1. Statistical Models and Identifiability

For each of the following examples, define a statistical model and check whether the parameter of interest is identifiable. Follow the definitions closely; it is helpful to consider the following: What is Θ and P_θ? What would it mean for the model to be identifiable?

(a)

4/4 points (graded)

1. One observes n i.i.d. Poisson random variables with unknown parameter λ.

● λ identifiable ✓

○ λ not identifiable

2. One observes n i.i.d. exponential random variables with parameter λ, which is unknown but a priori known to be no larger than 10.

● λ identifiable ✓

○ λ not identifiable

3. One observes n i.i.d. uniform random variables in the interval $[0, \theta]$, where θ is unknown.

● θ identifiable ✓

○ θ not identifiable

4. One observes n i.i.d. Gaussian random variables with unknown parameters μ, σ^2.

● (μ, σ^2) identifiable ✓

○ (μ, σ^2) not identifiable

Solution:

In question 1., λ is identifiable because it is the expectation of X_i, $\mathbb{E}[X_i] = \lambda$, where X_i denotes each of the Poisson random variables.

In question 2., λ is identifiable because the expectation of each variable X_i is $\mathbb{E}[X_i] = \lambda^{-1}$, which can be solved for λ.

In question 3., θ is identifiable because the expectation of X_i is $\mathbb{E}[X_i] = \theta/2$.

In question 4., (μ, σ^2) is identifiable because $\mathbb{E}[X_i] = \mu$, $\text{Var}(X_i) = \sigma^2$.

提交 你已经尝试了1次（总共可以尝试1次）

🛈 Answers are displayed within the problem

(b)

2/4 points (graded)

1. One observes the sign of n i.i.d. Gaussian random variables with unknown parameters μ, σ^2.

○ (μ, σ^2) identifiable

● (μ, σ^2) not identifiable ✓

2. *StatGen* is a statistical procedure to test the relevance of genes. When well calibrated, it outputs the (random) proportion of active genes in a (random) cell. We want to estimate the distribution of this proportion. To that end, we take n iid cells and submit them to *StatGen*. We model the output of *StatGen* as n random variables $X_1, \ldots, X_n$ that have uniform distribution on $[0, \theta]$ for some unknown theta.

● θ identifiable ✓

○ θ not identifiable

3. The US Census Bureau is interested in finding out the average commute time of Bostonians. To that end, it randomly selects n individuals, with replacement, among the people who work and live in the Boston area, and asks to each if their commute time is at least 20 minutes. The commute time of a random person is assumed to follow an exponential distribution with parameter λ.

○ λ identifiable ✓

● λ not identifiable ✗

4. Willy Wonka's contains 67 identical machines. Each machine has a lifetime that is modeled as an exponential random variable with some unknown parameter λ. After a certain time $T = 500$ days, one has observed the lifetimes of all machines that have stopped working before T. The parameter of interest is λ.

○ λ identifiable ✓

● λ not identifiable ✗

Solution:

In question 1. (μ, σ) are **not identifiable** . We can write the sign of a Gaussian variable X_i as a Rademacher random variable Y_i with $\mathbf{P}(Y_i = 1) = \mathbf{P}(X_i \geq 0) = \Phi(\mu/\sigma)$, where Φ denotes the cdf of a standard Gaussian variable. Hence, (μ, σ^2) and $(\tilde{\mu}, \tilde{\sigma}^2) = (2\mu, 4\sigma^2)$ will lead to the same distribution.

In question 2., θ is identifiable for the same reason as in the previous problem. Note that this was very much dependent on how we modeled the responses of the procedure.

In question 3., what we collect can be seen as Bernoulli random variables Y_i with hitting probability $p = \exp(-20\lambda)$, hence λ can be reconstructed by

$$\lambda = -\frac{\log \mathbb{E}[X_i]}{20},$$

so it is identifiable.

In question 4., λ is identifiable. The problem setting implies that we observe truncated Exponential variables, $Y_i = \min\{X_i, 500\}$ where $X_i \sim \mathrm{Exp}\,(\lambda)$ is not observed. In particular, one feature of the observed distribution is the proportion of machines that are still running, i.e.,

$$\mathbf{P}\,(Y_i = 500) = \mathbf{P}_\lambda\,(X_i \geq 500) = \exp\,(-500\lambda)\,.$$

This expression can be inverted, so that we get

$$\begin{aligned}\lambda = &\ -\frac{\log\left(\mathbf{P}_\lambda\,(X_i \geq 500)\right)}{500}\\ = &\ -\frac{\log\left(\mathbf{P}_\lambda\,(Y_i = 500)\right)}{500},\end{aligned}$$

by the definition of the observed variables Y_i .

提交 你已经尝试了1次（总共可以尝试1次）

Answers are displayed within the problem

讨论

显示讨论

主题: Unit 2 Foundation of Inference:Homework 1: Estimation, Confidence Interval, Modes of Convergence / 1. Statistical Models and Identifiability

认证证书是什么?