Chapter 8. Nonparametric bootstrap

Bootstrap is a computational method for estimating standard errors and confidence intervals.

Let $X_1, \cdots, X_n \sim_{iid} F$. We use statistic

$$T = g(X_1, \cdots X_n)$$

for inference (i.e. estimation or testing). It is important to know, e.g. the standard deviation or the standard error of T.

Bootstrap idea: Let $\widehat{F}_n(x) = n^{-1} \sum_i I(X_i \leq x)$.

$$\begin{aligned} \text{Real world: } F &\longrightarrow X_1, \cdots, X_n \longrightarrow T = g(X_1, \cdots X_n) \\ \text{Bootstrap world: } \widehat{F}_n &\longrightarrow X_1^*, \cdots, X_n^* \longrightarrow T^* = g(X_1^*, \cdots X_n^*) \end{aligned}$$

Although we do not know F, $\widehat{F}_n$ is known. Therefore we know the distribution of T^* (in principle), which is taken as an approximation for the distribution of T. We compute the distribution of T^* by simulation.

8.1 Bootstrap variance estimation

Suppose we need to know variance $v = \text{Var}(T) = \text{Var}\{g(X_1, \cdots X_n)\}$. The bootstrap scheme below provides an estimator v^* for v.

---

1. Draw $X_1^*, \cdots X_n^*$ independently from $\widehat{F}_n$.
2. Compute $T^* = g(X_1^*, \cdots X_n^*)$.
3. Repeat Steps 1 & 2 B times, to obtain $T_1^*, \cdots, T_B^*$.
4. Compute the sample variance $v^* = (B-1)^{-1} \sum_{1 \le i \le B} (T_i^* - \bar{T}^*)^2$,

where $\bar{T}^* = B^{-1} \sum_{1 \le i \le B} T_i^*$.

Remark. Step 1 can be easily implemented in R. Let `X` be n-vector $(X_1, \cdots, X_n)$, then a bootstrap sample is obtained using `sample` as follows:

```
> Xstar <- sample(X, n, replace=T)
```

Bootstrap MSE estimation. Let $T = g(X_1, \cdots X_n)$ be an estimator for $\theta = \theta(F)$. Let

$$m = \text{MSE}(T) = E\{(T - \theta)^2\} = \text{Var}(T) + (ET - \theta)^2.$$

The bootstrap scheme below provides an estimator m^* for m.

---

1. Draw $X_1^*, \cdots X_n^*$ independently from $\widehat{F}_n$.
2. Compute $T^* = g(X_1^*, \cdots X_n^*)$.
3. Repeat Steps 1 & 2 B times, to obtain $T_1^*, \cdots, T_B^*$.
4. Compute the sample MSE

