Non-Parametric Inference, Chapter 7

Elizabeth Purdom

This document has last been compiled on Sep 20, 2024.

Contents

Our basic overview in the previous section was general, but our examples all relied heavily on parametric models for our data. We will come back to parametric models through out much of this course, but we are first going to examine how one can do estimation and confidence intervals with non-parametric models; these are often called non-parametric estimates/inference.

Overview The most common approach for creating non-parametric estimates is to directly estimate the unknown distribution F from the data. This is true, even when your goal is to estimate only one aspect or parameter of the distribution. This approach will be the topic of chapters 7 and 8. Chapter 7 we will focus on how to make such estimates and their properties, and chapter 8 we will focus on using the bootstrap to use inference.

We will see later in hypothesis testing another non-parametric approach for certain types of hypothesis testing – permutation methods.

1 Point-estimation: The Plug–in (or Substitution) Principle

1.1 Statistical Functionals

Let $X_1, ..., X_n \overset{iid}{\sim} F$, where F can be parametric or nonparametric.

We will want to estimate quantities that are related to F, such as the mean, median, variance, quantiles, etc, by a nonparametric way. In other words, we can write the quantities of interest as a function of F, $T(F)$.

We call $T(F)$ a **Statistical Functional**. Note that $T(F)$ also defines what we would call a parameter (θ), so we might say we want to estimate $\theta = T(F)$.

To make this clear, here are some simple examples:

- the expected value of X_i is

$$T(F) = E_F(X_i) = \int x dF(x)$$

- the variance of X_i is

$$T(F) = var_F(X_i) = \int x^2 dF(x) - \left(\int x dF(x) \right)^2$$

- the variance of $\bar{X} = \frac{1}{n}\sum_i X_i$ is

$$T(F) = \frac{1}{n} var_F(X_i)$$

- The p^{th} quantile

$$F^{-1}(p) = \inf\{x : F(x) \geq p\}$$

Notice that writing our quantity of interest as a statistical functional does not depend on whether the true distribution F is parametric or non-parametric – it is still true.

1.2 The Empirical CDF

A natural estimate, then, of $\theta = T(F)$ is to have an estimate of F from the data, say $\hat{F}$ and "plug it in" to the equation,

$$\hat{\theta} = T(\hat{F})$$

A natural estimate of F is the **empirical CDF**.

Definition 1.1. Empirical CDF Let $X_1, \ldots, X_n \overset{iid}{\sim} F$. The **empirical CDF** $\hat{F}_n$ puts mass $1/n$ at each datapoint.

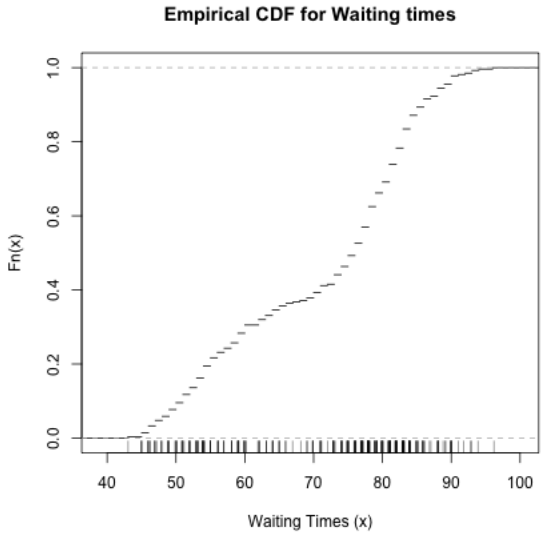
$$\begin{aligned} \hat{F}_n(x) &= \frac{\sum_{i=1}^n I(X_i \leq x)}{n} \\ &= \#\{X_i \leq x\}/n \end{aligned}$$

where $I(\cdot)$ is the indicator function,

$$I(E) = \begin{cases} 1 & E \text{ is true} \\ 0 & E \text{ is false} \end{cases}$$

We can visualize $\hat{F}_n$ for some data given in the book (p.105), which gives the eruption times and waiting times between eruptions of the Old Faithful geyser in Yellowstone.

```
##   eruptions waiting
## 1     3.600      79
## 2     1.800      54
## 3     3.333      74
## 4     2.283      62
## 5     4.533      85
## 6     2.883      55
```



