Applied Statistics 2013

Homework 4 - Due 8th of April

Exercise 1. ***Failure of the Bootstrap:*** *In this exercise we will see another example demonstrating the bootstrap can fail. Let $X_1, \ldots, X_n$ be i.i.d. samples from a uniform distribution over $[0, \theta]$, $\theta > 0$. For convenience let's denote this distribution by F_θ.*

a) Show that the maximum likelihood estimator for θ is given by $\hat{\theta}_n = X_{(n)} = \max_i X_i$.

b) Show that the distribution of $\hat{\theta}_n$ is

$$G(t) = \mathbb{P}_{F_\theta}(\hat{\theta}_n \leq t) = \begin{cases} 0 & \text{if } t < 0 \\ (t/\theta)^n & \text{if } 0 \leq t \leq \theta \\ 1 & \text{if } t > \theta \end{cases} .$$

c) Use this to derive the analytic expression for the variance of $\hat{\theta}_n$, that is $\mathbb{V}_{F_\theta}(\hat{\theta}_n)$.

d) Write a script that generates n observations from a uniformly distributed random variable on $[0, \theta]$ and implements the nonparametric bootstrap by drawing B bootstrap samples. Using this code with $n = 25$ and $B = 5000$, calculate $\hat{\theta}_n$ and also $\hat{\theta}^*_{n,b}$, $b = 1, \ldots, B$. Use the bootstrap samples to approximate $\mathbb{V}_{F_\theta}(\hat{\theta}_n)$ and compare it to your answer to (c). Repeat this experiment several times and experiment also with different values of B and n. What do you observe?

e) From your experiments in (d) you should notice that you are not able to use the non-parametric bootstrap to accurately estimate the variance. Let's try to understand what is going wrong. Here we'll use the notational convention of Theorems 3.19 and 3.21 of Wasserman's book. Define $T_n = n(\theta - \hat{\theta}_n)$ and $T^*_n = n(\hat{\theta}_n - \hat{\theta}^*_n)$ (the bootstrap analogue). Show that for $t \geq 0$

$$\mathbb{P}_{\hat{F}_n}(T^*_n \leq t) \geq \mathbb{P}_{\hat{F}_n}(T^*_n \leq 0) = 1 - (1 - 1/n)^n ,$$

the first inequality being trivial.

f) Show that

$$\liminf_{n \to \infty} \sup_t |\mathbb{P}_F(T_n \leq t) - P_{\hat{F}_n}(T^*_n \leq t)| \geq 1 - e^{-1} .$$

This means that the distribution of $\hat{\theta}_n$ computed by the non-parametric bootstrap procedure is significantly different than the real distribution. Refer back to question (d) and plot the histogram of $\hat{\theta}^*_{n,b}$ and check that this agrees with the above statement.

g) The parametric bootstrap generates bootstrap samples $X_1^*, \ldots, X_n^*$ by drawing from a uniform distribution on $[0, \hat{\theta}_n]$, instead of drawing from the empirical distribution function. Argue that

$$\sup_t |\mathbb{P}_F(T_n \leq t) - \mathbb{P}_{F_{\hat{\theta}_n}}(T_n^* \leq t)| \xrightarrow{P} 0 \ ,$$

as $n \to \infty$. Verify that this make sense by drawing $B = 5000$ bootstrap simulations.
Hint: *Show that T_n converges to an exponential distribution with mean θ.*

Exercise 2. *Exercise 15.8 from [Kvam and Vidakovic (2007)]:*

In a controlled clinical trial *Physician's Health Study I* which began in 1982 and ended in 1987, more that 22,000 physicians participated. The participants were randomly assigned to two groups: (i) *Aspirin* and (ii) *Placebo*, where the aspirin group have been taking 325 mg aspirin every second day. At the end of trial, the number of participants who suffered from Myocardial Infarction was assessed. The counts are given in the following table:

	MyoInf	No MyoInf	Total
Aspirin	104	10933	11037
Placebo	189	10845	11034

The popular measure in assessing results in clinical trials is Risk Ratio (RR) which is the ratio of proportions of cases (risks) in the two groups/treatments. From the table,

$$RR = R_a/R_p = \frac{104/11037}{189/11034} = 0.55.$$

Interpretation of RR is that the risk of Myocardial Infarction for the Placebo group is approximately $1/0.55 = 1.82$ times higher than that for the Aspirin group. With MATLAB, construct a bootstrap estimate for the variability of RR. *Hint:*

`The last sentence is to be replaced by:`
Use the bootstrap to compute the standard error for RR.

