In section so far, we have only worked with parametric distributions (i.e., where we assumed the specific distribution of the data). Now we will summarize nonparametric distributions and how we can estimate their parameters.

1 Nonparametric Estimation (Module 3)

1.1 Set-up

Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F$, where F can be parametric or nonparametric. The goal is to estimate some functional $T(F) = \theta$. A natural estimate is $\hat{\theta} = T(\hat{F}_n)$, where $\hat{F}_n$ is the empirical CDF, i.e.,

$$\begin{aligned} \hat{F}_n(x) &= \frac{\sum_{i=1}^n 1\{X_i \leq x\}}{n} \\ &= \frac{\#\{X_i \leq x\}}{n}. \end{aligned}$$

Note that this is an RV! We call it the plug-in estimator of $F(x)$.

Let X^* denote an RV distributed as $\hat{F}_n$; that is, we draw a values from $X_1, \ldots, X_n$, with each value selected with equal probability $1/n$. We can write probability and expectation statements using $\hat{F}_n$:

$$\begin{aligned} &\mathbb{P}_{\hat{F}_n}(X^* \leq x) \\ &\mathbb{E}_{\hat{F}_n}(X^*). \end{aligned}$$

What do these statements really mean? They are

$$\begin{aligned} &\mathbb{P}(X^* \leq x | X^* \sim \hat{F}_n, X_1, \ldots, X_n) \\ &\mathbb{E}(X^* | X^* \sim \hat{F}_n, X_1, \ldots, X_n). \end{aligned}$$

Keep in mind that these quantities are random, just like $\hat{F}_n$, since they are *conditional* on the data $X_1, \ldots, X_n$.

The *unconditional* distribution of X^* is the distribution of X^* considering both the randomness of sampling from $\hat{F}_n$ and from sampling $X_1, \ldots, X_n$. We can write the unconditional probability and expectation as $P(X^*)$ and $\mathbb{E}(X^*)$ (which are non-random). Some very helpful properties here are the Laws of Total Expectation and Variance:

$$\begin{aligned} \mathbb{E}(X^*) &= \mathbb{E}\left(\mathbb{E}(X^* | X_1, \ldots, X_n)\right) \\ \text{Var}(X^*) &= \mathbb{E}\left(\text{Var}(X^* | X_1, \ldots, X_n)\right) + \text{Var}\left(\mathbb{E}(X^* | X_1, \ldots, X_n)\right) \end{aligned}$$

1.2 What can we say about the distribution of $\hat{F}_n$?

For a fixed point x, define

$$\begin{aligned} Y_i(x) &= 1\{X_i \leq x\} \text{ which implies} \\ \hat{F}_n(x) &= \frac{\sum_{i=1}^n Y_i(x)}{n}. \end{aligned}$$

Observe that $Y_1(x), Y_2(x), \ldots, Y_n(x) \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$, where

$$p = \mathbb{P}(Y_i(x) = 1) = \mathbb{P}(X_i \leq x) = F(x).$$

Thus,

$$n\hat{F}_n(x) = \sum_{i=1}^n Y_i(x) \sim \text{Binomial}(n, F(x)).$$

We might wonder what properties $\hat{F}_n$ has as an estimator of F. Observe that for a fixed x,

$$\begin{aligned} \mathbb{E}(n\hat{F}_n(x)) &= np = nF(x) \\ \text{Var}(n\hat{F}_n(x)) &= np(1-p) = nF(x)(1 - F(x), \end{aligned}$$

since this is the mean and variance for a Binomial RV. Thus, for a fixed x,

$$\begin{aligned} \mathbb{E}(\hat{F}_n(x)) &= \frac{1}{n}\mathbb{E}(n\hat{F}_n(x)) = \frac{1}{n} \cdot nF(x) = F(x) \\ \text{Var}(\hat{F}_n(x)) &= \frac{1}{n^2} \cdot nF(x)(1 - F(x)) = \frac{F(x)(1 - F(x)}{n} \\ \text{MSE}(\hat{F}_n(x)) &= \text{Var}(\hat{F}_n(x)) \text{ (since } \hat{F}_n(x) \text{ is unbiased).} \end{aligned}$$

Let's look at some asymptotic properties. We have pointwise convergence, i.e., $\hat{F}_n(x) \xrightarrow{p} F(x)$. But this is just for one fixed value of x. A stronger statement is about uniform convergence:

Theorem 1.1. Glivenko-Cantelli. For the same set-up above, we have

$$\sup_x |\hat{F}_n(x) - F(x)| \xrightarrow{p} 0.$$

In math, there is pointwise vs. uniform convergence of (non-random) functions. Pointwise and uniform convergence in probability are basically the extensions of those notions to random functions, in this case, $\hat{F}_n(x)$.