Important Properties of $\hat{F}_n$

- $\hat{F}_n$ is random because it is a function of $X_1, \ldots, X_n$.
- Conditional on the data, $\hat{F}_n$ is a regular CDF. For example, we could calculate probabilities under $\hat{F}_n$ or sample from the distribution defined by $\hat{F}_n$, just like any distribution; we can write for a r.v. $W \sim \hat{F}_n$.
- Instead of using W, we will often write X^* for a r.v distributed according to $\hat{F}_n$.
- We will write probability and expectation statements using $\hat{F}_n$ just like other CDFs,

  $$P_{\hat{F}_n}(X^* \leq x)$$
  $$E_{\hat{F}_n}(X^*).$$

  The difference is that implicit in these statements is that we are conditioning on the data, i.e. this notation really means,

  $$P(X^* \leq x | X^* \sim \hat{F}_n, X_1, \ldots, X_n)$$
  $$E(X^* | X^* \sim \hat{F}_n, X_1, \ldots, X_n)$$

  In particular, these quantites are random, just like $\hat{F}_n$.
- We can also consider the *unconditional* distribution of X^*, and thus write statements unconditionally, like $E(X^*)$ or $P(X^*)$. This is the distribution of X^* considering both the randomness of sampling from $\hat{F}_n$ *and* the sampling

of $X_1 \ldots, X_n$ from F. Notice that the unconditional distribution is what we want to consider if we are considering how X^* varies over multiple draws of $X_1, \ldots, X_n$, and not only for the specific realization of data that we observe.

Note that by the tower rule or Law of total expectation, we have

$$E(X^*) = E(E(X^*|X_1, \ldots, X_n)).$$

Notice that conditional statements about X^* will be functions of the data $X_1, \ldots, X_n$, while unconditional statements about X^* will depend only on the unknown F.

- Conditional on $X_1, \ldots, X_n$, we can describe a process to generate X^* so that the resulting $X^* \sim \hat{F}_n$ by sampling from our X_i:

  Generate X^* by drawing a value from our $X_1, \ldots, X_n$, with each value being selected with equal probability. Then $X^* \sim \hat{F}_n$.

  This means that $P_{\hat{F}_n}(X^* = t) = \frac{\#X_i = t}{n} = \frac{1}{n}\sum_{i=1}^n I(X_i = t)$. Note this definition accounts for potential ties.

- The distribution defined by $\hat{F}_n$ is a discrete distribution – it takes on the *unique* values of $X_1, \ldots, X_n$.

- $\hat{F}_n$ is not F! This seems obvious, but it's important to keep the distinction. So that

  $$P_{\hat{F}_n}(X^* \leq x) = \hat{F}_n(x)$$

  is different from

  $$P_F(X \leq x) = F(x).$$

- If for a fixed x we define

  $$Y_i(x) = I(X_i \leq x), i = 1, \ldots, n$$

  then

  $$\hat{F}_n(x) = \frac{\sum_{i=1}^n Y_i(x)}{n}$$

  $Y_i(x)$ are each *iid* Bernoulli r.v.'s, with

  $$p = P(Y_i(x) = 1) = P(X_i \leq x) = F(x)$$

  So for at a fixed point x (i.e. not random),

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

  Is this a conditional or unconditional statement?

1.3 The plug-in estimator

Definition 1.2. Plug-in Estimator The plug-in estimator of $\theta = T(F)$ is given by

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

where $\hat{F}_n$ is the empirical CDF of the data.

