Smoothing concepts

Patrick Breheny

October 16

Introduction

- For the rest of the course, we will be interested in estimating curves
- We will be interested in two main types of curves:
  - Density estimation: $f(x)$
  - Regression: $f(x) = \mathbb{E}(y|x)$
- Because we anticipate that the real nature of these functions is some sort of smooth curve, it is desirable that our estimates be smooth as well – this is why another name for curve estimation is *smoothing*
- In this lecture, we will discuss histograms, the simplest method of density estimation, and introduce many of the main ideas that will come up consistently throughout the rest of the course

Histograms

- Let $[a, b]$ denote an interval which contains the data $\{x_i\}$, and let m be an integer which divides $[a, b]$ into m equal-width *bins*, $\{B_j\}_{j=1}^m$
- Let

$$h = \frac{b-a}{m}$$

denote the *binwidth* and let y_j denote the number of observations in B_j

- Finally, let $\hat{p}_j = y_j/n$, $p_j = \int_{B_j} f(u)du$, and

$$\hat{f}(x) = \frac{\hat{p}_j}{h}$$

for all $x \in B_j$, where $f(\cdot)$ is the true density of X

Expectation and variance of $\hat{f}$

Theorem: For a fixed x and m,

$$\mathbb{E}\hat{f}(x) = \frac{p_j}{h}$$

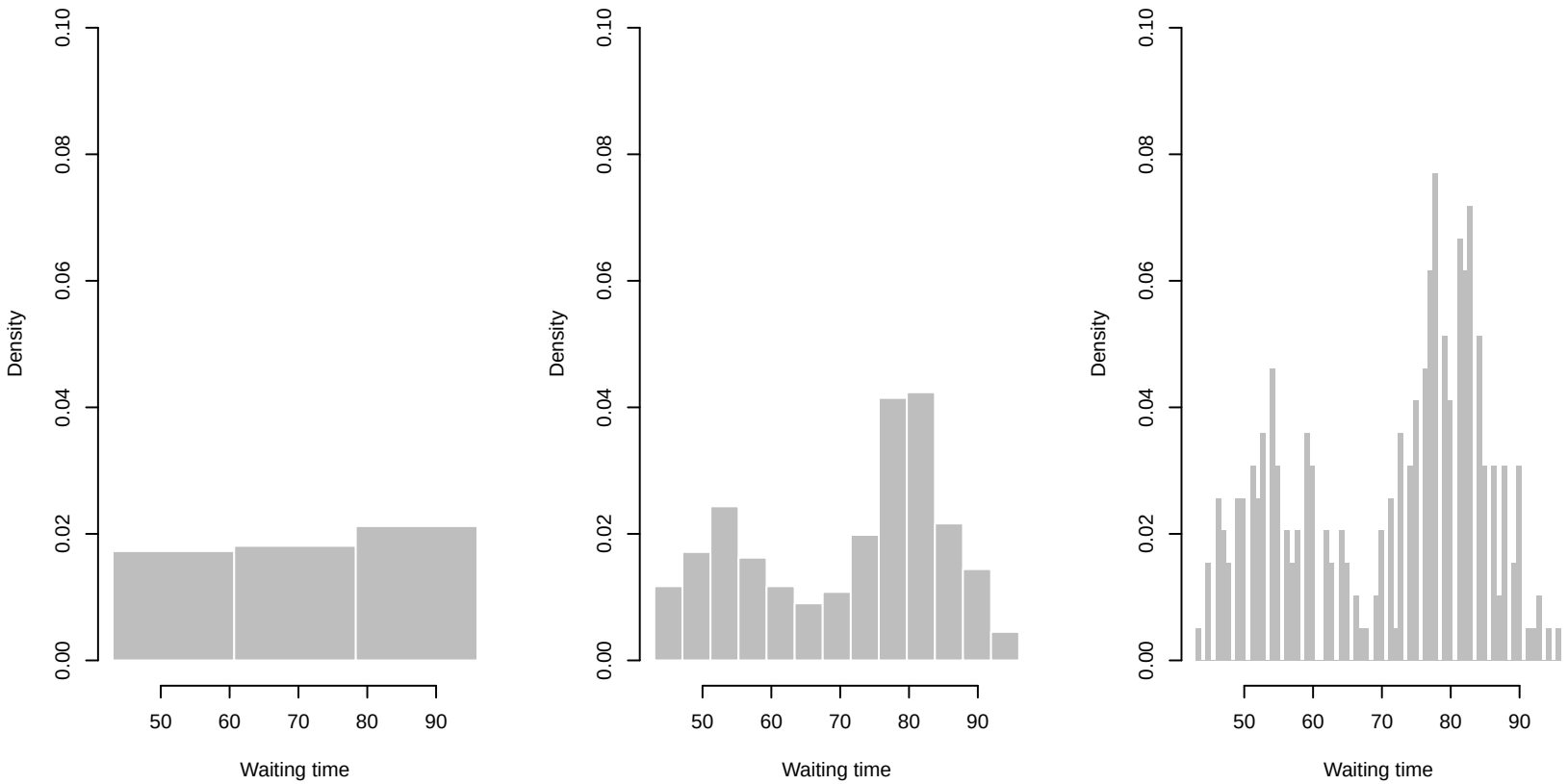
$$\mathbb{V}\hat{f}(x) = \frac{p_j(1-p_j)}{nh^2}$$

