Week 4

4.1 O notation

Throughout the course, we will consider the accuracy of the various nonparametric methods. For standard, parametric, statistical methods, measures of accuracy of different estimators tend to be very similar; for instance, the variance of an estimator $\widehat{\theta}$ of a single parameter θ is (nearly) always some constant divided by n, the sample size and the bias of an estimator is typically a constant divied by n (or 0, in the case of an unbiased estimator). As we will see, that is not the case with nonparametric methods.

Assessing the accuracy of nonparametric methods serves two important purposes. One is that it allows us to compare different methods of estimation, so that we can choose the most accurate one. A second purpose is that such measures of accuracy give important information about how the accuracy of an estimator relates to the properties of the quantities being estimated.

When analyzing the properties of many statistical methods, such as the nonparametric methods considered in this course, it is convenient to use *O notation*, read as "big-oh notation". This notation provides a simple way to describe the behavior of estimators, and other statistical quantities.

Consider a simple example. Let $Y_1, Y_2, \ldots, Y_n$ denote i.i.d. random variables, each mean μ and standard deviation σ; suppose that we are interested in the moments of the sample mean $\bar{Y}$. It is well-known that $\mathrm{E}\left(\bar{Y}\right) = \mu$ and

$$\mathrm{E}\left(\bar{Y}^2\right) = \mathrm{E}\left(\bar{Y}\right)^2 + \mathrm{Var}\left(\bar{Y}\right) = \mu^2 + \frac{\sigma^2}{n}.$$

Using O notation , the result for the second moment can be written

$$\mathrm{E}\left(\bar{Y}^2\right) = \mu^2 + O(\frac{1}{n}),$$

which means that the second moment of $\bar{Y}$ is μ^2 plus a term that is "of order $1/n$", that is, the term σ^2/n is of the form of a constant divided by n.

In that example, there is not much benefit to using O notation − $\mu^2 + O(1/n)$ is not any simpler than the exact expression $\mu^2 + \sigma^2/n$. However, consider $\mathrm{E}\left(\bar{Y}^3\right)$. By expanding $(\sum_{j=1}^{n} Y_j)^3$ in powers of $Y_1, \ldots, Y_n$ it may be shown that

$$\mathrm{E}\left(\bar{Y}^3\right) = \frac{n^2-1}{n^2}\mu^3 + \frac{3(n-1)}{n^2}\mu\sigma^2 + \frac{\mathrm{E}(Y_1^3)}{n^2}.$$

If we are interested in the general form of $\mathrm{E}\left(\bar{Y}^3\right)$, rather than an exact expression, we could write this result as

$$\mathrm{E}\left(\bar{Y}^3\right) = \mu^3 + \frac{3}{n}\mu\sigma^2 + O(\frac{1}{n^2}) \quad \text{as} \quad n \to \infty. \tag{4.1}$$

This is obtained by noting that

$$\frac{n^2-1}{n^2}\mu^3 = \mu^3 - \frac{1}{n^2}\mu^3$$

and

$$\frac{3(n-1)}{n^2}\mu\sigma^2 = \frac{3}{n}\mu\sigma^2 - \frac{3}{n^2}\mu\sigma^2.$$

It follows that

$$\mathrm{E}\left(\bar{Y}^3\right) = \mu^3 + \frac{3}{n}\mu\sigma^2 + \left(\mathrm{E}(Y_1^3) - \mu^3 - 3\mu\sigma^2\right)\frac{1}{n^2}$$

and the term

$$\left(\mathrm{E}(Y_1^3) - \mu^3 - 3\mu\sigma^2\right)\frac{1}{n^2}$$

can be written $O(1/n^2)$.

In general, consider two functions $g(\cdot)$ and $h(\cdot)$ of a positive argument. The statement

$$h(z) = O(g(z)) \quad \text{as} \quad z \to \infty$$

means, roughly, that $h(z)$ and $g(z)$ are of the same order as $z \to \infty$. More formally, $h(z) = O(g(z))$ as $z \to \infty$ if and only if there exists a constant M and real number $z_0 > 0$ such that

$$|h(z)| \leq M|g(z)| \quad \text{for all} \quad z > z_0;$$

however, this technical definition is rarely needed.

In statistical applications of this concept, the argument z is often the sample size n. For example, we might write

$$\frac{4}{n^2} + \frac{6}{n^{\frac{5}{2}}} = O(n^{-2}) \quad \text{as} \quad n \to \infty;$$

that is $4/n^2 + 6/n^{\frac{5}{2}}$ is of order $O(1/n^2)$ as $n \to \infty$. This holds because we may write

$$\frac{4}{n^2} + \frac{6}{n^{\frac{5}{2}}} = \frac{4}{n^2}(1 + \frac{6}{\sqrt{n}})$$

and

$$4(1 + \frac{6}{\sqrt{n}}) \leq 10 \quad \text{for} \quad n > 1.$$

Note that there is nothing special about the value 10 here; the important point is there *some* finite value that can serve as an upper bound.

