Practice problems–Bootstrap Jackknife

Problem 1

Given a dataset of n distinct values, show that the number of distinct bootstrap samples is

$$\binom{2n-1}{n}.$$

How many are there for $n = 15$?

Problem 2

Suppose that 50 people are given a placebo and 50 are given a new treatment. Thirty placebo patients show improvement, while 40 treated patients show improvement. Let p_2 be the probability of improving under treatment and p_1 the probability of improving under placebo.

a. Consider $\tau = p_1 - p_2$. Find an estimator for τ (the obvious one is good enough). Using bootstrap, estimate its standard error and bias. Construct a 90 percent confidence interval using the method of your choice.
b. Next, consider $\theta = p_1/p_2$. Find an estimator for θ (again, the obvious one is good enough). Using bootstrap, estimate its standard error and bias. Construct a 90 percent confidence interval using the method of your choice.
c. Consider the null hypothesis $H_0 : p_1 = p_2$. Which of the two confidence intervals above can be used to test this hypothesis? Discuss

Problem 3

Consider the `Rainfall` dataset in the `bootstrap` package:

```
rainfall data. The yearly rainfall, in inches, in Nevada City, California, USA, 1873 through 1978.
```

For more details about each variable, have a look at `?bootstrap::Rainfall`.

a. Compute the sample interquartile range (IQR) for this data.
b. Use bootstrap to estimate the bias and standard error of the sample IQR.
c. Construct a 95% confidence interval for the IQR.
d. (Bonus) Can you use the jackknife to estimate the bias and standard error of the IQR? Why?

Problem 4

Consider the following data, which represent 12 observations of failure times (in hours) of air-conditioning equipment:

```
library(boot)
str(aircondit)
```

```
## 'data.frame':    12 obs. of  1 variable:
##  $ hours: num  3 5 7 18 43 85 91 98 100 130 ...
```

We can model this data using an exponential distribution:

$$f(x) = \frac{1}{\beta} \exp\left(-\frac{x}{\beta}\right), \quad x > 0.$$

The theory of *Maximum Likelihood Theory* gives us an estimate of β and its standard error:

$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad SE(\hat{\beta}) = \frac{\hat{\beta}}{\sqrt{n}}.$$

In other words, $\hat{\beta}$ is the sample mean.

a. Use bootstrap to estimate the bias and standard error of the maximum likelihood estimate $\hat{\beta}$.
b. Compare the **normal** and **student** bootstrap 95% confidence intervals with the approximate 95% confidence interval from Maximum Likelihood theory.

Problem 5

For this problem, we will use the `olympic` dataset in the `ade4` package, which contains the decathlon score of 33 athletes:

```
library(ade4)
data("olympic")
# The dataset is the first element of the list
data_olympic <- olympic$tab

str(data_olympic)
```

```
## 'data.frame':    33 obs. of  10 variables:
##  $ 100 : num  11.2 10.9 11.2 10.6 11 ...
##  $ long: num  7.43 7.45 7.44 7.38 7.43 7.72 7.05 6.95 7.12 7.28 ...
##  $ poid: num  15.5 15 14.2 15 12.9 ...
##  $ haut: num  2.27 1.97 1.97 2.03 1.97 2.12 2.06 2 2.03 1.97 ...
##  $ 400 : num  48.9 47.7 48.3 49.1 47.4 ...
##  $ 110 : num  15.1 14.5 14.8 14.7 14.4 ...
##  $ disq: num  49.3 44.4 43.7 44.8 41.2 ...
##  $ perc: num  4.7 5.1 5.2 4.9 5.2 4.9 5.7 4.8 4.9 5.2 ...
##  $ jave: num  61.3 61.8 64.2 64 57.5 ...
##  $ 1500: num  269 273 263 285 257 ...
```