$$m^* = \frac{1}{B}\sum_{i=1}^{B}\{T_i^* - \theta(\widehat{F}_n)\}^2,$$

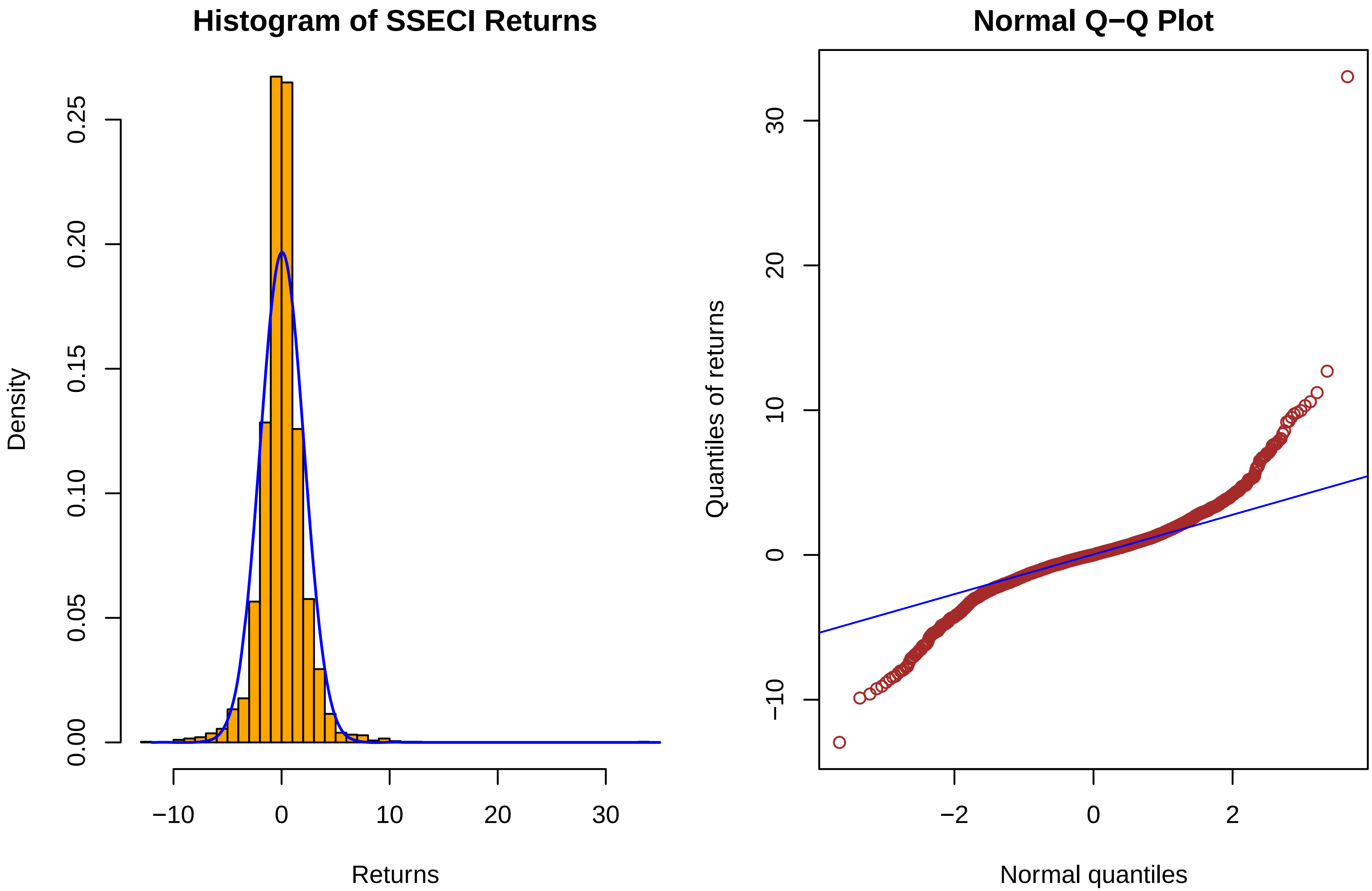
where $\widehat{F}_n(x) = n^{-1}\sum_i I(X_i \leq x)$ is the empirical distribution.

Example 1. Consider the daily returns of the Shanghai Stock Exchange Composite Index in December 1994 – September 2010

The data are saved in the file `shanghaiSECI.txt`. The sample size is $n = 3839$.

```
> x <- read.table("shanghaiSECI.txt", skip=3, header=T)
> x[1:4,]   # print out the first 4 rows
  idxcd          idxnmabbr             date            idxdret
1     8          SSE-Composite-Index   1994-12-08     -0.0165
2     8          SSE-Composite-Index   1994-12-09     -0.0014
3     8          SSE-Composite-Index   1994-12-12     -0.0085
4     8          SSE-Composite-Index   1994-12-13      0.0000
> dim(x)
[1] 3839    4
> y <- x[,4]*100   # daily return in percentages
> summary(y)
     Min.    1st Qu.    Median      Mean    3rd Qu.      Max.
```

```
-12.95000  -0.89000   0.01000   0.04994   0.96000  33.04000
> var(y)
[1] 4.111112
> hist(y, nclass=40, prob=T, main='Histogram of SSECI Returns')
> x <- seq(-12, 33, 0.1)
> lines(x, dnorm(x, mean(y), sqrt(var(y))), col='blue', lwd=2)
    # superimpose a normal PDF with the same mean/var onto the histogram
> qqnorm(y, xlab="Normal quantiles", ylab="Quantiles of returns")
                         # quantiles of the empirical distribution
                         # vs quantiles of Normal distribution
> qqline(y, col="blue")  # add a line passing through 1st and 3rd
                         # quartiles
```


Histogram of SSECI Returns
Density
0.00
0.05
0.10
0.15
0.20
0.25
−10
0
10
20
30
Returns
Normal Q−Q Plot
Quantiles of returns
−10
0
10
20
30
−2
0
2
Normal quantiles

The histogram shows that the returns do not follow a normal distribution, as the peak around 0 is much higher.

The Q-Q plot shows that both the tails of the return distribution are much heavier than the tails of normal distributions.

Recall: For any univariate CDF $F(\cdot)$, the quantile of F is defined as $F^{-1}(\alpha) = \inf\{x : F(x) \geq \alpha\},\ \alpha \in [0, 1]$.

Q-Q plot of two distributions F **and** G: the curves $\{(G^{-1}(\alpha), F^{-1}(\alpha)),\ \alpha \in [0, 1]\}$.

Lemma 1. Let F, G are two univariate CDFs, $b > 0$ and a are two constants. Then $G(x) = F(\frac{x-a}{b})$ for any x iff $G^{-1}(\alpha) = a + bF^{-1}(\alpha)$ for any $\alpha \in [0, 1]$.