A function of n that converges to a nonzero constant as $n \to \infty$ is $O(1)$. For example,

$$\frac{2n^2 + 1/n}{n^2} = O(1) \quad \text{and} \quad \frac{(\sqrt{n}+1)^4}{(n+1)^2} = O(1) \quad \text{as} \quad n \to \infty.$$

In the first case,

$$\frac{2n^2 + 1/n}{n^2} = 2 + \frac{1}{n^3} \to 2 \quad \text{as} \quad n \to \infty$$

and in the second case both

$$(\sqrt{n}+1)^4 \quad \text{and} \quad (n+1)^2$$

have n^2 as the largest power of n, with coefficients of 1 in both cases. Hence,

$$\frac{(\sqrt{n}+1)^4}{(n+1)^2} \to 1 \quad \text{as} \quad n \to \infty.$$

However, convergence is not needed to establish that a term is $O(1)$. For instance, suppose $c_1, c_2, \ldots,$ is a bounded sequence of real numbers that is bounded away from 0; that is, there exist constants a, b such that

$$0 < a \leq c_n \leq b < \infty \quad \text{for all } n.$$

Then

$$\frac{c_n n^2 + 1/n}{n^2} = O(1) \quad \text{as} \quad n \to \infty$$

without requiring that the sequence c_n has a limit or without us even knowing the exact form of c_n.

The same approach may be used to describe limiting behavior as $z \to 0$. In that case,

$$h(z) = O(g(z)) \quad \text{as} \quad z \to 0$$

means that there exists a constant M and real number $z_0 > 0$ such that

$$|h(z)| \leq M|g(z)| \quad \text{for all} \quad |z| < z_0;$$

in many cases, the index z is nonnegative, so that this condition may be written

$$|h(z)| \leq M|g(z)| \quad \text{for all} \quad z < z_0.$$

For example, suppose $\epsilon > 0$ and we are interested in the properties as $\epsilon \to 0$; then we can write

$$4\epsilon + \epsilon^2 + 12\epsilon^3 = O(\epsilon) \quad \text{as} \quad \epsilon \to 0.$$

This follows from the fact that

$$4\epsilon + \epsilon^2 + 12\epsilon^3 = (4 + \epsilon + 12\epsilon^2)\epsilon$$

which is bounded by 17ϵ for $\epsilon < 1$.

In many cases, the functions analyzed this way are more complicated than the simple polynomials considered in the examples here. For instance, because the exponential function $\exp(z)$ may be expanded

$$\exp(z) = 1 + z + z^2/2! + z^3/3! + \cdots,$$

if Z is a random variable with an exponential distribution with density $\exp(-t)$, $t > 0$, we may write

$$\Pr(Z \leq z) = \int_0^z \exp(-t)dt = 1 - \exp(-z) = z + O(z^2) \quad \text{as} \quad z \to 0. \tag{4.2}$$

That is, for small values of z, $\Pr(Z \leq z)$ is approximately equal to z. The reason for this conclusion is that, when z is near 0, the $O(z^2)$ term consisting of

$$z^2/2! + z^3/3! + z^4/4! + \cdots$$

will be smaller than the main term in the expression, z.

Because we have the exact value of $\Pr(Z \leq z)$ in this example, $1 - \exp(-z)$, we can evaluate z and the $O(z^2)$ in (4.2). For instance, for $z = 1/2$,

$$1 - \exp(-z) \doteq 0.393$$

so that the $O(z^2)$ term must be

$$0.393 - 0.5 = -0.107.$$

For $z = 0.1$,

$$1 - \exp(-z) \doteq 0.0952,$$

so the $O(z^2)$ term must be 0.0048. Thus, as z becomes smaller, the $O(z^2)$ term becomes smaller. Hence, if z is small, it may be appropriate to approximate $\Pr(Z \leq z)$ by z.

Although O-notation can be used when the exact expression for the quantity of interest is available, it is more often used when the exact expression is unavailable. The result is then an approximation to the quantity of interest, with the O-term representing the error in the approximation.

There are two important properties of O notation that make it particularly useful in deriving such approximations. One is that, often, it is possible to find the order of an expression without knowing its exact form. For instance, for any constants c_1, c_2, c_3, with $c_1 \neq 0$, we may write

$$\frac{c_1}{n^2} + \frac{c_2}{n^{\frac{5}{2}}} = O(n^{-2}) \quad \text{as} \quad n \to \infty$$

and

$$c_1\epsilon + c_2\epsilon^2 + c_3\epsilon^3 = O(\epsilon) \quad \text{as} \quad \epsilon \to 0$$

without knowing the values of c_1, c_2, c_3.

For example, if Z is a random variable with an exponential distribution with density $\lambda \exp(-\lambda t)$, $t > 0$, where $\lambda > 0$, it follows that

$$\Pr(Z \leq z) = 0 + O(z) = O(z) \quad \text{as} \quad z \to 0$$

for any value of λ.

The other important property of O notation is that is easy to apply algebraic operations to O terms. For instance, we may write

$$O(n)/n = O(1) \quad \text{as} \quad n \to \infty.$$

In general, if a_n and b_n are functions of n, then

$$O(a_n)/b_n = O(a_n/b_n).$$

In particular, multiplying (or dividing) a O-term by a constant does not change the order of the term.

The sum of $O(\cdot)$ terms is simply the term with the largest order. Therefore,

$$O(n^2) + O(n^2) + O(n) = O(n^2) \quad \text{as} \quad n \to \infty$$

and

$$O(\frac{1}{n}) + O(\frac{1}{n^2}) = O(\frac{1}{n}) \quad \text{as} \quad n \to \infty,$$

for example.