Bias

- The histogram is therefore an unbiased estimator of the average density over B_j
- But that isn't the same thing as an unbiased estimator of f (unless f happened to be constant over B_j)
- If f changes over B_j, $\hat{f}$ will be biased

Bias (cont'd)

The bias can be alleviated by choosing a smaller binwidth:



Variance

- However, recall that

$$\mathbb{V}\hat{f}(x) = \frac{p_j(1-p_j)}{nh^2}$$

- If we make the binwidth twice as small, we quadruple the variance of our estimator
- Therefore, we forced into a difficult tradeoff: if we try to reduce bias, we increase variance, and vice versa

The bias-variance tradeoff

- This *bias-variance tradeoff* is fundamental and occurs whenever we try to estimate curves: it will come up in every method we discuss from now until the end of the semester
- In most of these methods, there will be a parameter that controls this tradeoff
- This parameter is called the *smoothing parameter*, because it controls how smooth the curve is
- If the curve is too smooth, it risks bias; if it is very rough, it risks variance
- In the histogram example, the smoothing parameter is the binwidth

Loss functions

- In order to start thinking about an optimal way to balance bias and variance, we need to introduce a criterion that measures the overall quality of a curve
- This criterion is called a *loss function*; the most common loss function is the *squared-error loss function*:

$$L(f, \hat{f}) = \int \{\hat{f}(x) - f(x)\}^2 dx$$

Expected loss

- This leads us to a criterion for choosing the smoothing parameter to optimally balance variance and bias: the expected value of the loss function
- This criterion is known as the *expected loss* or *risk*
- Note that the random variable in this expectation is the estimate $\hat{f}$, which depends implicitly on the data

Expected squared-error loss

- For squared error loss, the bias-variance decomposition is explicit
- **Theorem:** For squared-error loss,

$$\mathbb{E}L(f, \hat{f}) = \int b(x)^2 dx + \int v(x) dx,$$

where $b(x) = \mathbb{E}\hat{f}(x) - f(x)$ is the bias of $\hat{f}(x)$ and $v(x) = \mathbb{V}\hat{f}(x)$ is the variance of $\hat{f}(x)$

- In words, expected loss is equal to (integrated) bias squared plus variance
- **Homework:** Prove the above theorem

Bias-variance decomposition for histograms

Theorem: Suppose f is twice differentiable with bounded support, $\int f'(u)^2 du < \infty$, and L is the squared error loss function. Then

$$\mathbb{E}L(f, \hat{f}) = \frac{h^2}{12} \int f'(u)^2 du + \frac{1}{nh} + o(h^2) + O\left(\frac{1}{n}\right).$$

