Introduction; The empirical distribution function

Patrick Breheny

August 23

Nonparametric vs. parametric statistics

- The main idea of nonparametric statistics is to make inferences about unknown quantities without resorting to simple parametric reductions of the problem
- For example, suppose $X \sim F$, and we wish to estimate, say $\mathbb{E}(X)$ or $\mathbb{P}(X > 1)$
- The approach taken by parametric statistics is to assume that F belongs to a family of distribution functions that can be described by a small number of parameters – *e.g.*, the normal distribution:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

- These parameters are then estimated, and we make inferences about the quantities we were originally interested in ($\mathbb{E}(X)$ or $\mathbb{P}(X > 1)$) based on assuming $X \sim N(\hat{\mu}, \hat{\sigma}^2)$

Parametric statistics (cont'd)

- Or suppose we wish to know how $\mathbb{E}(Y)$ changes with x
- Again, the parametric approach is to assume that

$$\mathbb{E}(Y|x) = \alpha + \beta x$$

- We estimate α and β, then base all future inference on those estimates

Shortcomings of the parametric approach

- Both of the aforementioned parametric approach rely on a tremendous reduction of the original problem
- They assume that all uncertainty regarding $F(x)$, or $\mathbb{E}(Y|x)$, can be reduced to just two unknown numbers
- If these assumptions are true, then of course, there is nothing wrong with making them
- If they are false, however:
  - The resulting statistical inference will be questionable
  - We might miss interesting patterns in the data

The nonparametric approach

- In contrast, nonparametric statistics tries to make as few assumptions as possible about the data
- Instead of assuming that $F(x)$ is normal, we will allow $F(x)$ to be any function (provided, of course, that it satisfies the definition of a cdf)
- Instead of assuming that $\mathbb{E}(y)$ is linear in x, we will allow it to be any continuous function
- Obviously, this requires the development of a whole new set of tools, as instead of estimating parameters, we will be estimating functions (which are much more complex)

The four main topics

We will go over four main areas of nonparametric statistics in this course:

- Estimating aspects of the distribution of a random variable
- Testing aspects of the distribution of a random variable
- Estimating the density of a random variable
- Estimating the regression function $\mathbb{E}(Y|x) = f(x)$

The empirical distribution function

- We will begin with the problem of estimating a CDF (cumulative distribution function)
- Suppose $X \sim F$, where $F(x) = \mathbb{P}(X \leq x)$ is a distribution function
- The *empirical distribution function*, $\hat{F}$, is the CDF that puts mass $1/n$ at each data point x_i:

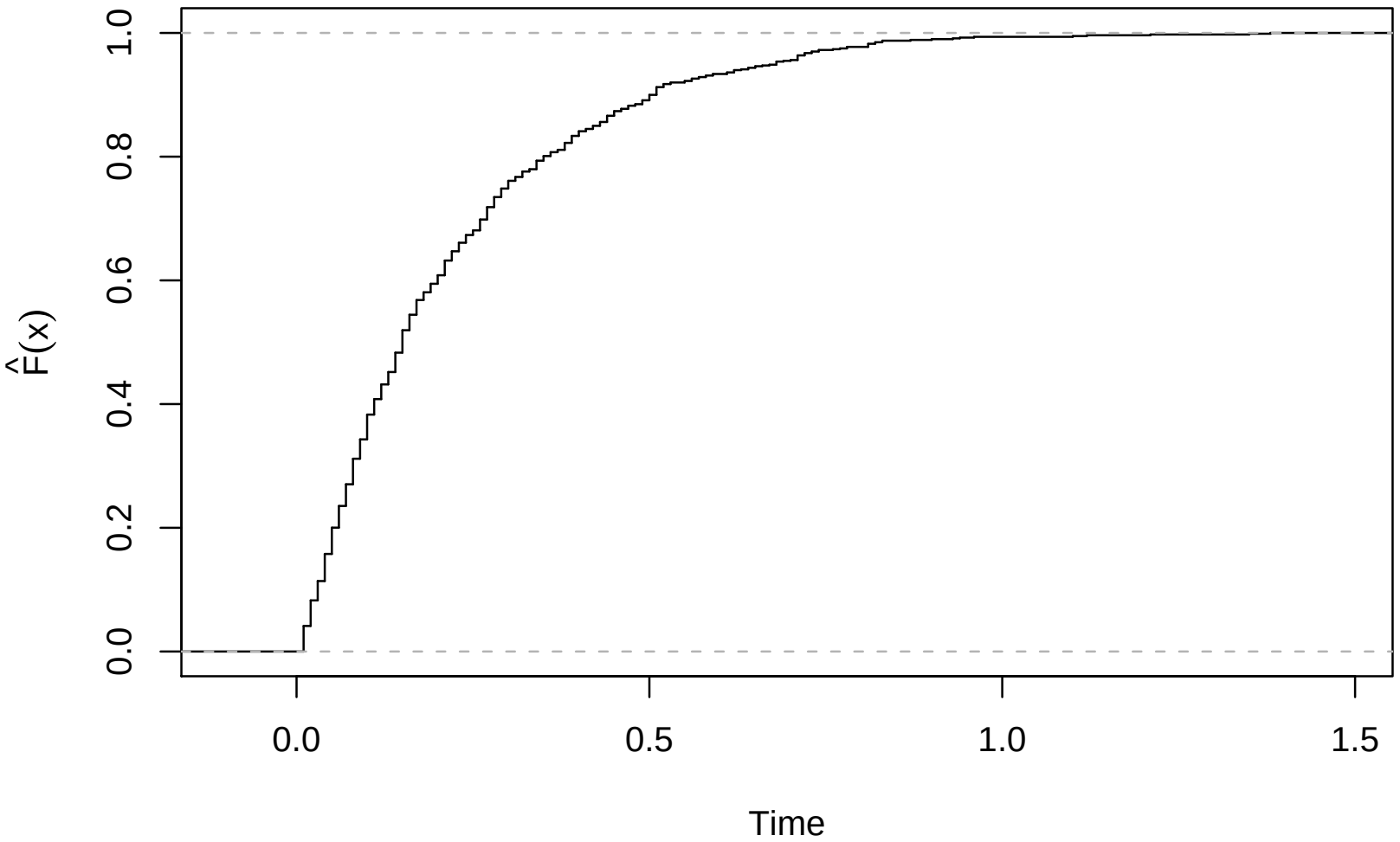
$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x)$$