We have already seen an example of this when deriving the expansion in (4.2); in that case, we used the fact that

$$z^2/2! + z^3/3! + z^4/4! + \cdots = O(z^2) \quad \text{as} \quad z \to 0.$$

For another example, consider the result

$$\mathrm{E}\left(\bar{Y}^3\right) = \mu^3 + \frac{3}{n}\mu\sigma^2 + O(\frac{1}{n^2}) \quad \text{as} \quad n \to \infty.$$

Because

$$\frac{3}{n}\mu\sigma^2 = O(n),$$

we could also write this as

$$\mathrm{E}\left(\bar{Y}^3\right) = \mu^3 + O(\frac{1}{n}) \quad \text{as} \quad n \to \infty.$$

The order of a product of a $O(\cdot)$ terms is the product of the orders. For instance,

$$O(n)O(\frac{1}{n^2}) = O(\frac{1}{n}) \quad \text{and} \quad O(1)O(\frac{1}{n}) = O(\frac{1}{n}) \quad \text{as} \quad n \to \infty.$$

It is important to keep in mind that in expressions containing several O-terms, each instance of the function $O(\cdot)$ refers to a different function. Thus, $O(n) - O(n) = O(n)$; it would be incorrect to conclude that $O(n) - O(n) = 0$. Also, we can write $2O(n) = O(n)$ and this does not imply that $O(n) = 0$ (as it would if $O(n)$ were a real number).

Often, the use of O-notation is tied to the use of a Taylor's series approximation. For instance, suppose X is a random variable with a continuous distribution with density

$$p(x) = 4x\exp(-2x), \quad x > 0$$

and let $F(\cdot)$ denote the distribution function of X. Consider the behavior of $F(x)$ for x near 0.

We know that

$$F(x) = \int_0^x 4t\exp(-2t)dt$$

so that $F(0) = 0$,

$$F'(0) = p(0) = 4(0)\exp(-0) = 0$$

and

$$F''(0) = p'(0) = 4\exp(-2x) + 4x\exp(-2x)\,\Big|_{x=0} = 4.$$

Using a Taylor's series expansion

$$\begin{aligned} F(x) &= F(0) + F'(0)x + \frac{1}{2}F''(0)x^2 + \frac{1}{6}F'''(0)x^3 + \cdots \\ &= 2x^2 + O(x^3) \quad \text{as} \quad x \to 0. \end{aligned}$$

That is, for small x,

$$\Pr(X \leq x) \doteq 2x^2.$$

A more refined approximation can be obtained by retaining more terms in the expansion. For example,

$$\Pr(X \leq x) = 2x^2 + \frac{8}{3}x^3 + 2x^4 + O(x^5) \quad \text{as} \quad x \to 0.$$

Application to the properties of $\widehat{p}_H$

We now use this notation to give more formal justifications to the results on the limiting behavior of $\mathrm{E}\left(\widehat{p}_H(y)\right)$ and $\mathrm{Var}\left(\widehat{p}_H(y)\right)$ described in Section 3.4. We assume that h is a function of n such that $h \to 0$ as $n \to \infty$ and $nh \to \infty$ as $n \to \infty$. These conditions ensure that the bias and variance of $\widehat{p}_H(y)$ as an estimator of $p(y)$ both approach 0 as $n \to \infty$.

That is, for larger sample sizes, we use a smaller bin width; the requirement that $nh \to \infty$ as $n \to \infty$ puts a limit on how fast h can approach 0; thus, we consider the case in which n is large and h is small, but not too small.

We have seen that, for $y \in (b_{k-1}, b_k]$,

$$\mathrm{E}\left(\widehat{p}_H(y)\right) = \frac{F(b_k) - F(b_{k-1})}{h}.$$

Writing $b_k = b_{k-1} + h$ and using a Taylor's series expansion around $h = 0$,

$$F(b_k) - F(b_{k-1}) = F(b_{k-1} + h) - F(b_{k-1}) = F'(b_{k-1})h + O(h^2) \quad \text{as} \quad h \to 0.$$

Recall that $F'(\cdot) = p(\cdot)$; then, as $h \to 0$,

$$F(b_k) - F(b_{k-1}) = p(b_{k-1})h + O(h^2).$$

Also, we can expand $p(b_{k-1})$ around $b_{k-1} = y$:

$$p(b_{k-1}) = p(y) + p'(y)\left(b_{k-1} - y\right) + \cdots .$$

Because $|y - b_{k-1}| \leq h$,

$$b_{k-1} - y = O(h) \quad \text{as} \quad h \to 0$$

so that

$$p(b_{k-1}) = p(y) + O(h) \quad \text{as} \quad h \to 0.$$

Combining these expansions, we have

$$\begin{aligned} F(b_k) - F(b_{k-1}) &= p(b_{k-1})h + O(h^2) \\ &= \left(p(y) + O(h)\right)h + O(h^2) \\ &= p(y)h + O(h^2) \quad \text{as} \quad h \to 0. \end{aligned}$$

It follows that

$$\begin{aligned}\mathrm{E}\left(\widehat{p}_H(y)\right) &= \frac{F(b_k) - F(b_{k-1})}{h} \\ &= \frac{p(y)h + O(h^2)}{h} \\ &= p(y) + O(h) \quad \text{as} \quad h \to 0.\end{aligned}$$

That is, for small bin widths, $\widehat{p}_H(y)$ is approximately unbiased as an estimator of $p(y)$, essentially the same result we obtained informally in Section 3.4.

Using the same basic approach but analyzing the terms a little more carefully, this result may be refined to

$$F(b_k) - F(b_{k-1}) = p(y)h - p'(y)(y - \bar{b}_k)h + O(h^3) \quad \text{as} \quad h \to 0 \tag{4.3}$$

and, hence,

$$\mathrm{E}\left(\widehat{p}_H(y)\right) = \frac{F(b_k) - F(b_{k-1})}{h} = p(y) - p'(y)(y - \bar{b}_k) + O(h^2) \quad \text{as} \quad h \to 0. \tag{4.4}$$

Here $\bar{b}_k = (b_k + b_{k-1})/2$, the midpoint of the kth interval.