Hence, a Q-Q plot is a straight line iff the two distributions are of the same form (i.e. one is a scale-location transformation of the other).

R-functions: `qqnorm`, `qqline`, `qqplot`

We introduce two measures related to the 3rd and 4th moments, which are often used as the measures for non-Gaussianality. Let $X \sim F$ and $E(X^4) < \infty$. Write $\mu = EX$ and $\sigma^2 = \text{Var}(X)$.

Skewness of F: $\gamma = E\{(X-\mu)^3\}/\sigma^3$.
Kurtosis of F: $\kappa = E\{(X-\mu)^4\}/\sigma^4$.

Remark. (i) The skewness is a measure for symmetry of distributions. If F is symmetric w.r.t the mean μ (such as $N(\mu, \sigma^2)$), $\gamma = 0$.

(ii) The kurtosis is a measure for tail-heaviness (i.e. fat-tails). For $N(\mu, \sigma^2)$, $\kappa = 3$. When $\kappa > (<) 3$, we say that the tails of F are heavier (lighter) than normal distributions.

(iii) Estimators for Skewness and Kurtosis: Let $\bar{X}$ and S^2 be the sample mean and the sample variance. Then

$$\widehat{\gamma} = \frac{1}{nS^3} \sum_{i=1}^{n} (X_i - \bar{X})^3, \qquad \widehat{\kappa} = \frac{1}{nS^4} \sum_{i=1}^{n} (X_i - \bar{X})^4.$$

Example 1 (Continue). We compute the estimates for skewness and kurtosis for the Shanghai SECI returns:

```
> mean((y-mean(y))^3) /var(y)^(1.5)
[1] 1.204415          # estimated skewness
> mean((y-mean(y))^4) /var(y)^2
[1] 25.05686         # estimated kurtosis
```

Since $\widehat{\gamma} = 1.204415 > 0$, the distribution is skewed to the right. The distribution is also heavy-tailed, since $\widehat{\kappa} = 25.05686$.

How accurate are those estimates? — use bootstrap to find the standard errors of the estimators.

```
> skew <- 1:1000
> kurt<- 1:1000
```

```
> for(i in 1:1000) {
+ ystar <- sample(y, 3839, replace=T)
+ skew[i] <- mean((ystar-mean(ystar))^3) /var(ystar)^(1.5)
+ kurt[i] <- mean((ystar-mean(ystar))^4) /var(ystar)^2
+ }
> sqrt(var(skew)); sqrt(var(kurt))
[1] 0.9514143  # bootstrap estimate for SE(estimated skewness)
[1] 13.96478   # bootstrap estimate for SE(estimated kurtosis)
```

Hence, the estimated skewness is 1.2044 with the standard error 0.9514, the estimated kurtosis is 25.06 with the standard error 13.97.

In the above we draw $B = 1000$ bootstrap samples. For this example, the results are insensitive for $B \geq 100$.

The analysis indicates that the returns are skewed to its right (unusual!) and heavy-tailed. Certainly their distribution is not normal.

8.2 Bootstrap confidence intervals

8.2.1 Approximate normal intervals

If $(\widehat{\theta}-\theta)/\{\text{Var}(\widehat{\theta})\}^{1/2} \xrightarrow{D} N(0,1)$, an approximate $(1-\alpha)$ confidence interval for θ is

$$\widehat{\theta} \pm Z_{\alpha/2}\{\text{Var}(\widehat{\theta})\}^{1/2},$$

where $Z_{\alpha/2}$ is the top-$\alpha/2$ point of $N(0,1)$.

However $\text{Var}(\widehat{\theta})$ is often unknown. Replacing it by its bootstrap estimate (see section 8.1 above), we obtain a bootstrap interval:

$$\widehat{\theta} \pm Z_{\alpha/2}\{\text{Var}(\theta^*)\}^{1/2}.$$

In practice, we repeat bootstrap sampling B times, obtaining bootstrap estimates $\theta_1^*, \cdots, \theta_B^*$. We take the sample variance of $\{\theta_1^*, \cdots, \theta_B^*\}$ as $\text{Var}(\theta^*)$.

8.2.2 Pivotal intervals

Let $X_1, \cdots, X_n$ be a sample from distribution F. We are interested in estimating a characteristics $\theta = \theta(F)$ (such as mean, skewness etc). Let $\widehat{\theta} = g(X_1, \cdots, X_n) = \theta(\widehat{F}_n)$ be the estimator for θ. Let r_α be the α-th percentile of the pivotal $\widehat{\theta} - \theta$, i.e.

$$\alpha = P(\widehat{\theta} - \theta \le r_\alpha).$$

Then

$$P(r_{\alpha/2} < \widehat{\theta} - \theta \le r_{1-\alpha/2}) = 1 - \alpha.$$

This gives a $(1 - \alpha)$-th confidence interval of θ:

$$(\widehat{\theta} - r_{1-\alpha/2}, \;\; \widehat{\theta} - r_{\alpha/2}).$$

This is a valid interval estimation if r_α does not depend on θ, i.e. the distribution of the pivotal $\widehat{\theta} - \theta$ does not depend on θ. However this requirement is not necessary if we adopt a bootstrap approach.

