Estimating CDF, Statistical Functionals and Nonparametric Bootstrap

Botond Szabo

Leiden University

Leiden, 18 March 2019
Outline

1. EDF
2. Statistical functionals
3. Nonparametric Bootstrap
4. Quiz
Empirical distribution function

**Definition**

Let $X_1, \ldots, X_n \overset{iid}{\sim} F$, where $F$ is a CDF on the real line. The **empirical distribution function** (EDF) is the CDF that puts mass $1/n$ at each data point. Formally,

\[ \hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x), \]

where

\[ I(X_i \leq x) = \begin{cases} 
1 & \text{if } X_i \leq x \\
0 & \text{if } X_i > x.
\end{cases} \]
Example

Cox and Lewis (1966) reported 799 waiting times between successive pulses along a nerve fiber. The solid line in the next plot is the corresponding EDF. The data points are represented by small vertical lines at the bottom.
Properties of EDF

**Theorem**

*At any fixed value of* $x$,*

$$
\mathbb{E}_F[\hat{F}_n(x)] = F(x),
$$

$$
\hat{F}_n(x) \xrightarrow{\mathbb{P}_F} F(x).
$$
Properties of EDF

**Theorem**

*At any fixed value of* \( x \),

\[
\mathbb{E}_F[\hat{F}_n(x)] = F(x),
\]

\[
\hat{F}_n(x) \xrightarrow{P_F} F(x).
\]

**Theorem**

*(Glivenko-Cantelli theorem)* Let \( X_1, \ldots, X_n \sim F \). Then

\[
\sup_x |\hat{F}_n(x) - F(x)| \xrightarrow{P_F} 0.
\]
Statistical functionals

**Definition**

A *statistical functional* $T(F)$ is any function of $F$.

**Example**

The mean $\mu = \int x dF(x)$, the variance $\sigma^2 = \int (x - \mu)^2 dF(x)$ and the median $m = F^{-1/2}(1/2)$ are all examples of statistical functionals.
Plug-in estimators and linear functionals

Definition

The plug-in estimator of $\theta = T(F)$ is $\hat{\theta}_n = T(\hat{F}_n)$.

Definition

If $T(F) = \int r(x)dF(x)$ for some function $r$, then $T$ is called a linear functional.

Lemma

If $T$ is a linear functional, then for any distribution functions $F$ and $G$ and any numbers $a$ and $b$

$$T(aF + bG) = aT(F) + bT(G).$$
Theorem

The plug-in estimator for a linear $T(F) = \int r(x)dF(x)$ is

$$T(\hat{F}_n) = \int r(x)d\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} r(X_i).$$
Confidence intervals

- Assume we know how to estimate the standard error of \( T(\hat{F}_n) \).
- In many cases, \( T(\hat{F}_n) \approx N(T(F), \hat{se}^2) \).
- An approximate \( 1 - \alpha \) confidence interval for \( T(F) \) is then
  \[
  T(\hat{F}_n) \pm z_{\alpha/2} \hat{se}.
  \]
  We will call it a Normal-based confidence interval.
- For \( \alpha = 0.05 \), \( z_{\alpha/2} = 1.96 \approx 2 \), so
  \[
  T(\hat{F}_n) \pm 2\hat{se}
  \]
  is an approximate 95% confidence interval.
Let $\mu = T(F) = \int x dF(x)$. The plug-in estimator of $\mu$ is

$$
\hat{\mu}_n = T(\hat{F}_n) = \int x d\hat{F}_n(x) = \bar{X}_n.
$$

The standard error of this estimator is

$$
se = \sqrt{\text{Var}[\bar{X}_n]} = \frac{\sigma}{\sqrt{n}}.
$$

Suppose we have an estimator of $\sigma$, called $\hat{\sigma}$. Then $\hat{se} = \hat{\sigma} / \sqrt{n}$, and a Normal-based confidence interval for $\mu$ is

$$
\bar{X}_n \pm z_{\alpha/2} \hat{se}.
$$
Example

Let \( \sigma^2 = T(F) = \int (x - \mu)^2 dF(x) \). We have
\[
\sigma^2 = \int x^2 dF(x) - \left( \int x dF(x) \right)^2 .
\]
The plug-in estimator of \( \sigma^2 \) is
\[
\hat{\sigma}^2 = \int x^2 d\hat{F}_n(x) - \left( \int x d\hat{F}_n(x) \right)^2
= \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \left( \frac{1}{n} \sum_{i=1}^{n} X_i \right)^2
= \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 .
\]
Another good estimator of $\sigma^2$ is the sample variance

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2.$$ 

In practice there is little difference between this and

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2,$$

except that the former is unbiased: $\mathbb{E}[S_n^2] = \sigma^2$.

