7 Data Science (djw1005)

We are given a numerical dataset $\{x_1, x_2, \ldots, x_n\}$. We wish to estimate the 99$^{\text{th}}$ percentile, and to find a confidence interval for it. Here are three approaches:

(a) We may decide to model the datapoints as independent samples from the Pareto$(1, \alpha)$ distribution. Then, the 99$^{\text{th}}$ percentile is the value q such that $\mathbb{P}\big(\text{Pareto}(1, \alpha) \leq q\big) = 0.99$.

(i) Find the maximum likelihood estimator for α. [3 marks]

(ii) Find q as a function of α. [2 marks]

(iii) Explain how to use parametric resampling to find a confidence interval for q. Give pseudocode. [4 marks]

(b) We may decide to estimate the 99$^{\text{th}}$ percentile by simply sorting the dataset and reading off the value in position int$(0.99n)$.

Explain how to use nonparametric resampling to find a confidence interval for it. Give pseudocode. Under what circumstances would you expect the result to be unreliable? [6 marks]

(c) We may decide to use computational Bayesian methods to find the confidence interval. Explain how, stating your model precisely. Give pseudocode. [5 marks]

Hint. If $X \sim \text{Pareto}(1, \alpha)$ *then it has cumulative distribution function*

$$\mathbb{P}(X \leq x) = \begin{cases} 1 - x^{-\alpha} & \text{if } x \geq 1 \\ 0 & \text{if } x < 1. \end{cases}$$