1.3 Confidence Intervals

Pointwise CIs. Recall that in Bernoulli/Binomial distributions, the sample proportion is asympotically normal:

$$\hat{p} \implies N(p, \sqrt{\hat{p}(1-\hat{p})/n}) \text{ which implies}$$
$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \text{ is an approx. } (1-\alpha)100\% \text{ confidence interval.}$$

In this case,

$$\hat{p} = \hat{F}_n(x) = \frac{\sum_{i=1}^n Y_i}{n} \text{ which implies}$$
$$\hat{F}_n(x) \pm z_{\alpha/2}\sqrt{\frac{\hat{F}_n(x)(1-\hat{F}_n(x))}{n}} \text{ is an approx. } (1-\alpha)100\% \text{ CI.}$$

Simultaneous CIs. Note though that this is just for a fixed x! If we form confidence intervals in this way for each unique x in our dataset, then the *total* confidence may be less than $1-\alpha$. We can adjust for this by forming *simultaneous* CIs: for each x, the CI is $(L(x), U(x))$, where

$$L(x) = \max\{\hat{F}_n(x) - \epsilon_n, 0\}$$
$$U(x) = \min\{\hat{F}_n(x) + \epsilon, 1\},$$

where $\epsilon = \sqrt{\log(2/\alpha)/2n}$. Each simultaneous CI will be wider than the corresponding pointwise CI.

1.4 Theoretical Properties of Plug-in Estimators

So we know that $\hat{F}_n$ is a good estimator of F for large enough sample sizes. But what about $\hat{\theta} = T(\hat{F}_n)$ for $T(F)$?

We need some notion that if $\hat{F}_n$ is sufficiently close to F, then also $T(\hat{F}_n)$ is close to $T(F)$. This is basically a notion of continuity, extended to the random function $\hat{F}_n$ and the (function of a) random function $T(\hat{F}_n)$.

How do we define “close” for a CDF? We use the sup-norm

$$||G - F||_\infty = \sup_x (G(x) - F(x)),$$

which is the largest distance between two CDFs G and F over their domain.

Theorem 1.2. Suppose $T(F)$ is continuous in the sup-norm. Then

$$T(\hat{F}_n) \xrightarrow{p} T(F).$$

2 Bootstrapping (Module 4)

In general in math, it can be difficult* to integrate functions. There are many different methods for numerical integration.

*By "difficult", I also mean that there may not even be a closed-form solution for the integral. Here's an interesting fact about that: Far more functions are integrable than differentiable, but closed-form derivatives are more common than closed-form integrals.

In statistics, we clearly often work with integrals. They could be difficult* to integrate, depending on the function. For example, if the parameter of interest is something like $\text{Var}\{\text{median}(X_1, \ldots, X_n)\}$, then it will be very tricky get with integration. We could instead numerically approximate the integral, but we have to use an algorithm that accounts for the randomness of the RVs. One such algorithm is called Monte Carlo (MC) Integration. Details are provided in the lecture notes.

But in nonparametric estimation, we do not know f or F, so we can't use MC integration. But we do have data drawn from F, and previously, we were using it to calculate $\hat{F}_n$ as an estimator of F. What if we performed MC integration on $\hat{F}_n$ instead of F? This is called bootstrapping, and it turns out to work well (under some assumptions).

So the bootstrap involves two types of approximation. Consider the example with $T_n = \text{median}(X_1, \ldots, X_n)$.

(a) **ECDF** (depends on n): We are approximating $\text{Var}_F(T)$ with $\text{Var}_{\hat{F}}(T)$.

(b) **MC integration** (depends on B): We are approximating $\text{Var}_{\hat{F}}(T)$ with $\hat{\text{Var}}_{\hat{F}}(T)$.

Bootstrap Algorithm

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$.