Furthermore, the value h^* that minimizes the expected loss is

$$h^* = \frac{1}{n^{1/3}} \left(\frac{6}{\int f'(u)^2 du} \right)^{1/3}$$

With this choice of binwidth,

$$\mathbb{E}L(f, \hat{f}) = O(n^{-2/3})$$

Comments

Three important things to note from this theorem:

- Note that the bias is proportional to h^2 (low binwidth is good for bias), while variance is proportional to h^{-1} (low binwidth is bad for variance)
- Bias depends on the non-constancy of f – the more f changes, the greater the impact of bias on expected loss
- The optimal expected loss converges to 0 at the rate $n^{-2/3}$; in our next lecture, we will introduce a type of density estimator with a superior convergence rate

Consistency of the histogram

It is also worth noting that the histogram is pointwise consistent:

Theorem: Suppose that f is continuous at x, that $h \to 0$, and that $nh \to \infty$ as $n \to \infty$. Then $\hat{f}(x) \xrightarrow{\mathsf{P}} f(x)$.

Introduction

- In practice, this theorem is not a useful way of choosing an optimal binwidth, as it depends on knowing f
- Instead, the optimal values of smoothing parameters must usually be estimated based on the observed data
- Writing the loss as a function of the smoothing parameter h, note that

$$L(h) = \int \hat{f}^2(x)dx - 2\int \hat{f}(x)f(x)dx + c,$$

where c is a constant with respect to h

Estimating $\mathbb{E}L(h)$

- We are interested in estimating $\mathbb{E}L(h)$
- An obvious step is to estimate $\mathbb{E}\hat{f}$ with $\hat{f}$
- Thus, we can estimate $\mathbb{E}L(h)$ if we can estimate $\int \hat{f}(x)f(x)dx$
- This is not trivial, however, since we don't know f