Exercise 3. *In this exercise we will give the necessary steps to deduce the bias corrected (BC) percentile confidence interval. Let $X_1, \ldots, X_n \sim F$ and suppose we wish to construct a confidence interval for $\theta = T(F)$. Let $\hat{\theta}_n = T(\hat{F}_n)$ be an estimator for θ ($\hat{F}_n$ denotes the empirical distribution function). Our approach is going to be very similar to the one for the percentile confidence intervals, but in this case we will take into account a possible bias. Let m be an increasing transformation (which may depend on the sample size n) and define $\hat{\psi}_n = m(\hat{\theta}_n)$, $\psi = m(\theta)$. Assume that*

$$\mathbb{P}_F(\hat{\psi}_n - \psi + b \leq x) = G(x) \ , \tag{1}$$

where b is a constant that may depend on F and n, and G is a continuous, invertible, and symmetric distribution around zero (meaning $G(x) = 1 - G(-x)$). Note that if $b = 0$ this is precisely the setup for deriving the percentile interval, so b plays the role of an unknown bias.

a) Show that if m and b were known, then a $100(1-\alpha)\%$-confidence interval for θ would be given by (L_n, U_n) where

$$L_n = m^{-1}(\hat{\psi}_n + G^{-1}(\alpha/2) + b) \text{ and } U_n = m^{-1}(\hat{\psi}_n - G^{-1}(\alpha/2) + b) \ .$$

b) Let $H(x) = \mathbb{P}_{\hat{F}_n}(\hat{\theta}_n^* \leq x)$ denote the bootstrap distribution of $\hat{\theta}_n^*$. Show that $H(L_n) \approx G\left(G^{-1}(\alpha/2) + 2b\right)$ and $H(U_n) \approx 1 - G\left(G^{-1}(\alpha/2) - 2b\right)$. Explain where the bootstrap approximation appears in the derivation.

c) Suppose we have constructed a large set of B bootstrap samples and computed $\hat{\theta}_{n,1}^*, \ldots, \hat{\theta}_{n,B}^*$. Conclude that an approximate bootstrap interval is given by $(\tilde{L}_n, \tilde{U}_n)$ where

$$\tilde{L}_n = [\hat{\theta}_n^*]_{(BG(G^{-1}(\alpha/2)+2b))} \ ,$$

and

$$\tilde{U}_n = [\hat{\theta}_n^*]_{(B(1-G(G^{-1}(\alpha/2)-2b)))} \ .$$

Here $[\hat{\theta}_n^*]_{(j)}$ is the j-th order statistic of $\hat{\theta}_{n,1}^*, \ldots, \hat{\theta}_{n,B}^*$.

d) Note that we still cannot use the above confidence interval, as neither b and G are known. Let's address the first problem: use the bootstrap approximation to show that

$$b \approx G^{-1}(H(\hat{\theta}_n)) \ .$$

Note that $H(\hat{\theta}_n)$ can be approximated by the fraction of bootstrap estimates $\hat{\theta}_{n,j}^*$ that are less of equal $\hat{\theta}_n$.

e) The BC-interval is obtained by setting $G = \Phi$, the cumulative distribution function of a standard normal random variable, and plugging in the above estimate of the bias, namely

$$\hat{b}^* = \Phi\left(\frac{1}{B}\sum_{i=1}^{B} \mathbf{1}\{\hat{\theta}_{n,j}^* \leq \hat{\theta}_n\}\right) \ .$$