Returning to estimation of $\mu$, we see that we can use $\hat{\sigma}/\sqrt{n}$ or $S_n/\sqrt{n}$ to estimate se of $\hat{\mu}_n$. 
We return to the nerve data example. Suppose we want to estimate the mean of $F$ and construct the 95% confidence interval. Now, with a bit of coding in $\mathbf{R}$, $\hat{\mu}_n = 0.2185732$, $\hat{\text{se}} = 0.00740056$, and the 95% confidence interval is $[0.2037721, 0.2333743]$. 
The bootstrap is a method for estimating standard errors and computing confidence intervals.

Let \( T_n = g(X_1, \ldots, X_n) \) be a statistic based on an IID sample \( X_1, \ldots, X_n \sim F \).

Suppose we want to know the variance of \( T_n \), \( \mathbb{V}_F[T_n] \).

We have written \( \mathbb{V}_F \) to emphasise dependence on the unknown CDF \( F \).

Example: if \( T_n = \bar{X}_n \), then \( \mathbb{V}_F(T_n) = \sigma^2/n \) with

\[
\sigma^2 = \int (x - \mu)^2 dF(x), \quad \mu = \int x dF(x).
\]

Thus \( \mathbb{V}_F(T_n) \) is a function of \( F \).

But \( F \) is unknown.
Informal description

- The bootstrap has to steps.
  1. **Estimate** $\mathbb{V}_F(T_n)$ with $\mathbb{V}_{\hat{F}_n}(T_n)$.
  2. If $\mathbb{V}_{\hat{F}_n}(T_n)$ is not explicitly computable, **approximate it with simulation**.

- For $T_n = \bar{X}_n$, we have that $\mathbb{V}_{\hat{F}_n}(T_n) = \hat{\sigma}^2/n$, where

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2.$$

  In this case Step 1 of the bootstrap is enough.

- In more complicated cases we cannot write down a formula for $\mathbb{V}_{\hat{F}_n}(T_n)$, which is why we need Step 2 (simulation).
Bootstrap variance estimation

- We want to approximate $\nabla \hat{F}_n[T_n(X_1, \ldots, X_n)]$ through simulation.

- We simulate $X_1^*, \ldots, X_n^*$ from $\hat{F}_n$ and compute $T_n^* = g(X_1^*, \ldots, X_n^*)$. This constitutes one draw from the distribution of $T_n$. Call it $T_{n,1}^*$.

- We repeat this process over and over again and compute $T_{n,2}^*, \ldots, T_{n,B}^*$. Then we set

$$v_{\text{boot}} = \frac{1}{B} \sum_{b=1}^{B} \left( T_{n,b}^* - \frac{1}{B} \sum_{r=1}^{B} T_{n,r}^* \right)^2.$$  

This is a bootstrap estimate of the variance of $T_n$. 
Simulation from $\hat{F}_n$

- Note that $\hat{F}_n$ puts mass $1/n$ at each data point $X_1, \ldots, X_n$.
- **Drawing** an observation from $\hat{F}_n$ is equivalent to drawing one point at random from $\{X_1, \ldots, X_n\}$.
- Thus to simulate $X_1^*, \ldots, X_n^* \sim \hat{F}_n$, it suffices to draw $n$ observations with replacement from $X_1, \ldots, X_n$. 
Example

**Bootstrap for the median**

Given data $X = (X(1), \ldots, X(n))$:

$T \leftarrow \text{median}(X)$

$T_{boot} \leftarrow \text{vector of length } B$

for $i = 1$ to $B$ do

    $X_{star} \leftarrow \text{sample of size } n \text{ from } X \text{ (with replacement)}$

    $T_{boot}[i] \leftarrow \text{median}(X_{star})$

end for

$se \leftarrow \text{sqrt(variance}(T_{boot}))$
Example

Consider the nerve data and suppose we want to estimate the skewness

\[ \theta = T(F) = \frac{1}{\sigma^3} \int (x - \mu)^3 dF(x). \]

The skewness measures assymetry of the distribution. For the normal distribution the skewness is zero.