Under some standard conditions,

$$P(\widehat{\theta} - \theta < r) \approx P(\theta^* - \widehat{\theta} < r \,\Big|\, X_1, \cdots, X_n)$$

when n is large, where $\theta^* = g(X_1^*, \cdots, X_n^*)$. Thus we may replace $r_{\alpha/2}$ and $r_{1-\alpha/2}$ by their bootstrap counterparts as follows:

> Repeat bootstrap sampling B times to form estimates $\theta_1^*, \cdots, \theta_B^*$. Let θ_α^* be the $[B\alpha]$-th smallest value among $\theta_1^*, \cdots, \theta_B^*$, where $[B\alpha]$ denotes the integer part of $B\alpha$ (i.e. $[a]$ is the largest integer smaller than a). Then
>
> $$r^*_{\alpha/2} = \theta^*_{\alpha/2} - \widehat{\theta}, \qquad r^*_{1-\alpha/2} = \theta^*_{1-\alpha/2} - \widehat{\theta}.$$

The $(1-\alpha)$ bootstrap pivotal interval for θ is:

$$(2\widehat{\theta} - \theta^*_{1-\alpha/2},\ \ 2\widehat{\theta} - \theta^*_{\alpha/2})$$

8.2.3 Percentile intervals

The $(1-\alpha)$ bootstrap percentile interval for θ is:

$$(\theta^*_{\alpha/2}, \quad \theta^*_{1-\alpha/2})$$

Example 2. We calculate the three bootstrap intervals for the median of the salary for the graduates in a business school based on data in `Jobs.txt` using the following R function:

```
jobsMedianCIs <- function(alpha, B) {
jobs <- read.table("Jobs.txt", header=T, row.names=1)
y <- jobs[,4]  # salary data
cat("Point estimate for median of salaries:", median(y), "\n\n")
my <- 1:B
for(i in 1:B) {
```

```
ystar <- sample(y, 253, replace=T)  # draw bootstrap sample
my[i] <- median(ystar) # bootstrap estimate for median
}
my <- sort(my)   # sort bootstrap estimates in ascending order
i <- as.integer(alpha*B/2)  # i=[B x alpha/2]
cat(1-alpha, "Bootstrap confidence intervals for median of salaries", "\n")
cat("Normal interval:", median(y)-qnorm(1-alpha/2)*sqrt(var(my)),
           median(y)+qnorm(1-alpha/2)*sqrt(var(my)), "\n")
cat("Pivotal interval:", 2*median(y)-my[B-i], 2*median(y)-my[i], "\n")
cat("Percentile interval:", my[i], my[B-i], "\n")
}
```

Calling `jobsMedianCIs(0.05, 5000)`, we obtain the results below. Note that the three intervals for this example are very similar.

```
Point estimate for median of salaries: 47

0.95 Bootstrap confidence intervals for median of salaries
Normal interval: 45.72511 48.27489
Pivotal interval: 46 48
Percentile interval: 46 48
```

Example 1 (Continue). We calculate the three bootstrap intervals for the skewness using the following *R*-function:

```
SSECIbootstrapCIs <- function(B) {
x <- read.table("shanghaiSECI.txt", skip=3, header=T)
y <- x[,4]*100
skew0 <- mean((y-mean(y))^3) /var(y)^(1.5)
cat("Point estimate for skewness:", skew0, "\n\n")
skew <- 1:B
for(i in 1:B) {
ystar <- sample(y, 3839, replace=T)   # draw bootstrap sample
skew[i] <- mean((ystar-mean(ystar))^3) /var(ystar)^(1.5)
}
skew <- sort(skew)   # sort the data in ascending order
i <- as.integer(0.025*B)  # i =[0.025B]
cat("95% Bootstrap confidence intervals for skewness", "\n")
cat("Normal interval:", skew0-2*sqrt(var(skew)),
                  skew0+2*sqrt(var(skew)), "\n")
cat("Pivotal interval:", 2*skew0-skew[B-i], 2*skew0-skew[i], "\n")
cat("Percentile interval:", skew[i], skew[B-i], "\n")
}
```

Call `SSECIbootstrapCIs(1000)`, yielding the following output:

```
Point estimate for skewness: 1.204415

95% Bootstrap confidence intervals for skewness
Normal interval: -0.6486737 3.057503
Pivotal interval: -0.6067615 2.577141
Percentile interval: -0.1683116 3.015591
```

Final Remark. All the bootstrap intervals work well when $\widehat{\theta}$ is asymptotically normal.