(a) For $b = 1, 2, \ldots, B$,

  (i) Draw $X^*_{1,b}, X^*_{2,b}, \ldots, X^*_{n,b} \overset{\text{iid}}{\sim} \hat{F}_n$ (i.e., sample n observations *with replacement* from the original data)

  (ii) Compute $T^*_{n,b} = g(X^*_{1,b}, X^*_{2,b}, \ldots, X^*_{n,b})$.

(b) We now have the B bootstrap samples of T_n,

$$T^*_{n,1}, \ldots, T^*_{n,B},$$

an iid sample from the sampling distribution for T_n under $\hat{F}_n$. Use this to approximate $\mathrm{Var}_{\hat{F}_n}(T_n)$ by MC integration. That is,

$$\mathrm{Var}_{\hat{F}_n}(T_n) \approx \frac{1}{B}\sum_{j=1}^{B}\left(T^*_{n,j} - \frac{1}{B}\sum_{k=1}^{B} T^*_{n,k}\right)^2 = v_{boot},$$

where v_{boot} is our bootstrap estimate of $\mathrm{Var}(T_n)$.

2.1 Confidence Intervals

Confidence intervals can be constructed from bootstrap samples. Here are three common types of $(1-\alpha)100\%$ bootstrap CIs:

(a) **Method 1: Normal-based Intervals:**

$$C_n = T(\hat{F}_n) \pm z_{\alpha/2}\hat{se}_{boot},$$

where $\hat{se}_{boot} = \sqrt{v_{boot}}$.

(b) **Method 2: Pivotal Intervals.** Let t^*_p be the p-th quantile of the bootstrap samples $(\hat{\theta}^*_{n,1}, \ldots, \hat{\theta}^*_{n,B})$. Then

$$C_n = (2\hat{\theta}_n - t^*_{1-\alpha/2}, 2\hat{\theta}_n - t^*_{\alpha/2}).$$

(c) **Method 3: Percentile Intervals.** Again, let t^*_p be the p-th quantile of the bootstrap samples $(\hat{\theta}^*_{n,1}, \ldots, \hat{\theta}^*_{n,B})$. Then

$$C_n = (t^*_{\alpha/2}, t^*_{1-\alpha/2}).$$

Here is the basic reasoning behind each of these CIs (see lecture notes for details):

(a) We have often seen confidence intervals based on the normal approximation. For example, if $T_{\hat{F}_n}(\hat{\theta}_n) \implies N(T(F), se^2)$, then hypothetically we could form an approximate $(1-\alpha)$ CI as

$$T(\hat{F}_n) \pm z_{\alpha/2}\hat{se},$$

where $\hat{se}$ was just the plug-in estimator for se. But we don't have a general formula for a non-parametric estimate of se, so we can instead use the bootstrap estimate of se.

(b) Let $R_n = \hat{\theta}_n - \theta$. It can be shown that $C_n = (\hat{\theta}_n - r_{1-\alpha/2}, \hat{\theta}_n - r_{\alpha/2})$ is a $(1-\alpha)$ CI, where $r_{\alpha/2}$ and $r_{1-\alpha/2}$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of R_n, respectively. Then this CI can be further simplified into the one with the quantiles $t^*_{\alpha/2}$ and $t^*_{1-\alpha/2}$.

(c) This one seems pretty intuitive. But notice how the quantiles are flipped, compared to the pivotal intervals! So why does this work? Actually, for this CI to be valid, we need to make the assumption that there exists a monotonic transformation m, so that $m(\hat{\theta}_n) - m(\theta)$ is pivotal.

3 When does the bootstrap work?

The power of the bootstrap is that we can use it to compute statistics for complicated parameters. It's quite a flexible method in that way. But it doesn't always work.

First of all, let me emphasize that the version presented here only works for iid data! There are extensions of the bootstrap for dependent data (such as the moving block bootstrap for time series data), but the algorithms are a bit different.

Secondly, it is fundamental to this method that the distribution of

$$\hat{\theta}^*_n - \hat{\theta}_n = T(\hat{F}^*_n) - T(\hat{F}_n)$$

is close to the distribution that we were originally interested in for nonparametric estimation:

$$\hat{\theta}_n - \theta = T(\hat{F}_n) - T(F),$$

where $\hat{F}^*_n$ is the empirical CDF of the bootstrap sample.

4 Problems

Problems 1 and 2 from Homework 4 – please see the homework folder on bCourses for the solutions.