focused solely on parametric estimators. However, exploring non-parametric estimators could be a valuable suggestion for the continuation of this work. One can suggest Kaplan-Meier estimator, the most common example in the actual state of art.

3.1 Kaplan-Meier estimator

The Kaplan Meier estimator or limit product estimator is named after two mathematicians Edward L. Kaplan and Paul Meier. It first appeared in their article "Nonparametric estimation from incomplete observations" [1] published in 1958 in the Journal of the American Statistical Association, vol. (Journal of the American Statistical Association, vol. 53, pp. 457-481). It is widely used in the medical field to perform survival analyses. It also has applications in industry, where it is used to estimate the time to failure of machine parts.

Definition The kaplan Meier estimator estimates the survival function S(t). S(t) is defined by

$$S(t) = P(\tau > t), t \in 0, 1, 3, ...$$

It is the probability that an event occurs after a certain time t.

The kaplan Meier estimator is the non parametric maximum likelyhood estimator of S(t). It is expressed by

$$\hat{S}(t) = \prod_{t_i \leq t} \frac{n_i - d_i}{n_i}$$

- t_i a time where at least one event happened
- n_i the number of elements for those no event happened at time t_i .
- d_i the number of events that happened at time t_i.

Remark : In case there is a censored event then it must be removed from n_i on t_i.

3.1.1 Properties

- Kaplan Meier estimator is non-parametric : that means that it does not make any assumptions about the distribution of the time before an event occurs.
- It can take into account some types of censored data : particularly right-censoring, which occurs if withdraws from a study, is lost to follow-up.

Note that the Kaplan-Meier estimator assumes that the hazard rate (i.e., the probability of dying given that the patient has survived up to that point) is constant over time. If this assumption is violated, the estimate of the survival function may be biased. Therefore, it's important to check the assumptions of the Kaplan-Meier estimator before using it to estimate the survival function.

References

[1] *KM*. URL: https://web.stanford.edu/~lutian/coursepdf/KMpaper.pdf.

[2] *Pareto - parameter estimation*. URL: https://www.casact.org/sites/default/files/database/astin_vol20no2_201.pdf.

[3] *The Veterans' Administration Lung Cancer Trial*. URL: https://scikit-survival.readthedocs.io/en/stable/user_guide/00-introduction.html#The-Veterans%E2%80%99-Administration-Lung-Cancer-Trial.