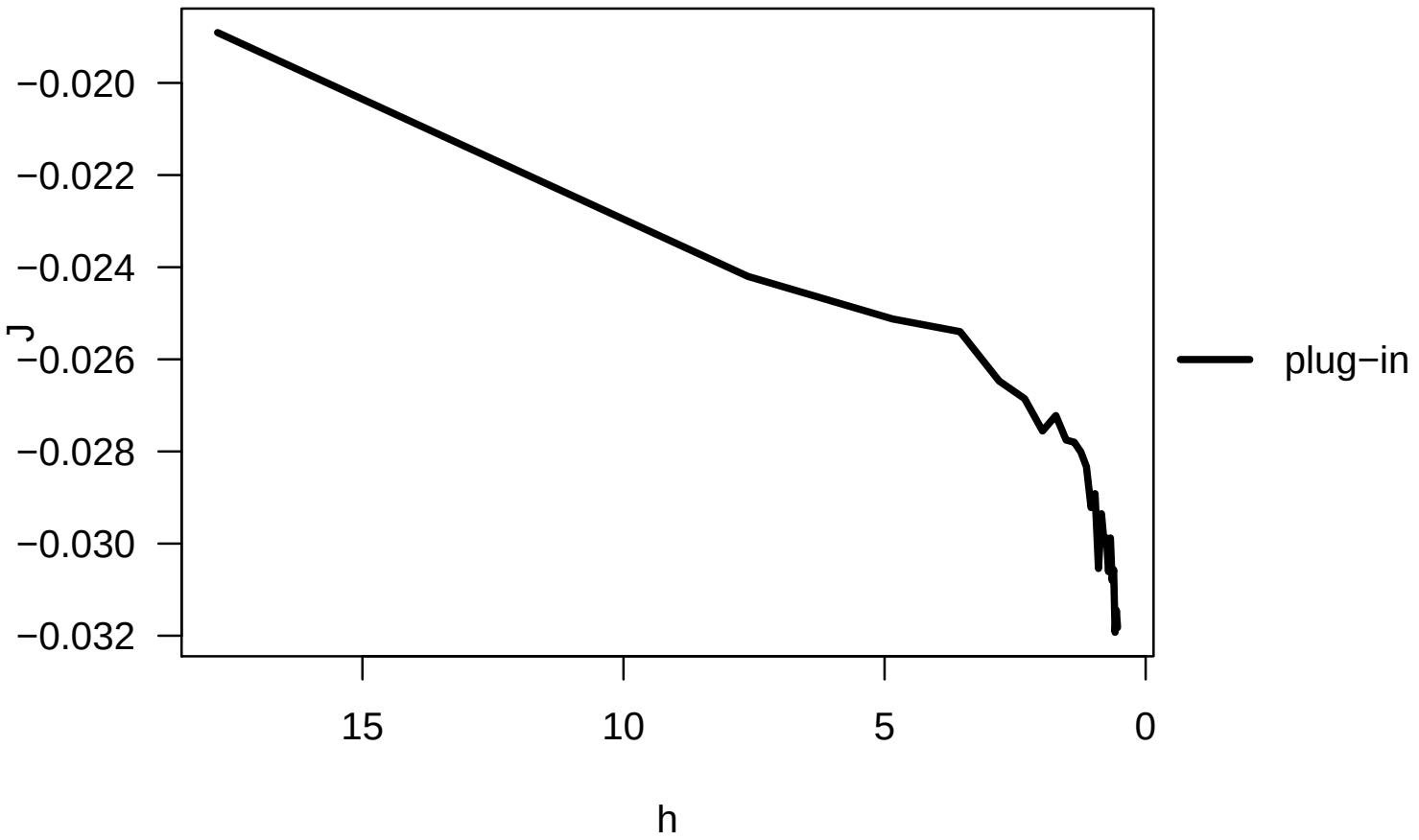
A proposal for $\hat{J}$

- One idea would be to use the plug-in principle: letting $J(h) = \mathbb{E}L(h) - c$,

$$\hat{J}(h) = \int \hat{f}^2(x)dx - 2\sum_i \frac{1}{n}\hat{f}(x_i)$$

- However, this turns out to be a poor estimator

A proposal for $\hat{J}$ (cont'd)



Overfitting

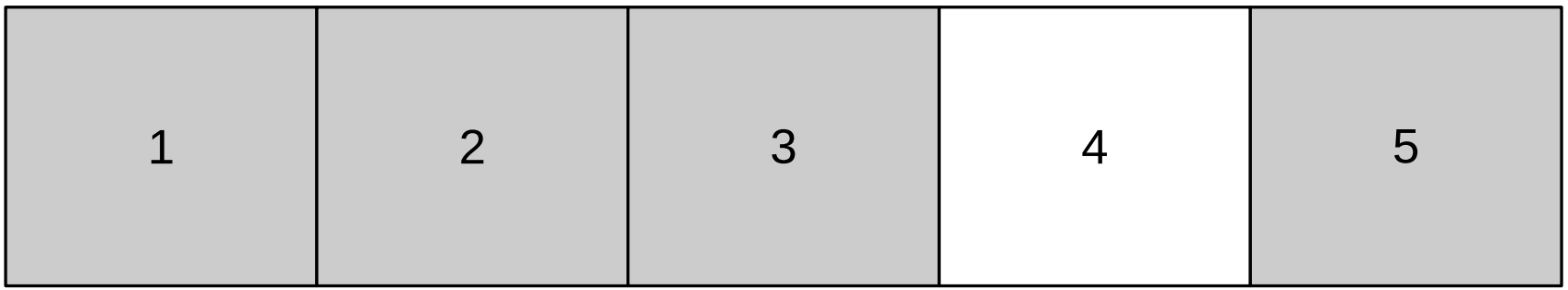
- The proposed $\hat{J}$ is biased downwards and will always indicate that the expected loss is minimized at $h \approx 0$
- The reason for the failure of this estimate is that we are using the data twice: once to fit $\hat{f}$, and then again to estimate the expected loss
- This will reward overfitting