where I is the indicator function

The empirical distribution function in R

R provides the very useful function ecdf for working with the empirical distribution function

```
Data <- read.delim("nerve-pulse.txt")
Fhat <- ecdf(Data$time)
> Fhat(0.1)
[1] 0.3829787
> Fhat(0.6)
[1] 0.933667
plot(Fhat)
```


$\hat{F}(x)$
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
Time

Properties of $\hat{F}$

- At any fixed value of x,

$$\mathbb{E}\{\hat{F}(x)\} = F(x)$$
$$\mathbb{V}\{\hat{F}(x)\} = \frac{1}{n}F(x)(1 - F(x))$$

- Note that these two facts imply that

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

for any given x

The Glivenko-Cantelli Theorem

- An even stronger proof of convergence is given by the *Glivenko-Cantelli Theorem*
- **Glivenko-Cantelli Theorem**: Suppose $X_1, X_2, \ldots$ are i.i.d. random variables with cdf F. Then

$$\sup_x \left| \hat{F}(x) - F(x) \right| \xrightarrow{\text{a.s.}} 0$$

- This theorem has been called the "fundamental theorem of (nonparametric) statistics"

$\hat{F}$ as a nonparametric MLE

- The empirical distribution function can be thought of as a nonparametric maximum likelihood estimator
- **Homework:** Show that, out of all possible CDFs, $\hat{F}$ maximizes

$$L(F|\mathbf{x}) = \prod_{i=1}^{n} \mathbb{P}_F(x_i)$$

Confidence intervals vs. confidence bands

- Before moving into the issue of calculating confidence intervals for F, we need to discuss the notion of a confidence interval for a function
- One approach is to fix x and calculate a confidence interval for $F(x)$ – *i.e.*, find a region $C(x)$ such that, for any CDF F,

$$\mathbb{P}\{F(x) \in C(x)\} \geq 1 - \alpha$$

- These intervals are referred to as *pointwise confidence intervals*

Confidence intervals vs. confidence bands (cont'd)

- Clearly, however, if there is a $1-\alpha$ probability that $F(x)$ will not lie in $C(x)$ at each point x, there is greater than a $1-\alpha$ probability that there exists an x such that $F(x)$ will lie outside $C(x)$
- Thus, a different approach to inference is to find a confidence region $C(x)$ such that, for any CDF F,

$$\mathbb{P}\{F(x) \in C(x) \quad \forall x\} \geq 1-\alpha$$

- These intervals are referred to as *confidence bands* or *confidence envelopes*

Inference regarding F

- We can use the fact that, for each value of x, $\hat{F}(x)$ follows a binomial distribution with mean $F(x)$ to construct pointwise intervals for F
- To construct confidence bands, we need a result called the *Dvoretzky-Kiefer-Wolfowitz inequality*, or *DKW inequality*:

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

- Note that this is a finite-sample, not an asymptotic, result

A confidence band for F

Thus, setting the right side of the DKW inequality equal to α,

$$\alpha = 2\exp(-2n\epsilon^2)$$

$$\log\left(\frac{\alpha}{2}\right) = -2n\epsilon^2$$

$$\log\left(\frac{2}{\alpha}\right) = 2n\epsilon^2$$

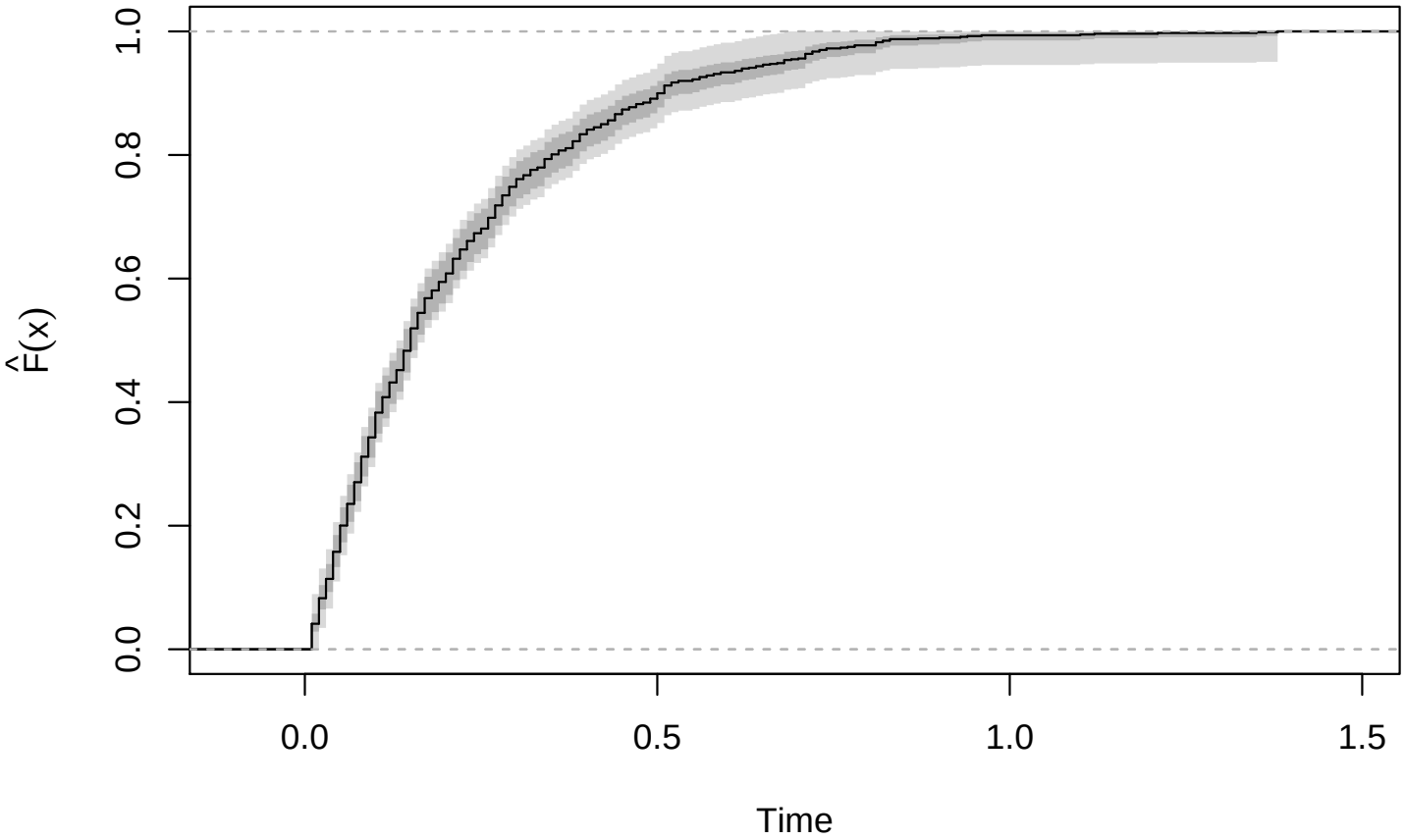
$$\epsilon = \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}$$

A confidence band for F (cont'd)

Thus, the following functions define an upper and lower $1-\alpha$ confidence band for any F and n:

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

Pointwise vs. confidence for nerve pulse data



Homework

- I claimed that the confidence band based on the DKW inequality worked for *any* distribution function . . . does it?
- **Homework:** Generate $X_1, X_2, \ldots, X_{100}$ independent observations and compute a 95 percent global confidence band for the CDF F based on the DKW inequality. Repeat this 1000 times and report the proportion of data sets for which the confidence band contained the true distribution function.
  (a) Carry out the above simulation with $F = N(0, 1)$.
  (b) Repeat using data generated from the standard Cauchy distribution.