Examples of Plug–in Estimators

- $\theta = T(F) = \mathrm{E}_F(X)$. Let $\mathcal{X}$ be all the unique values of $X_1, \ldots, X_n$. Then the plug–in estimator will be (from first principles of definition of expection for a discrete distribution):

$$\begin{aligned} \hat{\theta} = T(\hat{F}_n) &= \mathrm{E}_{\hat{F}_n}(X^*) \\ &= \textstyle\sum_{t\in\mathcal{X}} t P_{\hat{F}_n}(X^* = t) \\ &= \textstyle\sum_t t \cdot \frac{\sum_{i=1}^n I(X_i = t)}{n} \end{aligned} \tag{1}$$

  Notice that if there are unique X_i, this will further simplify to

$$\begin{aligned} &= \textstyle\sum_{i=1}^n X_i \frac{1}{n} \\ &= \bar{X}_n \end{aligned} \tag{2}$$

  But in fact, this is the answer regardless of ties because

$$\frac{1}{n}\sum_t t \cdot \{\#X_i = t\} = \frac{1}{n}\sum_{i=1}^n X_i$$

  But we can see this more formally by switching the order of the summations

$$\begin{aligned} &= \tfrac{1}{n}\textstyle\sum_{t\in\mathcal{X}} t \cdot \sum_{i=1}^n I(X_i = t) \\ &= \tfrac{1}{n}\textstyle\sum_{i=1}^n \sum_{t\in\mathcal{X}} t I(X_i = t) \\ &= \tfrac{1}{n}\textstyle\sum_{i=1}^n \left\{ \sum_{t=X_i} t \cdot 1 + \sum_{t\neq X_i} t \cdot 0 \right\} \\ &= \tfrac{1}{n}\textstyle\sum_i X_i \end{aligned} \tag{3}$$

- More generally, using the same logic, if we want to estimate $\theta = T(F) = E_{\hat{F}_n}(r(X^*))$, we have

$$E_{\hat{F}_n}(r(X^*)) = \sum_{t\in\mathcal{X}} r(t) P_{\hat{F}_n}(X^* = t) = \frac{1}{n}\sum_{i=1}^n r(X_i),$$

  i.e. not worry about the ties.

- $\theta = T(F) = var_F(X)$. Then the plug–in estimate will be

$$\begin{aligned} \hat{\theta} = T(\hat{F}_n) &= var_{\hat{F}_n}(X^*) = \mathrm{E}_{\hat{F}_n}(X^{*2}) - (\mathrm{E}_{\hat{F}_n}(X^*))^2 \\ &= \frac{1}{n}\sum_{i=1}^n X_i^2 - (\bar{X}_n)^2 \\ &= \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 \end{aligned} \tag{4}$$

- $\theta = T(F) = median(X) = \inf_t\{t|F(t) \geq \frac{1}{2}\}$.
  Then the plug–in estimate for the median is

$$\hat{\theta} = T(\hat{F}_n) = \inf_t\{t|\hat{F}_n(t) \geq \frac{1}{2}\} = median(X_1, \ldots, X_n)$$

Note: In general, how plug-in estimator works depending on the properties of $\hat{F}_n$ and also the property of θ function.

Is everything a Plug-in Estimator? The plug-in estimators result in natural estimators, which makes it seem like everything must be.

Consider however, how we normally estimate the variance, $\theta = T(F) = var_F(X)$. The standard estimate is given as:

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

while the plug-in estimator is

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

Similarly, the trimmed mean is a robust estimate of the mean, where you take the mean of the observations after discarding the top and bottom 5% of the data. It is not a plug-in estimator of the $\theta = T(F) = E_F(X)$

2 Theoretical properties of the Empirical CDF

How is $\hat{F}_n$ as an estimate of F? We can look at a number of properties of $\hat{F}_n$.

Finite Properties For any fixed x, we have already seen that $n\hat{F}_n \sim Binomial(n, F(x))$. This gives us the following statements we can make about $\hat{F}_n$:

$$\begin{aligned} E[\hat{F}_n(x)] &= F(x) \\ V[\hat{F}_n(x)] &= \frac{F(x)[1-F(x)]}{n} \\ MSE[\hat{F}_n(x)] &= V[\hat{F}_n(x)] \end{aligned}$$

Notice that these are *not* statements conditional on the data X, but taking the expected value over the data.

Asymptotics We can see that the MSE $\to 0$ as $n \to 0$.

We can also look at convergence in probability:

$$\hat{F}_n(x) \xrightarrow{P} F(x)$$

This last result is what we call **pointwise** convergence, meaning it is a statement about a *fixed* x. Indeed all of the above statements are regarding a fixed value x.