Thus, the bias of $\widehat{p}_H(y)$ as an estimator of $p(y)$ is of the form

$$-p'(y)(y - \bar{b}_k) + O(h^2) \quad \text{as} \quad h \to 0.$$

Note that $|y - \bar{b}_k| \leq h/2$ so that $(y - \bar{b}_k) = O(h)$ which leads to the previous result that the bias is of order $O(h)$ as $h \to 0$. However, the more refined expansion (4.4) gives us more information regarding the bias. Specifically, the bias of $\widehat{p}_H(y)$ is greater when y is farther from the midpoint of the bin it is in; the bias is also larger when $p'(y)$ is large, meaning that the value of $p(\cdot)$ changes rapidly near y.

Now consider $\mathrm{Var}\left(\widehat{p}_H(y)\right)$. In Section 3.4, we have seen that

$$\mathrm{Var}\left(\widehat{p}_H(y)\right) = \frac{F(b_k) - F(b_{k-1})}{h}\frac{1}{nh} - \left(\frac{F(b_k) - F(b_{k-1})}{h}\right)^2 \frac{1}{n}.$$

Using (4.3),

$$\frac{F(b_k) - F(b_{k-1})}{h} = p(y) + O(h) \quad \text{as} \quad h \to 0.$$

It follows that

$$\begin{aligned}\mathrm{Var}\left(\widehat{p}_H(y)\right) &= (p(y) + O(h))\frac{1}{nh} - (p(y) + O(h))^2 \frac{1}{n} \\ &= \frac{p(y)}{nh} - \frac{p(y)^2}{n} + \frac{O(h)}{nh} + \frac{O(h)}{n} + \frac{O(h^2)}{n} \\ &= \frac{p(y)}{nh} + O(n^{-1}). \end{aligned} \tag{4.5}$$

Note that the final line in (4.5) uses the facts that, because $h \to 0$ as $n \to \infty$, both h/n and h/n^2 are of smaller order than $O(1/n)$ and that $p(y)^2/n = O(1/n)$; hence,

$$O(1/n) + O(h/n) + O(h^2/n) = O(1/n).$$

Thus, the variance of $\widehat{p}_H(y)$ is smaller whenever $p(y)$ is smaller and the variance is larger whenever $p(y)$ is larger.

4.2 Accuracy of the histogram estimator

In the previous section consider the properties of the estimator $\widehat{p}_H(y)$ for a given value of the argument y. Here we consider measures of the accuracy of the function $\widehat{p}_H(\cdot)$ as an estimator of $p(\cdot)$.

Recall that if θ is a parameter and $\hat{\theta}$ is an estimator of θ, then the accuracy of $\hat{\theta}$ as an estimator of θ is often measured by its mean squared error (MSE):

$$\begin{aligned} \text{MSE} &= \text{E}\left((\hat{\theta} - \theta)^2\right) \\ &= \left(\text{E}(\hat{\theta}) - \theta\right)^2 + \text{Var}(\widehat{\theta}) \\ &= (\text{bias})^2 + \text{Var}(\widehat{\theta}). \end{aligned}$$

Fix a value of y and consider the histogram estimator $\widehat{p}_H(y)$ as an estimator of $p(y)$. The MSE of $\widehat{p}_H(y)$ is given by

$$\text{MSE}(y) = \text{E}\left((\widehat{p}_H(y) - p(y))^2\right)$$

which may be written

$$\text{MSE}(y) = \left(\text{E}\left(\widehat{p}_H(y)\right) - p(y)\right)^2 + \text{Var}\left(\widehat{p}_H(y)\right).$$

For $b_{k-1} < y \leq b_k$, (4.4) shows that

$$\text{E}\left(\widehat{p}_H(y)\right) - p(y) = -p'(y)(y - \bar{b}_k) + O(h^2)$$

and, according to (4.5),

$$\text{Var}\left(\widehat{p}_H(y)\right) = \frac{p(y)}{nh} + O(n^{-1}).$$

It follows that, for $b_{k-1} < y \leq b_k$,

$$\text{MSE}(y) = p'(y)^2(y - \bar{b}_k)^2 + O(h^3) + p(y)\frac{1}{nh} + O(n^{-1}). \tag{4.6}$$

Although, for some purposes, a measure of accuracy that depends on the value of y under consideration is useful, for other purposes, it is desirable to have a meaure of accuracy of the function $\widehat{p}_H(\cdot)$ as an estimator of the density function $p(\cdot)$.

One such measure may be obtained by summarizing $\text{MSE}(y)$ by its integral

$$\int_{-\infty}^{\infty} \text{MSE}(y) dy,$$

which may be written as

$$\int_{-\infty}^{\infty} \text{MSE}(y) dy = \int_{-\infty}^{\infty} \text{E}\left((\widehat{p}_H(y) - p(y))^2\right) dy;$$

this quantity is known as the *integrated mean squared error* (IMSE).

Integrating the expansion in (4.6), we obtain the following expansion for the ISE:

$$\text{IMSE} = \sum_{k=1}^{m} \int_{b_{k-1}}^{b_k} p'(y)^2 (\bar{b}_k - y)^2 dy + \frac{1}{nh} \int_{-\infty}^{\infty} p(y) dy + O(h^3) + O(n^{-1}). \tag{4.7}$$

This result makes the assumption that the integral of the $O(\cdot)$ terms are of the same order as the term itself; e.g.,

$$\int_{-\infty}^{\infty} O(h^3) dy = O(h^3);$$

note that the $O(h^3)$ term depends on y so that such a result is plausible. Finally, using the fact that

$$\int_{-\infty}^{\infty} p(y) dy = 1,$$

(4.7) becomes

$$\text{IMSE} = \sum_{k=1}^{m} \int_{b_{k-1}}^{b_k} p'(y)^2 (\bar{b}_k - y)^2 dy + \frac{1}{nh} + O(h^3) + O(n^{-1}). \tag{4.8}$$

To simplify the expression for the IMSE, we can approximate the integrals in (4.8). Here we consider give the leading terms in the approximation; it may be shown that the remaining terms are of smaller order and, more importantly, they do not change to order of the remainder terms ($O(h^3)$ and $O(n^{-1})$) in the expansion (4.8).