Cross-validation

- A solution is to split the data set into two fractions, then use one portion to fit $\hat{f}$ and the other to evaluate how well $\hat{f}$ seemed to estimate the density of the observations in the second portion
- The problem with this solution is that we rarely have so much data that we can freely part with half of it solely for the purpose of choosing a smoothing parameter
- To finesse this problem, *cross-validation* splits the data into K folds, fits the data on $K-1$ of the folds, and evaluates risk on the fold that was left out

Cross-validation figure

This process is repeated for each of the folds, and the risk averaged across all of these results:



How many folds?

- Common choices for K are 5, 10, and n
- n-fold cross-validation is also known as *leave-one-out* cross-validation
- n-fold cross-validation has the attractive property that it doesn't depend on how the data was randomly split into K folds, although it carries a downside of increased computational burden, as $\hat{f}$ must be fit n separate times

Leave-one-out cross-validation for histograms

- The leave-one-out cross-validation estimate of expected loss for the histogram problem is

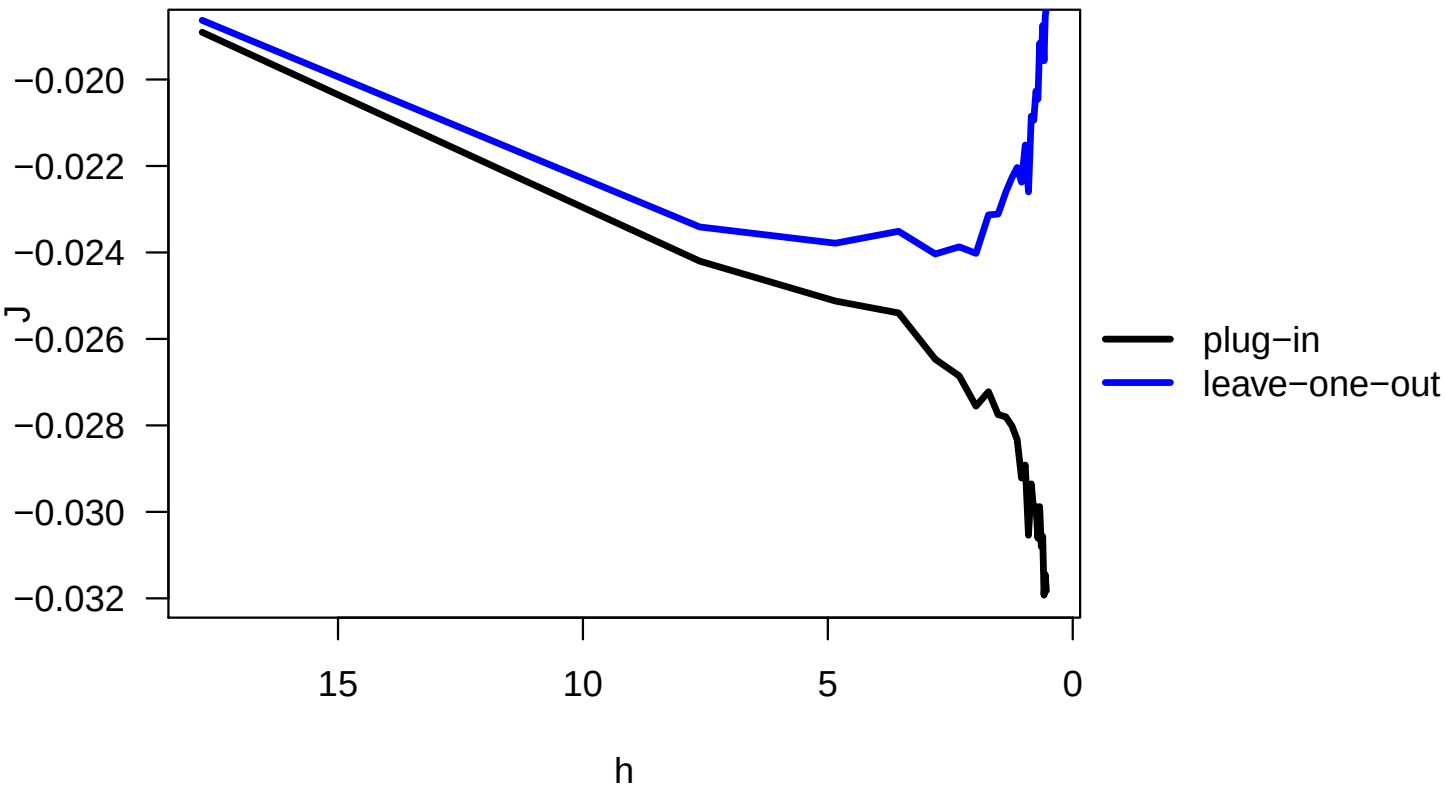
$$\hat{J}(h) = \int \hat{f}^2(x)dx - 2\sum_i \frac{1}{n}\hat{f}_{(-i)}(x_i),$$

