STAT/Q SCI 403: Introduction to Resampling Method
Spring 2019
Homework 06

Instructions:

- You have to submit all your answers in a single PDF file generated by either LaTeX or *Rmarkdown*.
- You may use the LaTeXtemplate `HW_template.tex` to submit your answer.
- For questions using R, you have to attach your code in the PDF file. If the question ask you to plot something, you need to attach the plot in the PDF as well.
- If the question asks you to show a figure, the clarity of the figure will also be graded.
- The total score of this homework is 8 points.
- Questions with ♠ will be difficult questions.

Questions:

1. Assume that your data consists of $x_1, \cdots, x_n$, n values. When we generate the bootstrap sample, we sample with replacement of these n points to obtain a set of IID new points $X_1^*, \cdots, X_n^*$ such that

$$P(X_\ell^* = x_1) = P(X_\ell^* = x_2) = \cdots = P(X_\ell^* = x_n) = \frac{1}{n} \tag{1}$$

   for each ℓ. This new dataset, $X_1^*, \cdots, X_n^*$, is called a bootstrap sample.

   (a) **(1 pt)** Show that the bootstrap sample is an IID random sample from $\widehat{F}_n$, where

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \le x),$$

   is the EDF formed by the original data points $x_1, \cdots, x_n$.

   (b) **(1 pt; ♠)** Assume we want to use the bootstrap to estimate the variance of the sample mean. It is well-known that the variance of the sample mean can be approximated by the sample variance divided by n, the sample size. Namely,

$$\widehat{\sigma}^2 = S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2, \quad \bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Let $\bar{X}_n^* = \frac{1}{n}\sum_{i=1}^n X_i^*$ be the sample mean of a *bootstrap sample*. Given the original data $x_1, \cdots, x_n$ being fixed, show that

$$\mathsf{Var}(\bar{X}_n^*) = \frac{1}{n^2}\sum_{i=1}^n (x_i - \bar{x}_n)^2 = \frac{n-1}{n^2} \cdot S_n^2.$$

(this implies $\mathsf{Var}(\bar{X}_n^*) \approx S_n^2/n$ when the sample size is large)

2. In this problem, we will use the bootstrap technique to analyze the `faithful` dataset (a built-in dataset in R). We focus on the standard deviation (SD) of the variable `waiting`.

   (a) **(1 pt)** Apply the (empirical) bootstrap $10,000$ times to show the bootstrap distribution of the SD. Attach a vertical line indicating the original value of the sample SD of the `waiting`.

   (b) **(0.5 pt)** What are the bootstrap estimate of the variance and MSE of the sample SD?

   (c) **(1 pt)** Use both the asymptotic normality method and the quantile method to construct 95% CIs of the SD. You should report two confidence intervals.

   (d) **(1 pt)** Let σ be the true SD of variable `waiting`. Assume we want to test the hypothesis:

   $$H_0 : \sigma = 15$$

   using the bootstrap sample. There are many ways to test this hypothesis using the bootstrap approach. Here we simply use the bootstrap variance estimate to test this hypothesis. What is the p-value?

   (e) **(1 pt)** Briefly explain the benefit(s) of increasing the number of bootstrap samples (in this problem the number of bootstrap samples is $10,000$). Hint: the bootstrap is a Monte Carlo method.

3. ♠♠ In this problem, we will use the bootstrap to analyze the odds ratio of UC Berkeley's admission dataset, a built in dataset in R. In particular, we will focus on the department `A`. To obtain this dataset, use the command `UCBAdmissions[,,1]` in R. It is a 2 by 2 contingency table as the follows:

	Male	Female
Admitted	512	89
Rejected	313	19

The product of the diagonal terms (512 and 19) divided by the product of the off-diagonal terms (313 and 89), is called the odds ratio. In this case, the odds ratio $OR = \frac{512 \cdot 19}{313 \cdot 89} \approx 0.349212$.

(a) **(0.5 pt)** Use the bootstrap to compute the MSE of the odds ratio OR.

(b) **(0.5 pt)** If there is no gender bias, the odds ratio will be 1. Use the bootstrap to compute the p-value of testing

$$H_0 : \text{no gender bias in this contingency table.}$$

(c) **(0.5 pt)** In this case, the parametric bootstrap and the empirical bootstrap are the same procedure. Explain why.