Catam Project Report
PartII Additional Projects
(July 2020 Edition)

Freud Fan

19/03/2021

1 Bootstrap Estimation of Standard Error

1.1 Question 1

Assume $X_i \neq X_j$ if $i \neq j$. Let $\Omega \supset A = \{x_i\}$,

$$\mathbf{P}(Y_j = x_i) = \hat{F}(\{x_i\}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}_{\{X_i = x_i\}} = \begin{cases} 1/n, & \text{if } \mathrm{X}_i = x_i \text{ for some i} \\ 0, & \text{if } \mathrm{X}_i \neq x_i \text{ for all i} \end{cases}$$

for all i, j. I.e., $\mathbf{Y}$ is the same as a random sample of size n, drawn with replacement from the actual sample $\mathbf{X}$. We have

$$\mathbf{E}(Y_j) = \sum_{i=1}^{n} x_i \mathbf{P}(X_i = x_i) = \frac{1}{n}\sum_{i=1}^{n} x_i \xrightarrow[SLLN]{a.s.} \mathbf{E}(X_i) \quad as\ n \to \infty$$

and

$$\begin{aligned} Var(Y_j) =& \mathbf{E}(Y_j^2) - \mathbf{E}^2(Y_j) = \sum_{i=1}^{n} x_i^2 \mathbf{P}(Y_j = x_i) - \mathbf{E}^2(Y_j) \\ \xrightarrow[CMT]{a.s.}& \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \mathbf{E}^2(X_i) = Var(X_i) \quad as\ n \to \infty. \end{aligned}$$

Hence $\mathbf{Y}$ is asymptotically unbiased with an asymptotically equal variance - the bootstrap estimate is reasonable.

1.2 Question 2

1.2.1 Estimating T's Distribution: Programming Task

A program in R using the bootstrap method is listed on page 8, named **histo(n,m)**. (Data set imported from *II-10-3-2020.csv*.) n is the number of copies of Y_i, m is the total frequency of the experiment in terms of T(Y_b).

The result is in Fig.1, taking n=200, m=1000. These bootstrap values are normally distributed with the mean at around 0.525.

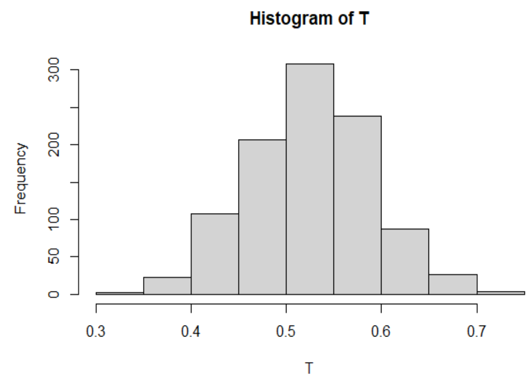

Figure 1: The histogram of the bootstrap values T($\mathbf{Y}_b$) with n=200, m=1000

1.3 Question 3

1.3.1 Estimating $\hat{\sigma}$'s Distribution: Programming Task

A program using the bootstrap method is listed on page 8, named **histo2(n,B,m)**. n is the number of copies of Y_i, B is the number of independent bootstrap samples $\mathbf{Y}_b$, m is the total frequency of the experiment in terms of $\hat{\sigma}_B$.

Fix n=200, m=1000. See results in Fig.2-5.