Note that, using properties of integrals,

$$\int_{b_{k-1}}^{b_k} p'(y)^2 (\bar{b} - y)^2 dy \doteq p'(\bar{b}_k)^2 \int_{b_{k-1}}^{b_k} (\bar{b} - y)^2 dy = p'(\bar{b}_k)^2 \frac{h^3}{12}$$

and that

$$\int_{b_{k-1}}^{b_k} p'(y)^2 dy \doteq p'(\bar{b}_k)^2 h;$$

it follows that

$$\int_{b_{k-1}}^{b_k} p'(y)^2(\bar{b}_k - y)^2 dy \doteq \int_{b_{k-1}}^{b_k} p'(y)^2 dy \frac{h^2}{12}.$$

Using this result, it may be shown that

$$\text{IMSE} = \frac{h^2}{12}\int_{-\infty}^{\infty} p'(y)^2 dy + \frac{1}{nh} + O(h^3) + O(n^{-1}). \tag{4.9}$$

The leading terms in the expansion (4.9) give what is sometimes called the "asymptotic" IMSE (AIMSE); thus,

$$\text{AIMSE} = \frac{h^2}{12}\int_{-\infty}^{\infty} p'(y)^2 dy + \frac{1}{nh}.$$

Note that, if the density $p(\cdot)$ is such that $p(y) > 0$ if and only if y is in an interval (a, b) and $p(\cdot)$ is differentiable on (a, b) but not necessarily at $y = a$ or $y = b$, then the expression for the AIMSE is given by

$$\text{AIMSE} = \frac{h^2}{12}\int_{a}^{b} p'(y)^2 dy + \frac{1}{nh}.$$

For simplicity, we will continue to write the integral in the expression for the AIMSE as an integral over the entire real line, with the understanding that if $p(y) > 0$ if and only if $y \in (a, b)$, the limits on the integral should be a and b.

The AIMSE gives a relatively simple expression for the accuracy of a histogram as an estimator of a density function and it can be used to derive important qualitative information regarding histograms.

(1) The AIMSE approaches 0 as $n \to \infty$ (so that, in the limit, $\widehat{p}_H(\cdot)$ is a "perfectly accurate" estimator of $p(\cdot)$) only if $h \to 0$ as $n \to \infty$ **and** $nh \to \infty$ as $n \to \infty$. There are two important implications of this. One is that, in general, we should use a smaller bin width for large sample sizes than for small sample sizes. The other is that it is important that the bin width is not too small; that is, h should be small, but nh should be large; otherwise, the bias of the estimator will be small, but its variance will be large.

(2) The accuracy of $\widehat{p}_H(\cdot)$ as an estimator of $p(\cdot)$ depends on $p(\cdot)$ primarily through the quantity

$$\int_{-\infty}^{\infty} p'(y)^2 dy,$$

which may be viewed as a measure of the local variation in $p(\cdot)$. In particular, this term is minimized if $p(\cdot)$ is constant on the interval on which $p(y) > 0$; that is, the histogram estimator of a density is particularly accurate if $p(\cdot)$ is the density of the uniform distribution on some interval.

(3) The first term in the AIMSE,

$$\frac{h^2}{12}\int_{-\infty}^{\infty} p'(y)^2 dy$$

is a measure of the asymptotic squared bias of $\widehat{p}_H(\cdot)$, while the second term, $1/(nh)$, is a measure of the asymptotic variance. Note that the asymptotic variance does not depend on the density $p(\cdot)$.

If

$$\frac{h^2}{12}\int_{-\infty}^{\infty} p'(y)^2 dy > \frac{1}{nh},$$

that is, if

$$h^3 > \frac{12}{\int_{-\infty}^{\infty} p'(y)^2 dy}\frac{1}{n},$$

then the bias makes a larger contribution to the AIMSE than does the variance; otherwise, the variance makes the larger contribution (or, if the two terms are equal, then the contributions are equal).

Thus, if the bin width is larger than a certain multiple of $1/n^{\frac{1}{3}}$, then the error in the estimator is attributable more to bias than to variance; otherwise, it is attributable more to the variance than to the bias.

(4) The rate at which the AIMSE approaches 0 is maximized by choosing $h = O(n^{-\frac{1}{3}})$ and, for such a choice, the squared bias and the variance of the histogram estimator are both of order $O(n^{-\frac{2}{3}})$. That is, the rate at which the AIMSE approaches 0 is $O(n^{-\frac{2}{3}})$.

Note that, when estimating the parameter of a parametric model, the usual rate of convergence of the MSE is $O(n^{-1})$; for instance, for estimating the mean of a distribution using a sample of size n, the MSE is σ^2/n, where σ^2 is the variance of the distribution. However, in nonparametric estimation problems, the rate of convergence of the mean squared error, or a similar measure such as the AIMSE, is slower than $O(n^{-1})$; for instance, for the histogram estimator, it is $O(n^{-\frac{2}{3}})$. This is one way in which nonparametric estimation is fundamentally "more difficult" than parametric estimation.

Example 4.1 Suppose that the density $p(\cdot)$ is given by

$$p(y) = \frac{1}{(1+y)^2}, \quad y > 0;$$

this is the density of a type of Pareto distribution.

