Stevens Institute of Technology

SYS-611 Homework #7

Due Apr. 21 2021

Submit the following using the online submission system: 1) Cover sheet with name, date, and collaborators, 2) Written responses in PDF format, 3) All work (e.g. .xlsx or .py files).

7.1 A Jolly Guess [10 points]

Recall the challenge from lecture: use a computational model to estimate the number of Jolly Rancher candies in the jar.

(a) 5 pts Develop process generators for the following quantities:

(i) Jar usable volume ("Le Parfait Super Terrines" 500 ml, overfilled)

(ii) Candy volume (regular Jolly Ranchers, approx. 1 cm diameter, 2.0–2.5 cm length)

(iii) Packing factor (note: appears to be pretty low)

(b) 5 pts Perform a Monte Carlo simulation and report your best estimate for the number of candies in the jar along with an estimate of the 5th percentile and 95th percentile values as bounds on plausible counts.

Optional: Submit your answer at `https://bit.ly/2TWocG1` (one entry per person, must be received by 11:59pm ET Apr. 21 2021). The closest guess without exceeding the actual count wins a batch of wrapped candies (shipped within the U.S.).

7.2 Modeling Old Faithful Eruptions [15 points]

Given a proposed distribution with probability density function (PDF) $f(x)$ for random variable X with a set of n observations $\{X_1, \ldots, X_n\}$, the likelihood function

$$L(X_1, \ldots, X_n) = \prod_{i=1}^{n} f(X_i) = f(X_1) \cdot f(X_2) \cdot \cdots \cdot f(X_n)$$

measures the product of the PDF evaluated at each sample observation X_i to measure how well a proposed distribution explains the observed data. Often it is computationally easier to compute the log likelihood function

$$\log L(X_1, \ldots X_n) = \log \prod_{i=1}^{n} f(X_i) = \sum_{i=1}^{n} \log f(X_i) = \log f(X_1) + \log f(X_2) + \cdots + \log f(X_n)$$

to accommodate very small numerical values.

The `faithful` data set records observations for the Old Faithful geyser with two random variables: the eruption duration D and the waiting time between eruptions W.

(a) 2 PTS To model the eruption duration as a uniform distribution $D \sim \text{uniform}(a, b)$, find the parameters a (lower bound) and b (upper bound) that maximize the likelihood function. Report a, b, and $\log L$. (*Hint*: there is an intuitive solution for a and b).

(b) 2 PTS To model the eruption duration as a normal distribution $D \sim \text{normal}(\mu, \sigma)$, find the parameters μ (mean) and σ (standard deviation) that maximize the likelihood function. Report μ, σ, and $\log L$. (*Hint*: there is an intuitive solution for μ and σ).

(c) 3 PTS To model the eruption duration D as a custom V-shaped distribution, find the parameters a (lower bound), b (upper bound), and c (vertex) for the PDF and CDF:

$$f(d) = \begin{cases} \frac{2(c-d)}{(c-a)(b-a)} & a \le d < c \\ \frac{2(d-c)}{(b-c)(b-a)} & c \le d \le b \end{cases}, \qquad F(d) = \begin{cases} \frac{(c-a)^2-(c-d)^2}{(c-a)(b-a)} & a \le d < c \\ \frac{(d-c)^2+(b-c)(c-a)}{(b-c)(b-a)} & c \le d \le b \end{cases}$$

that maximize the likelihood function. Report a, b, c and $\log L$. (*Hint*: there is an intuitive solution for a and b but you should optimize c).

(d) 3 PTS Plot and visually compare the cumulative distribution functions (CDFs) for the fitted uniform (a), normal (b), and V-shaped distribution (c) to the sample data for eruption duration D. Which distribution has the best fit?

(e) 5 PTS Evaluate the uniform, normal, and **either** the V-shaped distributions **or** propose a new mathematical distribution for the waiting time between eruptions W. Use maximum likelihood estimation (MLE) to find the distribution parameters.