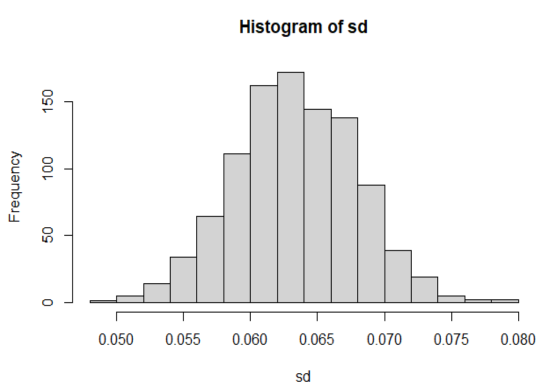

Figure 2: B=100

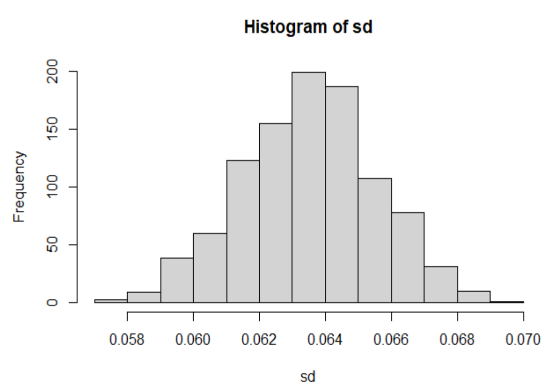

Figure 3: B=500

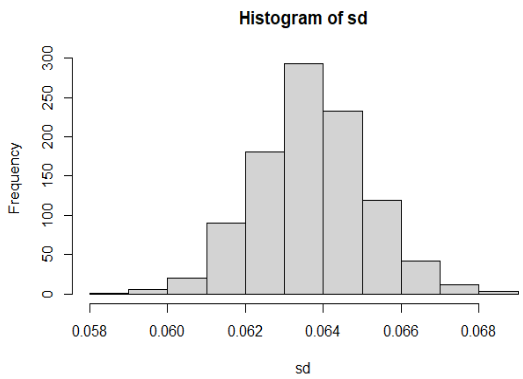

Figure 4: B=1000

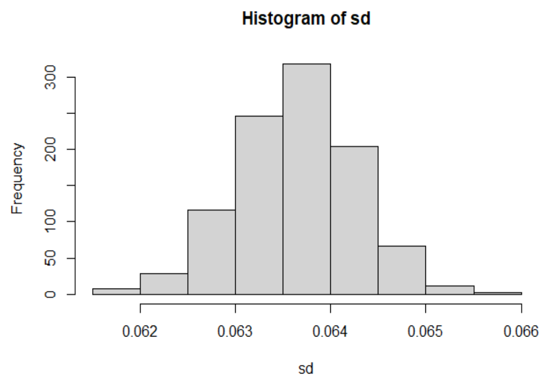

Figure 5: B=5000

As B increases, more $\hat{\sigma}_B$ accumulate at around 0.0635 with less deviation (spotted from broader bar with smaller range), i.e, $\hat{\sigma}_B \to \hat{\sigma} \approx 0.0635$. Large B with reasonable operating time is preferred for estimation (according to central limit theorem), e.g., B=1000.

1.4 Question 4

Fix B=1000, m=1000. Run the program **histo2** with various n. Results in Tbl.1.

n	100	500	1000	5000
Observed sd $\hat{\sigma}_B$ (approx.)	0.091	0.040	0.02825	0.01265
Theoretical sd $\sigma(= 1/\sqrt{n-3})$	0.10153	0.04486	0.03167	0.01415

Table 1: Comparison of values of $\hat{\sigma}_B$ and σ for various n

σ corrected to 5 decimal places. Given IQ scores are normally distributed, from the table we can see consistency in the trend and value of numerical $\hat{\sigma}_B$ and theoretical σ. Hence there is no sufficient evidence to reject the bivariate normal hypothesis.

1.5 Question 5

Notice for $\hat{\sigma}$, the expectation and variance summing over x_i, hence WLOG, reorder each sample $\mathbf{X}$ in ascending order, i.e.,

$$x_1 \leq x_2 \leq ... \leq x_n.$$

Then

$$\begin{aligned}
\mathbf{P}(T(\mathbf{Y}) = x_i) &= \mathbf{P}(max\{Y_1, ..., Y_n\} = x_i) \\
&= \mathbf{P}(max\{Y_1, ..., Y_n\} \leq x_i) - \mathbf{P}(max\{Y_1, ..., Y_n\} \leq x_{i-1}) \\
&\overset{iid}{=} \prod_{j=1}^{n} \mathbf{P}(Y_j \leq x_i) - \prod_{j=1}^{n} \mathbf{P}(Y_j \leq x_{i-1}) \\
&\overset{iid}{=} (\frac{i}{n})^n - (\frac{i-1}{n})^n \quad since \quad \mathbf{P}(Y_j = x_i) = \frac{1}{n} \quad \forall j.
\end{aligned}$$

Calculate

$$\mathbf{E}(T(\mathbf{Y})) = \sum_{i=1}^{n} x_i \mathbf{P}(T(\mathbf{Y}) = x_i) = \sum_{i=1}^{n} x_i((\frac{i}{n})^n - (\frac{i-1}{n})^n)$$

$$\mathbf{E}(T(\mathbf{Y})^2) = \sum_{i=1}^{n} x_i^2 \mathbf{P}(T(\mathbf{Y}) = x_i) = \sum_{i=1}^{n} x_i^2((\frac{i}{n})^n - (\frac{i-1}{n})^n)$$

hence

$$\begin{aligned}
\sigma(T; \hat{F}) &= \sqrt{var_{\hat{F}} T(\mathbf{Y})} \\
&= \sqrt{\mathbf{E}(T(\mathbf{Y})^2) - \mathbf{E}^2(T(\mathbf{Y}))} \\
&= \sqrt{\sum_{i=1}^{n} x_i^2((\frac{i}{n})^n - (\frac{i-1}{n})^n) - (\sum_{i=1}^{n} x_i((\frac{i}{n})^n - (\frac{i-1}{n})^n))^2}.
\end{aligned}$$