The plug-in estimator is

\[ \hat{\theta} = T(\hat{F}_n) = \frac{1}{\hat{\sigma}^3} \int (x - \hat{\mu})^3 d\hat{F}_n(x) = \frac{\sum_{i=1}^{n} (X_i - \bar{X}_n)^3}{\hat{\sigma}^3} = 1.76. \]

To estimate the standard error, we use bootstrap with \( B = 1000 \) replications, yielding an estimated standard error of 0.16.
Methods for confidence interval

- **Normal method:**
  - The simplest method for constructing a bootstrap confidence interval is the normal interval

  \[ T_n \pm z_{\alpha/2} \hat{s}_{\text{e boot}}, \]

  where \( \hat{s}_{\text{e boot}} = \sqrt{v_{\text{boot}}} \) is the bootstrap estimate of the standard error.

  - The method is not accurate. unless the distribution of \( T_n \) is close to normal.
Methods for confidence interval

- **Normal method:**
  - The simplest method for constructing a bootstrap confidence interval is the normal interval
  
  \[ T_n \pm z_{\alpha/2} \hat{s}_\text{e}_{\text{boot}}, \]

  where \( \hat{s}_\text{e}_{\text{boot}} = \sqrt{v_{\text{boot}}} \) is the bootstrap estimate of the standard error.
  - The method is not accurate. unless the distribution of \( T_n \) is close to normal.

- Percentile interval
  - The bootstrap percentile interval is defined by
  
  \[ C_n = (\theta_{\alpha/2}^*, \theta_{1-\alpha/2}^*). \]
Question 1

Which of the following statements is not correct about the empirical distribution function (EDF)?

Answers:

1. It puts $1/n$ weight to each observation.
2. It is a consistent estimator of the cumulative distribution function (CDF).
3. It converges to the CDF in every point $x$ simultaneously.
4. It is continuous.
Question 1

Which of the following statements is not correct about the empirical distribution function (EDF)?

Answers:

1. It puts $1/n$ weight to each observation.
2. It is a consistent estimator of the cumulative distribution function (CDF).
3. It converges to the CDF in every point $x$ simultaneously.
4. It is continuous.
Question 2

How to construct an estimator for the mean if we do not know the distribution family?

Answers:

1. Use MM estimator.
2. Use the ML estimator.
3. Use nonparametric plug-in estimator.
4. Use the CLT.
Question 2

How to construct an estimator for the mean if we do not know the distribution family?

Answers:

1. Use MM estimator.
2. Use the ML estimator.
3. Use nonparametric plug-in estimator.
4. Use the CLT.
Question 3

How to construct the $1 - \alpha$ confidence interval for the statistical functional $T(F)$, if we can estimate the standard error?

Answers:

1. $T(\hat{F}) \pm z_{\alpha/2} \hat{se}$.
2. $T(\hat{F}) \pm 2 \hat{se}$.
3. $\bar{X}_n \pm z_{\alpha/2} \hat{se}$.
4. $\bar{X}_n \pm sd(X)$. 
Question 3

How to construct the $1 - \alpha$ confidence interval for the statistical functional $T(F)$, if we can estimate the standard error?

Answers:

1. $T(\hat{F}) \pm z_{\alpha/2} \hat{se}$.
2. $T(\hat{F}) \pm 2\hat{se}$.
3. $\bar{X}_n \pm z_{\alpha/2} \hat{se}$.
4. $\bar{X}_n \pm sd(X)$. 
Question 4

Which of the following is not true for the nonparametric bootstrap?

Answers:

1. It is used to construct confidence intervals and estimating the standard error.
2. We are resampling from the observation $X_1, \ldots, X_n$.
3. We estimate the parameters of the distribution with the help of it.
4. Useful when estimator of the variance of the functional $\hat{V}_{f_n}(T_n)$ is not computable.
Question 4

Which of the following is not true for the nonparametric bootstrap?

Answers:

1. It is used to construct confidence intervals and estimating the standard error.
2. We are resampling from the observations $X_1, \ldots, X_n$.
3. We estimate the parameters of the distribution with the help of it.
4. Useful when estimator of the variance of the functional $\hat{V}_{F_n}(T_n)$ is not computable.
Question 5

How do we compute $T_n^*$ for $T_n = g(X_1, ..., X_n)$?

Answers:

1. $T_n^* = g(X_1^*, ..., X_n^*)$.
2. $T_n^* = g(X_1^*, ..., X_m^*)$.
3. $T_n^* = g(X_1, ..., X_n)$.
4. $T_n^* = \bar{X}_n$. 
Question 5

How do we compute $T^*_n$ for $T_n = g(X_1, ..., X_n)$?

Answers:

1. $T^*_n = g(X^*_1, ..., X^*_n)$.
2. $T^*_n = g(X^*_1, ..., X^*_m)$.
3. $T^*_n = g(X_1, ..., X_n)$.
4. $T^*_n = \bar{X}_n$. 