For this density,

$$\int_0^{\infty} p'(y)^2 dy = \int_0^{\infty} \frac{4}{(1+y)^6} dy = \frac{4}{5}$$

so that the AIMSE is given by

$$\frac{h^2}{15} + \frac{1}{nh}.$$

Following point (3) above, if

$$\frac{h^2}{15} > \frac{1}{nh},$$

that is, if

$$h^3 > \frac{15}{n}$$

then the bias makes a larger contribution to the AIMSE than does the variance. □

4.3 Choosing the bin width of the histogram

The results on the accuracy of the histogram estimator, given in the previous section, may be used to suggest an appropriate value for the bin width h.

Let

$$R_1 = \int_{-\infty}^{\infty} p'(y)^2 dy.$$

Then the AIMSE of the histogram estimator based on a bid width of h is given by

$$\text{AIMSE} = \frac{h^2}{12} R_1 + \frac{1}{nh}.$$

Suppose we choose h to minimize the AIMSE. Note that

$$\frac{d}{dh}\text{AIMSE} = \frac{h}{6} R_1 - \frac{1}{nh^3}.$$

Setting this equal to 0 and solving for h yields the solution

$$h_{opt} = \left(\frac{6}{R_1}\right)^{\frac{1}{3}} \frac{1}{n^{\frac{1}{3}}}.$$

Note that

$$\frac{d^2}{dh^2}\text{AIMSE} = \frac{1}{6} R_1 + \frac{3}{nh^4} > 0$$

for all $h > 0$ so that h_{opt} minimizes the AIMSE.

Example 4.2 Consider the Pareto density analyzed in Example 4.1. For this density, we have seen that

$$R_1 = \frac{4}{5}$$

so that

$$h_{opt} = \left(\frac{15}{2}\right)^{\frac{1}{3}} \frac{1}{n^{\frac{1}{3}}} \doteq \frac{1.96}{n^{\frac{1}{3}}}.$$

For instance, if $n = 100$, then

$$h_{opt} = \left(\frac{15}{2}\right)^{\frac{1}{3}} \frac{1}{(100)^{\frac{1}{3}}} \doteq 0.422.$$

□

In practice, R_1 is, of course, unknown, and because the density $p(\cdot)$ is unknown, estimation of R_1 is a difficult problem (although one that we will consider in the following chapter).

An important factor in R_1 is the variability of the distribution, as estimated, for example, by the sample standard deviation of the data or its "interquartile range" (IQR), the difference between the upper and lower quartiles of the data. Hence, here we consider an approach to choosing a value for R_1 based on estimating the variability of the distribution.

Consider a family of density functions with "scale parameter" $\sigma > 0$:

$$p(y;\sigma) = \frac{1}{\sigma} p_0(y/\sigma), \quad -\infty < y < \infty$$

where p_0 is a density function with standard deviation 1. That is, if $p(\cdot)$ is the density of a random variable X with standard deviation σ, then $p_0(\cdot)$ is the density of X/σ.

Then

$$p'(y;\sigma) = \frac{1}{\sigma^2} p_0'(y/\sigma)$$

and

$$R_1 = \frac{1}{\sigma^4} \int_{-\infty}^{\infty} p_0'(y/\sigma)^2 dy.$$

Using the change-of-variable $u = y/\sigma$,

$$R_1 = \frac{1}{\sigma^3} \int_{-\infty}^{\infty} p_0'(u)^2 du = \frac{R_{10}}{\sigma^3} \tag{4.10}$$

where

$$R_{10} = \int_{-\infty}^{\infty} p_0'(u)^2 du.$$

With this expression for R_1, h_{opt} may be written

$$h_{opt} = \left(\frac{6}{R_{10}}\right)^{\frac{1}{3}} \frac{\sigma}{n^{\frac{1}{3}}}.$$

Consider estimation of σ using an estimator based on the IQR of the data. Such an estimator is of the form c_0IQR for a constant c_0 that depends on the shape of the distribution under consideration; for instance, for normally distributed data, σ can be estimated by IQR/1.349. Therefore, an estimator of the optimal bin width is of the form

$$c\frac{\text{IQR}}{n^{\frac{1}{3}}}$$

for some constant c, which depends on the shape of the distribution with density $p(\cdot)$ (but not its standard deviation). Of course, in our context, the density $p(\cdot)$ is unknown. However, it has been shown empirically that $c = 2$ is often a good choice, leading to a bin width of

$$h_{FD} = 2\frac{\text{IQR}}{n^{\frac{1}{3}}};$$

this is known as the *Freedman-Diaconis* bin width choice.

Example 4.3 Consider the data in the variable `speed`, described in Example 3.6. The command

```
> hist(speed, breaks="FD", freq=F)
```

computes a histogram of the data using the bin width based on the Freedman-Diaconis rule. A plot of the histogram is given in Figure 4.1. However, as noted earlier, the `hist` function modifies the breaks specified to improve the appearance of the histogram.

The bins that were used in the histogram construction can be obtained from the component `$breaks` of the result of the `hist` function:

```
> hist(speed, breaks="FD", freq=F)$breaks
 [1]  600  650  700  750  800  850  900  950 1000 1050 1100
```

Thus, the Freedman-Diaconis rule chose of a bin width of approximately 50, leading to 10 bins.