We can be even stronger than this last statement with the Glivenko-Cantelli Theorem:[1]

Theorem 1 (Glivenko-Cantelli)**.** *Let* $X_1, \ldots X_n \overset{iid}{\sim} F$*. Then*

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

This is even stronger because it is *uniform convergence*, which intuitively we can think of as being able to find a *rate* of convergence that works for all x – there is a single n will be large enough to ensure that $\hat{F}_n(x)$ will be within a specified ϵ of $F(x)$ regardless of x.

3 Confidence Intervals for F

So far we have just focused on $\hat{F}_n$ as a estimator of F, but we can also create confidence intervals as well.

[1] Actually, the Glivenko-Cantelli theorem gives almost sure convergence, which is even stronger than convergence in probability

Specifically, for a particular point x, we can create a confidence interval for $\hat{F}_n(x)$. There are two ways we can think about confidence intervals: pointwise and simultaneous.

3.1 Pointwise CIs

We have already seen that for a fixed value of x,

$$\hat{F}_n(x) = \frac{\sum_{i=1}^n Y_i(x)}{n}$$

and the $Y_i(x)$ are each *iid* Bernoulli r.v.'s, with

$$p = P(Y_i(x) = 1) = P(X_i \leq x) = F(x)$$

This means that $\sum_{i=1}^n Y_i(x) \sim Binomial(n, F(x))$.

We could create binomial confidence interval for $F(x)$ *for a particular* x based on based on our normal approximation to the binomial:

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

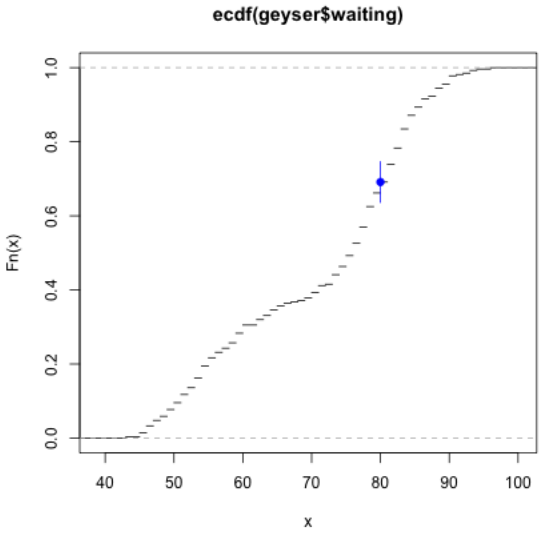
What is $\hat{p}$ in this case? The standard estimate of $p = F(x)$ for a binomial is

$$\hat{p}_n = \hat{F}_n(x) = \frac{\sum_{i=1}^n Y_i(x)}{n}$$

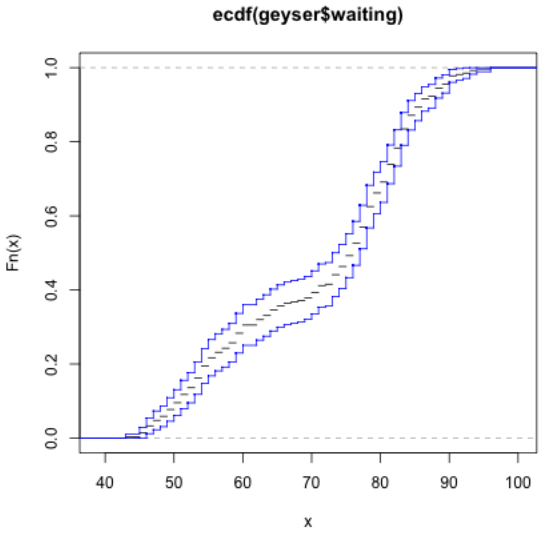
so we have

$$\hat{F}_n(x) \pm 1.96\sqrt{\frac{\hat{F}_n(x)(1-\hat{F}_n(x))}{n}}$$

which could look something like this for $x = 80$:



We could imagine doing this for all x,



However, these are a bunch of CI, each considered pointwise, meaning this would only guarantee your confidence intervals have 95% coverage at a *particular* x; it does not bound the probability that across all of the x values you do not cover the true $F(x)$. So if you use such a pointwise method over and over again for all x, then the error could add up across x to make the chance of one those CI not covering $F(x)$ be much more than 0.05.

