1 Foundations of Data Science (DJW)

Let $x_1, \ldots, x_n$ be a sample of values that we believe is drawn from $X \sim \text{Uniform}[\mu - \theta, \mu + \theta]$, for some unknown parameters $\mu \in \mathbb{R}$ and $\theta > 0$.

(*a*) Calculate the likelihood function $\text{lik}(\mu, \theta \mid x_1, \ldots, x_n)$. [2 marks]

(*b*) We wish to calculate the maximum likelihood estimator $(\hat{\mu}, \hat{\theta})$. The solution for $\hat{\mu}$ is $\hat{\mu} = (m + M)/2$ where $m = \min_i x_i$ and $M = \max_i x_i$. Show that $\hat{\theta} = \max(\hat{\mu} - m, M - \hat{\mu})$. [2 marks]

(*c*) Explain what is meant by (i) resampling, (ii) the error probability of an output procedure. [3 marks]

(*d*) Give pseudocode that uses resampling to plot a histogram of the distribution of values of $\hat{\mu}$ that we might see, if we were to collect a new dataset and repeat the experiment. Explain your resampling method. [5 marks]

(*e*) I propose to use $[\hat{\mu} - \delta, \hat{\mu} + \delta]$ as a confidence interval for μ, where δ is given. Explain how to estimate the error probability of my confidence interval. [5 marks]

(*f*) Discuss briefly how to find a confidence interval for θ. [3 marks]

2 Foundations of Data Science (DJW)

Let $X_1, \ldots, X_n$ be independent random variables drawn from the distribution $\text{Uniform}[\mu - \theta, \mu + \theta]$, for some unknown parameters $\mu \in \mathbb{R}$ and $\theta > 0$.

(*a*) Calculate the likelihood function $\text{lik}(\mu, \theta \mid x_1, \ldots, x_n)$. [2 marks]

(*b*) We wish to calculate the maximum likelihood estimator $(\hat{\mu}, \hat{\theta})$. The solution for $\hat{\mu}$ is $\hat{\mu} = (m + M)/2$ where $m = \min_i x_i$ and $M = \max_i x_i$. Calculate $\hat{\theta}$. [4 marks]

(*c*) Using $\text{Normal}(\mu_0, \sigma_0^2)$ as the prior distribution for μ, and $\text{Exp}(\lambda_0)$ as the prior distribution for θ, calculate the posterior distribution

$$\Pr(\mu, \theta \mid x_1, \ldots, x_n).$$

[4 marks]

Suppose we have a random number generator `r_mu_theta()` that samples (μ, θ) from the distribution you found in part (*c*).

(*d*) Give pseudocode that uses `r_mu_theta()` to estimate the posterior mean of μ. [4 marks]

(*e*) Let X' be a new value drawn from $\text{Uniform}[\mu - \theta, \mu + \theta]$. Calculate $\mathbb{P}(X' \leq y | \mu, \theta)$, where y is given. Hence, give pseudocode to estimate

$$\mathbb{P}(X' \leq y \mid x_1, \ldots, x_n).$$

[6 marks]

Hint. The $\text{Normal}(\mu, \sigma^2)$ *distribution has density*

$$\Pr(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2 / 2\sigma^2}, \quad x \in \mathbb{R}$$

and the $\text{Exp}(\lambda)$ *distribution has density*

$$\Pr(x) = \lambda e^{-\lambda x}, \quad x > 0.$$

3 Foundations of Data Science (DJW)

We are given a dataset of police stop-and-search records. These include the ethnicity of the suspect, and whether or not something suspicious was found in the search. Let E be the number of ethnicities represented in the dataset, let n_e be the number of stops of suspects of ethnicity $e \in \{1, \ldots, E\}$, let x_e be the number of these in which something suspicious was found, and assume it comes from the distribution $X_e \sim \text{Binom}(n_e, \beta_e)$ for $\beta_e \in [0, 1]$.

We propose to measure ethnic bias in policing using the metric

$$d(\beta) = \max_{e,e'} |\beta_e - \beta_{e'}|$$

where $\beta = (\beta_1, \ldots, \beta_E)$.

(a) Find the maximum likelihood estimator $\hat{\beta}$. Explain your reasoning. [3 marks]

(b) Take as a prior distribution E independent random variables $\beta_e \sim \text{Beta}(\delta, \delta)$ where $\delta = 1/2$. Calculate the posterior distribution of β. [4 marks]

(c) Give pseudocode to compute a 95% posterior confidence interval for $d(\beta)$. [4 marks]

(d) Explain what is meant by *resampling*. Given $c > 0$, explain how to compute the error probability of the confidence interval "$d(\beta) \le d(\hat{\beta}) + c$". [6 marks]

(e) We'd like to pick c such that the error probability of the confidence interval in (d) is 5%. Give pseudocode to do this. [3 marks]

Hint. The $\text{Beta}(\alpha, \beta)$ *distribution has density*

$$\Pr(x) = \binom{\alpha + \beta - 1}{\alpha - 1} x^{\alpha-1}(1 - x)^{\beta-1}, \quad x \in [0, 1].$$