In order to use the actual value of h given by the Freedman-Diaconis rule, we can calculate it directly and then use the function `truehist` in the package “MASS”. The value of h_{FD} for the `speed` data is given by

```
> 2*IQR(speed)/length(speed)^(1/3)
[1] 36.63
```

To plot a histogram using that value of h, we can use the commands

```
> library(MASS)
> truehist(speed, h=36.63, ylab="density", col=0)
```

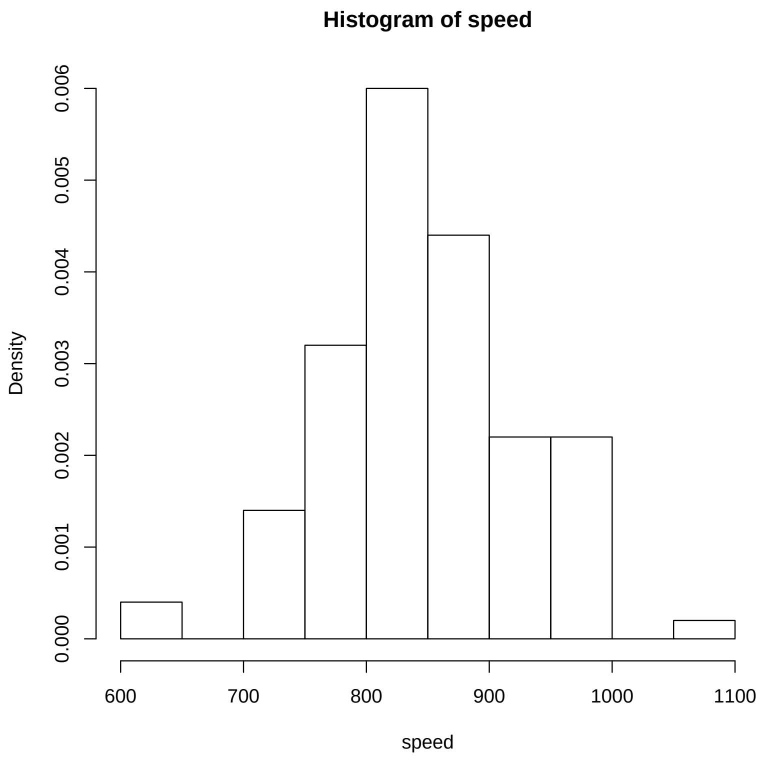

Figure 4.1: Histogram of the Michelson Speed of Light Data Based on the Freedman-Diaconis Rule

The argument `ylab="density"` specifies the label for the y-axis and `col=0` specifies the color to use for the shading of the bars in the histogram (the default is blue). The histogram is displayed in Figure 4.2; note that here there are 14 bins used. □

Example 4.4 For the Pareto density described in Example 4.1, we have seen that the optimal choice of h is, approximately, $2/n^{\frac{1}{3}}$. The lower and upper quantiles of the distribution may be shown to be 1/3 and 3, respectively, so that the IQR of the distribution is 8/3. Therefore, we expect that the Freedman-Diaconis rule will choose a value of h of about

$$\frac{(16/3)}{n^{\frac{1}{3}}} \doteq \frac{5.3}{n^{\frac{1}{3}}};$$

thus, for this distribution, the Freedman-Diaconis rule will tend to choose fewer bins than the optimal number. □

4.4 Kernel Density Estimators

One drawback of histograms as estimators of a density function $p(\cdot)$, and an important source of their inaccuracy, is that they are discontinuous functions. It is easy to see the reason for

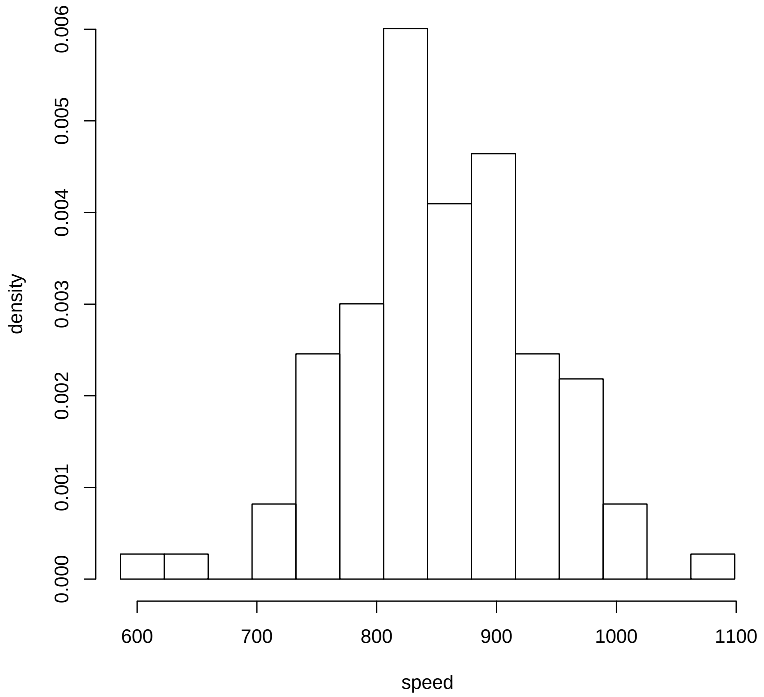

Figure 4.2: Histogram of the Michelson Based on the Freedman-Diaconis Rule Using the truehist Function

the discontinuities. Suppose that break points $b_0, b_1, \ldots, b_m$ are used, with $h = b_j - b_{j-1}$ and let f_j denote the number of observations falling in the j bin, $(b_{j-1}, b_j]$.

If, for example, y is in the interval $(b_0, b_1]$ then the estimate of $p(y)$ is $f_1/(nh)$, while if y is in the interval $(b_1, b_2]$ then the estimate is $f_2/(nh)$. Thus, a histogram density estimator is, in general, discontinuous at each of $b_1, b_2, \ldots, b_{m-1}$.

