CS109 February 29, 2024

Sampling and Bootstrapping, MLE

Before you leave lab, make sure you click here so that you're marked as having attended this week's section. The CA leading your discussion section can enter the password needed once you've submitted.

1 Warmups

1.1 Sample and Population Mean

Computing the sample mean is similar to the population mean: sum all available points and divide by the number of points. However, sample variance is slightly different from population variance.

1. Consider the equation for population variance, and an analogous equation for sample variance.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \tag{1}$$

$$S^2_{biased} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{2}$$

   S^2_{biased} is a random variable to estimate the constant σ^2. Because it is biased, $E[S^2_{biased}] \neq \sigma^2$. Is $E[S^2_{biased}]$ greater or less than σ^2?

2. Consider an alternative Random Variable, $S^2_{unbiased}$ (known simply as S^2 in class). The technique of un-biasing variance is known as *Bessel's correction*. Write the $S^2_{unbiased}$ equation.

1.2 MLE

Suppose $x_1, \ldots, x_n$ are iid (independent and identically distributed) values sampled from some distribution with density function $f(x|\theta)$, where θ is unknown. Recall that the likelihood of the data is

$$L(\theta) = f(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

Recall we solve an optimization problem to find $\hat{\theta}$ which maximizes $L(\theta)$, i.e., $\hat{\theta} = \arg\max_\theta L(\theta)$.

1. Write an expression for the log-likelihood, $LL(\theta) = \log L(\theta)$.
2. Why *can* we optimize $LL(\theta)$ rather than $L(\theta)$?
3. Why *might* we optimize $LL(\theta)$ rather than $L(\theta)$?

2 Problems

2.1 Variance of Hemoglobin Levels

A medical researcher treats patients with dangerously low hemoglobin levels. She has formulated two slightly different drugs and is now testing them on patients. First, she administered drug A to one group of 50 patients and drug B to a separate group of 50 patients. Then, she measured all the patients' hemoglobin levels post-treatment. For simplicity, assume that all variation in the patient outcomes is due to their different reactions to treatment.

The researcher notes that the sample mean is similar between the two groups: both have mean hemoglobin levels around 10g/dL. However, drug B's group has a **sample variance** that is 3 $(g/dL)^2$ **greater** than drug A's group. The researcher thinks that patients respond to drugs A and B differently. Specifically, she wants to make the scientific claim that drug A's patients will end up with a significantly different spread of hemoglobin levels compared to drug B's.

You are skeptical. It is possible that the two drugs have practically identical effects and that the observed different in variance was a result of chance and a small sample size, i.e. the **null hypothesis**. Calculate the probability of the null hypothesis using bootstrapping. Here is the data. Each number is the level of an independently sampled patient:

Hemoglobin Levels of Drug A's Group ($S^2 = 6.0$): 13, 12, 7, 16, 9, 11, 7, 10, 9, 8, 9, 7, 16, 7, 9, 8, 13, 10, 11, 9, 13, 13, 10, 10, 9, 7, 7, 6, 7, 8, 12, 13, 9, 6, 9, 11, 10, 8, 12, 10, 9, 10, 8, 14, 13, 13, 10, 11, 12, 9

Hemoglobin Levels of Drug B's Group ($S^2 = 9.1$): 8, 8, 16, 16, 9, 13, 14, 13, 10, 12, 10, 6, 14, 8, 13, 14, 7, 13, 7, 8, 4, 11, 7, 12, 8, 9, 12, 8, 11, 10, 12, 6, 10, 15, 11, 12, 3, 8, 11, 10, 10, 8, 12, 8, 11, 6, 7, 10, 8, 5

Complete the prompts in this Colab notebook to investigate this question using bootstrapping.

2.2 Parameter Estimation and Wealth Distribution

The broader field of economics also relies on likelihood estimation and parameter estimation. One continuous probability distribution—one with a long tail as x approaches infinity—is used to model wealth inequality and the socioeconomic problems that stem from it. This probability distribution is given as:

$$f(x|\omega) = \frac{\omega 3^{\omega}}{x^{\omega+1}}, \text{ where } x \geq 3, \omega > 1$$

Assume that you've observed a sample of iid random variables $(X_1, X_2, X_3, ..., X_n)$, where each of the X_i is modeled according to the above probability distribution function.

a. What is the log-likelihood function $LL(\omega)$ of the sample $(X_1, X_2, X_3, ..., X_n)$? Simplify using properties of logarithms wherever possible.

b. Set up the equation that would need to be solved in order to compute $\hat{\omega}_{MLE}$. Once you arrive at the equation and have worked through any calculus, you can stop and simply present the equation that can be solved via simple algebraic manipulation.