3.2 Simultaneous CIs

We could also ask for CI bands around our entire function, so that the probability that the entire CI band covers the true F is 0.95.

To create such simultaneous CI, we need to use the following theorem that improves upon the Glivenko-Cantelli Theorem by giving specific, uniform bounds for how likely it is *for a particular value* n to see values of $\hat{F}_n(x)$ that are far from $F(x)$:

Theorem 2 (Dvoretzky-Kiefer-Wolfowitz Inequality:)**.** *Let* $X_1, \ldots, X_n \overset{iid}{\sim} F$*. For any* $\epsilon > 0$*,*

$$P\left(\sup_x |F(x) - \hat{F}_n(x)| > \epsilon\right) \leq 2e^{-2n\epsilon^2}$$

Notice that Glivenko-Cantelli Theorem tells us that for every ϵ

$$\lim_{n\to\infty} P\left(\sup_x |F(x) - \hat{F}_n(x)| > \epsilon\right) = 0$$

The Dvoretzky-Kiefer-Wolfowitz Inequality gives precise bounds for how big that probability can be for any particular n.

How does this give us confidence intervals? We want a statement of the following form:

$$P\left(L(x) \leq F(x) \leq U(x) \text{ for all } x\right) \geq 1 - \alpha$$

If we set

$$\epsilon = \epsilon_n = \sqrt{\log(2/\alpha)/(2n)}$$

then the right hand side of the inequality is equal to α, i.e.

$$P\left(\sup_x |F(x) - \hat{F}_n(x)| > \sqrt{\log(2/\alpha)/(2n)}\right) \leq \alpha$$

This means that

$$P\left(\sup_x |F(x) - \hat{F}_n(x)| < \sqrt{\log(2/\alpha)/(2n)}\right) \geq 1 - \alpha$$

because this is just the complement. So we have

$$P\left(\hat{F}_n(x) - \sqrt{\log(2/\alpha)/(2n)} < F(x) < \hat{F}_n(x) + \sqrt{\log(2/\alpha)/(2n)} \text{ for all } x\right) \geq 1 - \alpha$$

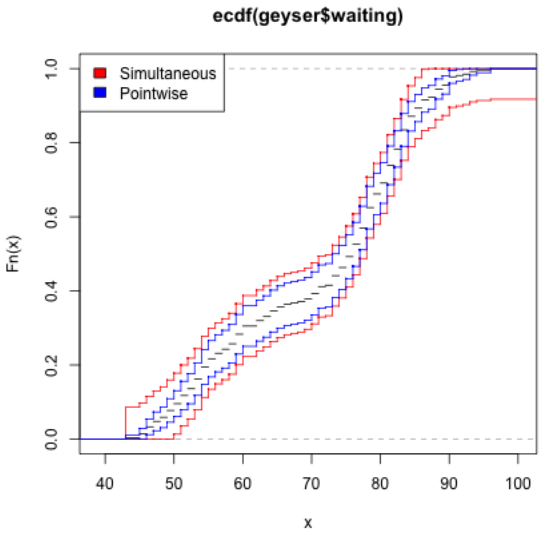
Notice that this a global probability statement over all x – i.e. the probability that the confidence band doesn't cover $F(x)$ anywhere. Thus this is a global or **simultaneous** $1 - \alpha$ confidence band for F.

This gives us simultaneous confidence intervals for all x with lower and upper bounds

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

(notice we can cap $L(x)$ and $U(x)$ with the values 0 and 1, respectively, since $F(x)$ is between $0, 1$!)

We can compare the simultaneous CI with the previous pointwise CI.



4 Theoretical Properties of Plug–in estimators

What about plug-in estimators $\hat{\theta} = T(\hat{F}_n)$ – what can we say about them? Specifically, the previous results tell us that $\hat{F}_n(x)$ is a good estimate of F for large enough sample size. But what about $T(\hat{F}_n)$ as an estimate of $T(F)$?