1.6 Question 6

For $X_i \sim U[0, \theta]$, $\theta > 0$, we have the cdf

$$\begin{aligned}
\mathbf{P}(T(\mathbf{X}) \leq t) &= \mathbf{P}(max\{X_1, ..., X_n\} \leq t) \\
&\overset{iid}{=} \prod_{i=1}^{n} \mathbf{P}(X_i \leq t) \\
&= (\frac{t}{\theta})^n \quad for \ 0 \leq t \leq \theta,
\end{aligned}$$

i.e.,

$$F(t) = \begin{cases} 0 & t < 0 \\ (\frac{t}{\theta})^n & 0 \leq t \leq \theta \\ 1 & t > \theta \end{cases}$$

and the pdf

$$f(t) = F'(t) = \frac{nt^{n-1}}{\theta^n}\ \mathbb{I}_{\{t\in[0,\theta]\}}.$$

Calculate

$$\begin{aligned}
\mathbf{E}(T(\mathbf{X})) &= \int_{-\infty}^{\infty} tf(t)dt = \int_0^\theta n(\frac{t}{\theta})^n dt = \frac{n\theta}{n+1} \\
\mathbf{E}(T(\mathbf{X})^2) &= \int_{-\infty}^{\infty} t^2 f(t)dt = \int_0^\theta nt(\frac{t}{\theta})^n dt = \frac{n\theta^2}{n+2}
\end{aligned}$$

hence

$$\begin{aligned}
\sigma(T;F) &= \sqrt{var_F T(\mathbf{X})} \\
&= \sqrt{\mathbf{E}(T(\mathbf{X})^2) - \mathbf{E}^2(T(\mathbf{X}))} \\
&= \sqrt{\frac{n\theta^2}{n+2} - \frac{n^2\theta^2}{(n+1)^2}} \\
&= \frac{\theta}{n+1}\sqrt{\frac{n}{n+2}}.
\end{aligned}$$

1.7 Question 7

1.7.1 Generating Bootstrap Samples: Programming Task

A program generating sample $\mathbf{X}$ is listed on page 9, named **bs(n)**. n is the sample size of $\mathbf{X}$ under uniform distribution U[0,5]. A set of 50 $\hat{\sigma}$ is performed for each n to calculate the average $\hat{\sigma}$.

n	100	500	1000	5000
$\hat{\sigma}_1$	0.0294645	0.006942944	0.004091652	0.0007380898
$\hat{\sigma}_2$	0.06493861	0.01053255	0.002186515	0.0002077604
average $\hat{\sigma}$	0.04518773	0.0103667	0.004796646	0.0009716006
σ	0.04901721	0.00996014	0.004990017	0.0009996001
percentage error	7.81	4.08	3.88	2.80

Table 2: Comparison of values of $\hat{\sigma}$ and σ for various n

$\hat{\sigma}$ corrected up to computer precision. Percentage error calculated by $\frac{|\bar{\hat{\sigma}}-\sigma|}{\sigma}\times 100\%$, corrected to 2 decimal places. From Tbl.2 we can see that the bootstrap method is unstable when computing $\hat{\sigma}$. Its average, however, gives a good approximation of σ with an overall decreasing percent error (from 7.81% to 2.80%) as n increases (from 100 to 5000), showing a trend of converging to σ.

The instability of the measure is inherited from the bias-variance trade-off. T($\mathbf{X}$) is an asymptotic unbiased mle, i.e.,

$$\mathbf{E}(T(\mathbf{X})) = \frac{n\theta}{n+1} \to \theta \quad as\ n \to \infty.$$

Hence there is a possible error coming from the spread of the data, resulting in the unsatisfying behavior of the bootstrap method, while small n enhances the instability.

2 Programs

Note: Some programs listed on this pdf have *'return'* added after excessively long texts, which needs to be removed before testing.

2.1 Question 2

2.1.1 histo(n,m)

```
mydata<-read.csv(file.choose())
histo<-function(n,m){
  T<-c(0,m);I<-c(1:m)
  for (i in I){
    Xmat<-as.matrix(mydata)
    Yb<-Xmat[sample(nrow(Xmat),n,replace=TRUE),]
    VIQ<-Yb[,1];PIQ<-Yb[,2]
    r<-cov(VIQ,PIQ)/(sqrt(var(VIQ)*var(PIQ)))
    T[i]<-log((1+r)/(1-r))/2
  }
  hist(T)
}
```

2.2 Question 3

2.2.1 histo2(n,B,m)

```
mydata<-read.csv(file.choose())
histo2<-function(n,B,m){
  sd<-c(0,m);J<-c(1:m)
  for (j in J){
    T<-c(0,B);I<-c(1:B)
    for (i in I){
      Xmat<-as.matrix(mydata)
      Yb<-Xmat[sample(nrow(Xmat),n,replace=TRUE),]
      VIQ<-Yb[,1];PIQ<-Yb[,2]
      r<-cov(VIQ,PIQ)/sqrt(var(VIQ)*var(PIQ))
      T[i]<-log((1+r)/(1-r))/2
      }
```

```
    sd[j]<-sd(T)
  }
  hist(sd)
}
```

2.3 Question 7

2.3.1 bs(n)

```
bs<-function(n){
  X=runif(n,0,5)
  X=sort(X, decreasing = FALSE)
  M=integer(n)
  for (i in c(1:n)){
    M[i]=(i/n)^n-((i-1)/n)^n
  }
  D=(sum(X^2*M)-sum(X*M)^2)^0.5
  D
}
```