10 Statistics

10.3 Bootstrap Estimation of Standard Error (5 units)

This project is not strongly related to particular courses in Part II, though practice in statistical thinking will obviously be helpful.

Background

Bootstrap methods are procedures for the empirical estimation or approximation of sampling distributions and their characteristics. Their primary use lies in estimating the accuracy (e.g., bias or variance) of parameter estimators, and in constructing confidence sets or hypothesis tests. They are applied in circumstances where the form of the population from which the observed data was drawn is unknown.

The general bootstrap method was formalized by Efron [1], [2]. In this project, the bootstrap method will be used to estimate the standard error of certain statistics derived from a sample of independent, identically distributed random variables.

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be an IID sample on some sample space Ω, drawn from a distribution F, and let $T(\mathbf{X})$ be a statistic of interest. The standard error of T is

$$\sigma(T; F) = \sqrt{\mathrm{Var}_F\, T(\mathbf{X})}\,.$$

The non-parametric bootstrap estimate of the standard error is

$$\sigma(T; \hat{F}) = \sqrt{\mathrm{Var}_{\hat{F}}\, T(\mathbf{Y})}\,,$$

where $\mathbf{Y}$ is an IID sample of size n drawn from the empirical distribution

$$\hat{F}(A) = \frac{1}{n} \sum_{i=1}^{n} 1\,[X_i \in A] \qquad \text{for } \; A \subset \Omega\,.$$

> **Question 1** Show that $\mathbf{Y}$ is the same as a random sample of size n, drawn with replacement from the actual sample $\mathbf{X}$. Comment on the reasonableness of the bootstrap estimate, such as its bias, ergodic variance, etc.

Often there will be no simple expression for $\sigma(T; \hat{F})$. It is, however, simple to estimate it numerically by means of simulation. The algorithm proceeds in three steps:

1. Draw a large number B of independent bootstrap samples $\mathbf{Y}_1, \ldots, \mathbf{Y}_B$.
2. For each bootstrap sample, evaluate the statistic $T(\mathbf{Y}_b)$.
3. Calculate the sample standard deviation of the $T(\mathbf{Y}_b)$ values:

$$\hat{\sigma_B} = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} \big(T(\mathbf{Y}_b) - \bar{T}\big)^2}\,, \qquad \text{where} \qquad \bar{T} = \frac{1}{B} \sum_{b=1}^{B} T(\mathbf{Y}_b)\,.$$

As $B \to \infty$, $\hat{\sigma_B}$ will approach $\sigma(T; \hat{F})$.

1 Correlation Coefficient

Suppose each sample point X_i consists of a pair $X_i = (Y_i, Z_i)$. The *variance-stabilized correlation coefficient* is

$$T(\mathbf{X}) = \frac{1}{2} \log\left(\frac{1 + r(\mathbf{X})}{1 - r(\mathbf{X})}\right),$$

where

$$r(\mathbf{X}) = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(Z_i - \bar{Z})}{\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}\sqrt{\sum_{i=1}^{n}(Z_i - \bar{Z})^2}},$$

and $\bar{Y} = n^{-1}\sum Y_i$ and $\bar{Z} = n^{-1}\sum Z_i$.

The data set `2-10-3-2016.csv` on the CATAM website contains IQ data from 120 people. Each data point X_i consists of two statistics $X_i = (\mathrm{VIQ}_i, \mathrm{PIQ}_i)$. The first measures the verbal IQ score (based on verbal questions and verbal responses); the second measures the performance IQ score (based on picture arrangement, object assembly and other nonverbal tasks).

Question 2 Use the bootstrap method to estimate the distribution of T, i.e., plot a histogram of the bootstrap values $T(\mathbf{Y}_b)$. Comment on any interesting features.

Question 3 Use the bootstrap method to estimate $\sigma(T; F)$ by finding $\hat{\sigma}_B$ for a reasonable value of B. Repeat this experiment several times, for the same value of B, and plot a histogram of the values of $\hat{\sigma_B}$ you obtain. How does this histogram change with B? What value of B would you advise?

Question 4 Theory tells that for a bivariate normal distribution F, the standard error of T equates to

$$\sigma(T; F) = \frac{1}{\sqrt{n-3}}.$$

It has been suggested that IQ scores are normally distributed. Does your analysis provide sufficient evidence to reject the hypothesis that the Verbal and Performance IQ data are bivariate normal?

2 Uniform Data

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a sample of real-valued random variables. Suppose the distribution of the statistic

$$T(\mathbf{X}) = \max\{X_1, \ldots, X_n\}$$

is of interest. For this very simple statistic, it is possible to calculate $\sigma(T; \hat{F})$ exactly.

Question 5 Calculate $\sigma(T; \hat{F})$. (You may wish to assess your answer by comparing it to what you obtain using the bootstrap algorithm, for some sample $\mathbf{X}$, but you do not need to include any such tests in your final report.)

Suppose the sample $\mathbf{X}$ comes from the uniform distribution on $[0, \theta]$ for some $\theta > 0$. (Then $T(\mathbf{X})$ is the maximum likelihood estimator for θ.)

Question 6 Calculate $\sigma(T; F)$.

Question 7 Generate a sample $\mathbf{X}$ with $\theta = 5$ and $n = 100$. Compare $\sigma(T; \hat{F})$ to $\sigma(T; F)$. Repeat for increasing n. How well does the bootstrap method perform? Why?

References

[1] B. Efron,
Bootstrap Methods: Another Look at the Jackknife,
Annals of Statistics **7** (1979) 1–26.

[2] B. Efron, R.J. Tibshirani,
An Introduction to the Bootstrap,
Chapman and Hall (1993), ISBN 0-412-44980-3.

[3] A.C. Davison, D.V. Hinkley,
Bootstrap Methods and Their Application
Cambridge Series in Statistical and Probabilistic Mathematics, No 1,
Cambridge University Press, 1997, ISBN 0521574714.