To be able to say that, we have to put some limits on our functional T – we need to know that if two distributions F and G are "close" to each other, then $T(F)$ and $T(G)$ are close to each other.

It's a bit trickier with CDFs, since we need a definition of distance between functions. The one that works is the sup-norm, basically the largest distance between the functions over all the domain of the function,

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

Then our result we can state is that,

Theorem 3. *Suppose the function $T(F)$ is continuous in the sup-norm:*

$$\forall \epsilon > 0, \exists \delta > 0 \text{ such that } ||G - F||_\infty < \delta \text{ implies } |T(G) - T(F)| < \epsilon.$$

Then,

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

The condition that starts our theorem is our restriction on T: if F and G are two distributions that are close to each other, then $T(G)$ is also close to $T(F)$. If that is the case, the theorem tells us that we can "transfer" the convergence of $\hat{F}_n(x)$ to F into convergence of $T(\hat{F}_n)$ to $T(F)$.

Example of a "badly behaved" functional Let's consider an example of T that is not well-behaved. Let $T_{max}(F)$ be the maximum, i.e. the largest possible value. For example, if $F = Uniform(0,1)$ then $T_{max}(F) = 1$. If we have a sample of data $X_1, \ldots, X_n$, then the support of $\hat{F}_n$ is the maximum value you can take on $\hat{F}_n$, so we have

$$T_{max}(\hat{F}_n) = max(X_1, \ldots, X_n)$$

This is an example of T that is not well-behaved. More precisely, you can find examples of two distributions F_n, G_n where the cdfs F_n and G_n can be arbitrarily close to each other, but the max is not. You can do this by for example setting $F_n = Uniform(0,1)$ and letting G_n be $Uniform(0,1)$, except for a a small point mass on a value, say 2. If we let the probability of that point mass be $1/n$, then we can make F_n and G_n very close to each other by setting a very large n, but $T_{max}(F_n) = 1$ and $T_{max}(G_n) = 2$ – so $T_{max}(F_n) - T_{max}(G_n) = 1$ and their difference is not going to zero, and never will.

This seems like a really artificial choice of F and G as a counter example, it demonstrate how T_{max} is not a "smooth" function – it's like having a discontinutity in a function. We will see when we look at the bootstrap, that it's not just an artificial example, but that indeed we do not get good results in trying to estimate $T_{max}(F)$.

4.1 Non-parametric Estimator

Let's step back and remember what this module of the class is about – non-parametric inference. We defined this as estimating a quantity of interest (θ) from an infinite class of models $\mathcal{F}$.

How is what we are doing non-paramtric estimation?

Notice that both in the construction of the plug-in estimator and the statement of its theoretical properties, we've made very little assumptions about the true distribution F. We have not done is made any assumptions about the *form* of F, and certainly no assumption that it can be written in terms of a finite number of parameters. So these are non-parametric methods.

That doesn't mean we have made no assumptions about the data. The big assumption we have made is that our data $X_1, \ldots, X_n$ are $i.i.d.$[2] Now that a fundamental assumption, and all we have said above can fall apart if it's not true. Nor do we have any effective generally way to check if it's true.[3] And data may not be $i.i.d.$ There are many ways in which data may be dependent or have slightly different distributions so they are not identically distributed. We might collect data over time or space with complicated dependencies between different responses. We might be unknowing collecting related observations (people from the same families or communities such that their responses influence each other) that induce dependencies or create groups of slightly differently distributed data.

Common Estimators Note also, that we get pretty standard estimators of things like the mean and variance. We'll see that we can get these same estimators from more constrained parametric models too. This is a common feature of these kinds of simple quantities – by taking different routes we wind up on these same estimators. Knowing that we can justify the estimator on non-parametric grounds, however, justifies the estimator as more *robust*, meaning it is not sensitive to the assumptions used to develop it.

[2]We also make assumptions about the functional T which we want to estimate, but that's separate from assumptions about F

[3]We can check for certain kinds of dependency if we know they might exist, for example temporal or spatial dependency. But we don't have a generic way to determine if an arbitrary set of data $X_1, \ldots, X_n$ are generated independently from each other.