  where $\hat{f}_{(-i)}$ denotes the density estimate obtained after removing the ith observation

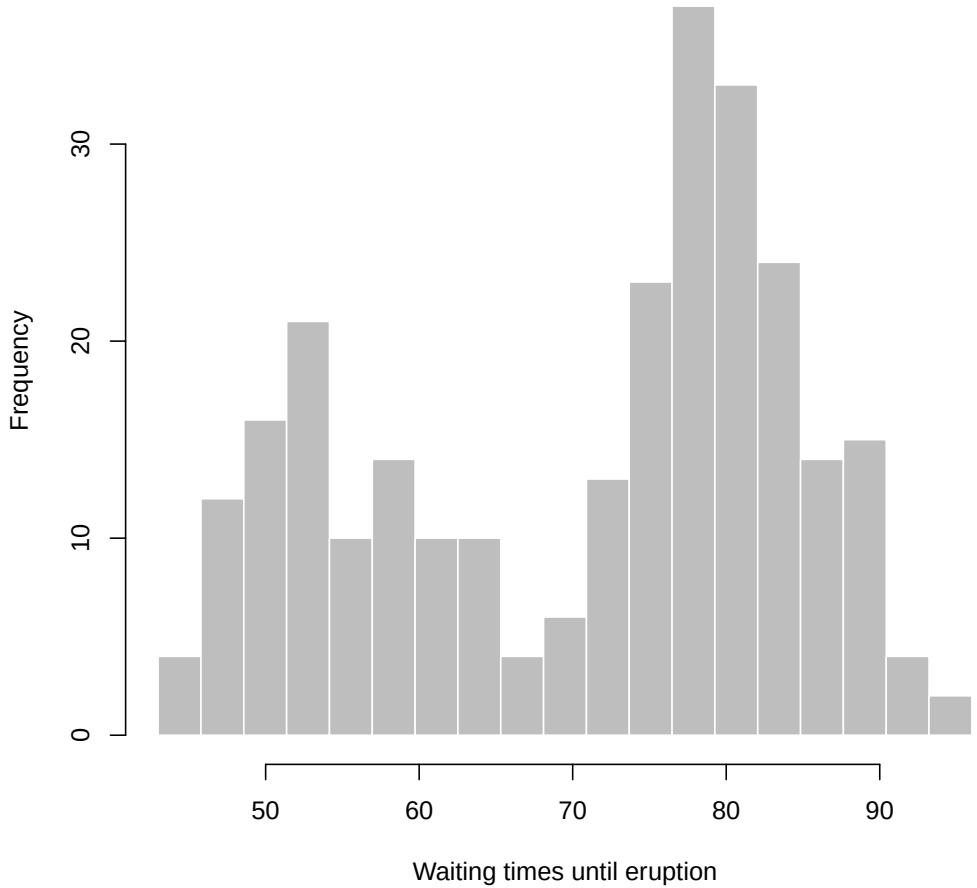
- Because the histogram is such a simple estimator, this expression can actually be worked out in closed form:

$$\hat{J}(h) = \frac{2}{h(n-1)} - \frac{n+1}{h(n-1)}\sum_{j=1}^{m}\hat{p}_j^2$$

Estimated risk: cross-validation



Histogram with optimal number of bins

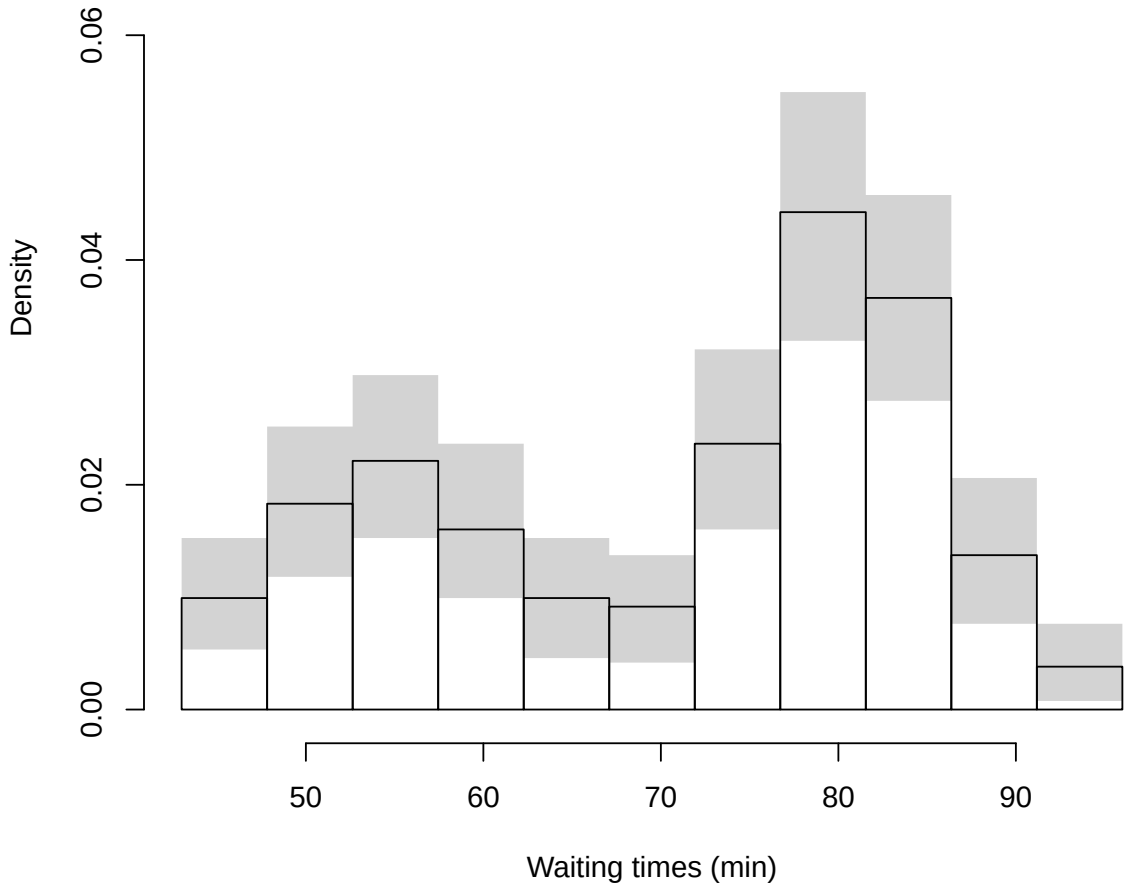


Introduction

- Finally, let's discuss the issue of confidence bands for f
- In general, it is not actually possible to construct confidence bands for f itself – we must settle for confidence bands for $\bar{f}$, the piecewise constant function of average density over the histogram bins
- Recall also the distinction between pointwise confidence intervals (which have 1-α coverage only at a given x) versus confidence bands (which have 1-α coverage for containing the entire $\bar{f}$ over all x)

