STAT 102B: Sample Exam Questions

George Michailidis

Problem 1:

(1) Consider θ to be a parameter of interest for the population distribution F.

Suppose a random independent and identically distributed sample of size n, $\mathbf{x} = (X_1, \cdots, X_n)$ has been drawn from the population distribution F.

Part (a): Assume that for the sample estimate $\hat{\theta}$, the asymptotic sampling distribution is a normal distribution.

Explain how you can use the bootstrap to construct an $(1 - \alpha)$ level confidence interval.

Write down the formula for the confidence interval.

Part (b): Write pseudo-code that implements the confidence interval proposed in Part (a).

Problem 2: Let $\mathbf{x} = (X_1, X_2, \cdots, X_n)$ be an independent and identically distributed random sample drawn from the Bernoulli distribution.

Recall that the Bernoulli distribution is defined as:

$$\mathbb{P}(X = 1) = p, \quad \mathbb{P}(X = 0) = 1 - p.$$

Part (a): Use the central limit theorem to propose a confidence interval for the population parameter p.

Part (b): Write pseudo-code that implements the confidence interval proposed in Part (a).

Part (c): Propose another type of confidence interval for the population parameter p that is based on the bootstrap sampling distribution of the sample estimate $\hat{p}$.

Part (d): Write pseudo-code that implements the confidence interval proposed in Part (c).

Problem 3: Give a brief and clear statement of the central limit theorem. Explain how it can be used to obtain a confidence interval for the corresponding population parameter.

Also, provide an interpretation of the confidence interval proposed above; i.e., if a confidence interval $CI = [C_L, C_U]$ is obtained, what does it say about the corresponding parameter of interest?

Problem 4: What information shall one give you to be able to construct a percentile bootstrap confidence interval for the mean of a Gamma(α, β) distribution?

Write down the form of the confidence interval.

Problem 5: One has obtained an estimate of the correlation coefficient r between two random variables, from an independent and identically distributed random sample $\mathbf{x}$ of size n.

Let $\hat{r} = 0.60$ in the sample $\mathbf{x}$.

Further, suppose that the standard deviation of the bootstrap sampling distribution of $\hat{r}$ is 0.10.

Explain which type of bootstrap based confidence intervals you can construct based on this information. Write down the upper and lower bounds of the corresponding confidence interval.

Problem 6: One has obtained an estimate of the correlation coefficient r between two random variables, from an independent and identically distributed random sample $\mathbf{x}$ of size n.

Let $\hat{r} = 0.60$ in the sample $\mathbf{x}$.

Part (a): Further, suppose that the 2.5-th percentile of the bootstrap sampling distribution of $\hat{r}$ is 0.39 and the 97.5-th percentile of the bootstrap sampling distribution is 0.77, respectively.

Propose and construct two types of bootstrap based confidence intervals for the population correlation coefficient r. What is the level α of the confidence intervals constructed?

Part (b): What other information do you require to construct a third type of bootstrap based confidence interval?

problem 7: A data scientist has access to a *small* data set of size $n = 350$ of single family house prices in Southern California for the year 2019.

The data scientist is interested in constructing an $(1 - \alpha)$ confidence interval for the *median* price of single family houses in Southern California in 2019.

Part (a): Can you help the data scientist to construct two different types of confidence intervals based on the bootstrap?

Specifically, provide the formulas for these two types of confidence intervals.

Part (b): What assumption(s) do you need to make for the confidence intervals to be valid?

Part (c): Provide pseudo-code that would construct the two types of confidence intervals proposed in Part (a).

Part (d): The data scientist would also like to construct a bootstrap based confidence interval for the *minimum* value of house prices in Southern California for the year 2019. Can you provide assistance?

Problem 8: A data scientist has obtained an independent and identically distributed sample $\mathbf{x} = (X_1, \cdots, X_n)$ of size n of positive and negative answers regarding the quality of a product; i.e., X_i =positive, if respondent i thinks the underlying product is of good quality and X_i=negative, if respondent i thinks the underlying product is *not* of good quality.

Part (a): The data scientist wants your help to construct a *normal bootstrap confidence interval* for the proportion of positive opinions about the quality of the product in the population of all consumers of that product.

Specifically, write pseudo-code that constructs a $(1-\alpha)$ *normal confidence interval.*

Part (b): Another data scientist wants to construct a *bootstrap-t (studentized) confidence interval* for the proportion of positive opinions about the quality of the product in the population of all consumers of that product.

Explain what changes you need to make to your pseudo-code in Part (a).