We will use bootstrap and jackknife to study **Principal Component Analysis** (PCA). PCA is a dimension reduction method, where we summarise a dataset by taking linear combinations of the columns that have maximal variance: the first **principal component** is the linear combination with maximal variance, the second principal component has maximal variance among all linear combinations that uncorrelated with the first principal component, etc. The details of how to achieve this are not important for this course (see STAT 3690 for more details); we will use the function `prcomp` in `R`:

```r
# PCA of data_olympic
fit <- prcomp(data_olympic)
summary(fit)
```

```
## Importance of components:
##                             PC1    PC2     PC3     PC4     PC5
## Standard deviation     13.7807 5.8043 2.96066 0.94062 0.69811
## Proportion of Variance  0.8116 0.1440 0.03746 0.00378 0.00208
## Cumulative Proportion   0.8116 0.9555 0.99300 0.99678 0.99886
##                             PC6     PC7     PC8    PC9    PC10
## Standard deviation     0.34653 0.24744 0.23147 0.1563 0.08279
## Proportion of Variance 0.00051 0.00026 0.00023 0.0001 0.00003
## Cumulative Proportion  0.99938 0.99964 0.99987 1.0000 1.00000
```

As we can see, there are 10 principal components, and for each component, we have the `Proportion of Variance`. Use the function `extract_prop_var` below to extract this information from `fit`:

```r
extract_prop_var <- function(fit) {
    if (inherits(fit, "prcomp")) {
        result <- fit$sdev^2/sum(fit$sdev^2)
        names(result) <- colnames(fit$rotation)
    } else {
        stop("This function only works with the output of prcomp.")
    }
    return(result)
}
# All proportions
extract_prop_var(fit)
```

```
##          PC1          PC2          PC3          PC4          PC5
## 8.115669e-01 1.439729e-01 3.745910e-02 3.780982e-03 2.082723e-03
##          PC6          PC7          PC8          PC9         PC10
## 5.131703e-04 2.616439e-04 2.289692e-04 1.043846e-04 2.929297e-05
```

```r
# Or if you want just the first value
extract_prop_var(fit)["PC1"]
```

```
##       PC1
## 0.8115669
```

These proportions of variance are related to the eigenvalues of the sample covariance matrix, and they can be used to construct a test of independence between the variables in the data.

For this problem, we are interested in the first element of the vector above, which is the **proportion of variance explained by the first principal component**.

a. Use jackknife to estimate both the bias and standard error of this quantity.
b. Use bootstrap to estimate both the bias and standard error of this quantity.
c. Compare the jackknife 95% confidence interval with the basic bootstrap 95% confidence interval. Which one do you think is the most accurate? (*Hint*: Look at a histogram of the bootstrap samples.)

Problem 6

Conduct a Monte Carlo study to estimate the coverage probabilities of the standard normal bootstrap confidence interval, the basic bootstrap confidence interval, and the percentile confidence interval. Sample from a **log-normal** population of your choice and check the empirical coverage rates for the **sample mean**. This means you need to choose μ, σ^2, and the sample size n.

Find the proportion of times that the confidence intervals miss on the left (i.e. the true value is to the left of both bounds), and the proportion of times that the confidence intervals miss on the right (i.e. the true value is to the right of both bounds). Note that μ is **not** equal to the mean of a log-normal distribution.

Discuss the results.

(Bonus): Repeat for different sample sizes. What happens when n increases?

Problem 7

Efron and Tibshirani discuss the `scor` dataset (available in the package `bootstrap`) on 88 students who took examinations in five subjects. The first two tests (mechanics, vectors) were closed book and the last three tests (algebra, analysis, statistics) were open book.

Each row of the data frame is a set of scores $(x_{i1}, \ldots, x_{i5})$ for the i-th student.

a. Create a pairs plot for this dataset, in order to look at all pairwise comparisons. (See function `pairs`.)
b. Compare the plot with the sample correlation matrix.
c. Obtain bootstrap estimates of the standard errors for each of the following correlations: $\hat{\rho}_{12} = \rho(\text{mec}, \text{vec})$, $\hat{\rho}_{34} = \rho(\text{alg}, \text{ana})$, $\hat{\rho}_{35} = \rho(\text{alg}, \text{sta})$, and $\hat{\rho}_{45} = \rho(\text{ana}, \text{sta})$.