Put all the pieces together to show that the Bias-Corrected confidence interval is given by

$$\left([\hat{\theta}_n^*]_{(B\Phi(\Phi^{-1}(\alpha/2)+2\hat{b}^*))}, [\hat{\theta}_n^*]_{(B(1-\Phi(\Phi^{-1}(\alpha/2)-2\hat{b}^*)))}\right) \ .$$

Remarks: *if we estimate b by zero, then the interval coincides with the percentile confidence interval you are familiar about from class. An improvement of the BC interval is the BC_a confidence interval (the a stands for accelerated). This is obtained by changing our initial assumption (1) to*

$$\mathbb{P}_F\left(\frac{\hat{\psi}_n - \psi}{1 + a\psi} + b \leq x\right) = \Phi(x) \ ,$$

and finding an appropriate value for a. We do not go into details of this derivation here, but these are similar in spirit to what you have done.

Exercise 4.

Suppose $X_1, \ldots, X_n$ is a random sample from an exponential distribution with parameter λ. So its density is given by $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$.

a) Show that the maximum likelihood estimator is given by $\hat{\lambda}_n = 1/\bar{X}_n$.

b) Show that $\sqrt{n}(\hat{\lambda}_n - \lambda) \xrightarrow{D} \mathcal{N}(0, \lambda^2)$. See for example chapter one in Wasserman. Conclude with the Delta-method that $\sqrt{n}(\log \hat{\lambda}_n - \log \lambda) \xrightarrow{D} \mathcal{N}(0, 1)$.

c) Show that an asymptotic CI for λ is given by

$$\left(\hat{\lambda}_n e^{-z_{\alpha/2}/\sqrt{n}}, \hat{\lambda}_n e^{z_{\alpha/2}/\sqrt{n}}\right) ,$$

with z_α denoting the $1-\alpha$ quantile of the standard normal distribution.

d) Note that $\lambda \sum_{i=1}^n X_i$ has a Gamma distribution with parameters n and 1. Deduce from this that an exact confidence interval for λ is given by

$$\left(\hat{\lambda}_n G^{-1}(\alpha/2)/n, \hat{\lambda}_n G^{-1}(1-\alpha/2)/n\right) ,$$

where G denotes the CDF of a Gamma distribution with parameters n and 1.

e) Do a simulation study to compare the coverage and length of the exact and asymptotic confidence intervals as derived above, and the BC_a confidence interval. The BC_a confidence intervals are easily obtained from the `boot`*-package. Consider sample sizes $n = 10, 25, 100$ and generate data from an exponential distribution with $\lambda = 5$. You can use the following code to help you.*

```
# obtain exact CI
ci.exact <- function(x,alpha)
{
   n <- length(x)
   ss <- sum(x)
   c(qgamma(alpha/2,n,1)/ss, qgamma(1-alpha/2,n,1)/ss)
}

# obtain asymptotic CI
ci.asympt <- function(x,alpha)
{
   d <- exp(qnorm(alpha/2,lower=F)/sqrt(n))
   av <- mean(x)
   c(1/(av*d), d/av)
}

# obtain nonparametric bootstrap CI (bca interval)
library(boot)
bootfun <- function(x,i)
```

```
{
   d <- x[i]
   1/mean(d)
}

ci.bca <- function(x,alpha)
{
   resboot <- boot(x,bootfun,R=1000)
   boot.ci(resboot,conf=1-alpha,type='bca')$bca[4:5]
}
```

f) Modify your code to assess also the length and coverage of other bootstrap confidence intervals, namely the normal, pivotal, and percentile intervals (using the R command `boot.ci` *these correspond respectively to the CI types* `normal,basic` *and* `percent`*). How do these compare with the BC_a intervals? Note that, regardless of the distribution, we are computing bootstrap confidence intervals for $1/\mathbb{E}[X_1]$. Repeat these experiments with distributions other than the exponential, and see how each bootstrap interval behaves.*

References

[Kvam and Vidakovic (2007)] Kvam, P.H. and Vidakovic, B. (2007) *Nonparametric Statistics with Applications to Science and Engineering*, Wiley.