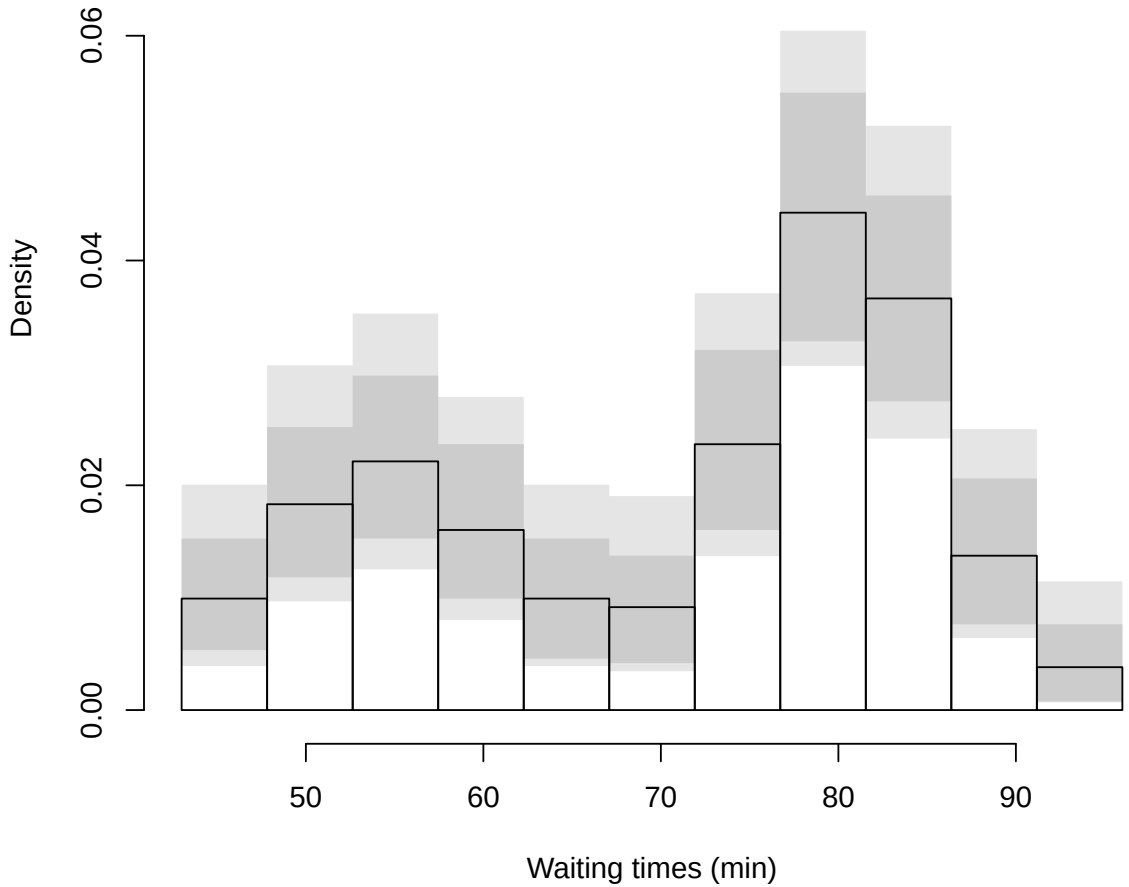
Pointwise binomial intervals

One can construct pointwise confidence intervals based on the binomial distribution:



Bonferroni bands

To obtain confidence bands, one could use a Bonferroni approach, with $\alpha^* = \alpha/m$:



Bootstrap bands

- Another approach is to use the bootstrap
- The two approaches yield similar answers here, but the bootstrap approach will be useful later on when we discuss more complicated methods than the histogram

Bootstrap bands (cont'd)