A better approach would be to estimate the density at a point y by using the number of observations in the interval $(y - h/2, y + h/2)$ where $h > 0$ is a given value. To do this, let

$$K_0(y) = \begin{cases} 1 & \text{if } |y| < 1/2 \\ 0 & \text{otherwise} \end{cases}.$$

Define an estimator $\widehat{p}_N(\cdot)$ by

$$\widehat{p}_N(y) = \frac{\sum_{j=1}^{n} \frac{1}{h} K_0(\frac{y - Y_j}{h})}{n}, \quad -\infty < y < \infty.$$

Fix a value y. Then

$$K_0(\frac{y-Y_j}{h}) = \begin{cases} 1 & \text{if } \left|\frac{y-Y_j}{h}\right| < 1/2 \\ 0 & \text{otherwise} \end{cases}$$

$$= \begin{cases} 1 & \text{if } y - h/2 < Y_j < y + h/2 \\ 0 & \text{otherwise} \end{cases}.$$

Therefore,

$$\sum_{j=1}^{n} K_0\left(\frac{y-Y_j}{h}\right)$$

simply counts the number of observations that fall in the interval $(y - h/2, y + h/2)$. It follows that

$$\widehat{p}_N(y) = \frac{1}{nh}\#\left(Y_j \text{ in } (y - h/2, y + h/2)\right)$$

so that $\widehat{p}_N(\cdot)$ is like a "moving histogram"; it is sometimes called the *naive* density estimator.

Note that, because it is based on counting, $\widehat{p}_N(\cdot)$, like the histogram estimator $\widehat{p}_H(\cdot)$, is discontinuous. However, there is sense in which $\widehat{p}_N(\cdot)$ is "smoother" than $\widehat{p}_H(\cdot)$.

Example 4.5 Consider the speed-of-light data analyzed in Example 3.6. Figure 4.3 contains a plot of the naive kernel density estimator corresponding to $h = 50$, so that the naive kernel estimator uses the same effective bin length as that selected by the Freedman-Diaconis rule; see Example 4.3.

Comparing the estimate in Figure 4.3 to the one in Figure 4.1, we see that the naive kernel estimate is smoother than the histogram estimate but it still has a somewhat "ragged" appearance. □

The reason for the ragged appearance of $\widehat{p}_N(\cdot)$ is that, when calculating $\widehat{p}_N(y)$, each Y_j is either close enough to y to be counted or it is not close enough to be counted. More formally, in estimating $p(y)$, observations Y_j satisfying $|y - Y_j| < h/2$ contribute 1 to the sum

$$\sum_{j=1}^{n} K_0(\frac{y-Y_j}{h})$$

and observations Y_j satisfying $|y - Y_j| \geq h/2$ contribute 0 to that sum. That is, the function $K_0(\cdot)$ is not continuous and, hence, the resulting density estimator is not continuous.

This discussion suggests that a better approach to estimating a density function might be to use a weighted sum in forming the density estimator such that, when estimating $p(y)$, observations Y_j that are closer to y contribute more to the sum and observations farther from y contribute less to the sum. But instead of using only the values 0 and 1, as in the naive

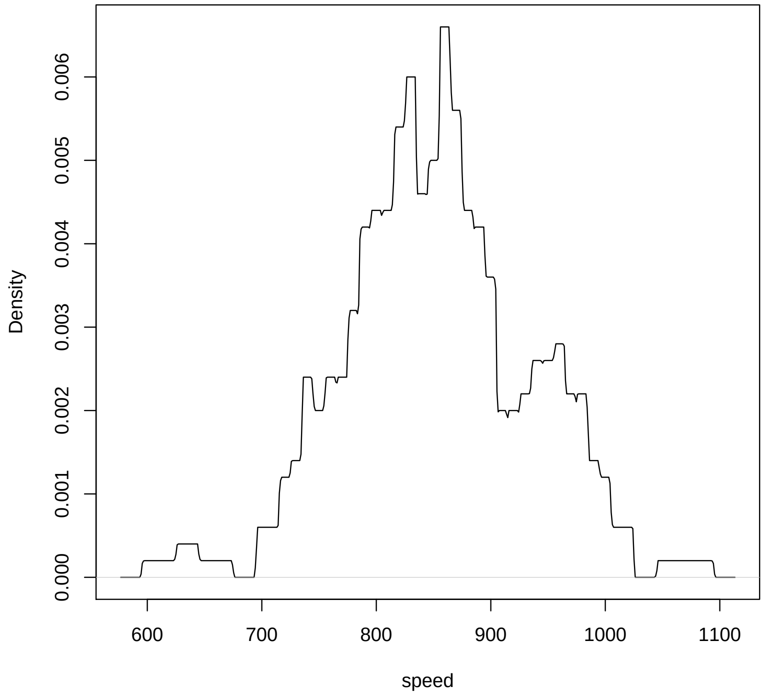

Figure 4.3: Naive Density Estimate of the Michelson Speed of Light Data

estimator, the weights could vary smoothly with the distance $|y - Y_j|$; that is, we could base the estimator on a continuous function $K(\cdot)$, leading to an estimator that is a continuous function.

Let $K(\cdot)$ denote a continuous, nonnegative function satisfying

$$\int_{-\infty}^{\infty} K(u)du = 1;$$

That is, $K(\cdot)$ is a density function; in this context, the function $K(\cdot)$ is known as a *kernel.*

For a positive real number h, define a *kernel density estimator* $\widehat{p}(\cdot)$ by

$$\widehat{p}(y) = \frac{1}{nh}\sum_{j=1}^{n} K(\frac{y - Y_j}{h}), \quad -\infty < y < \infty;$$

h is known as the *smoothing parameter* or the *bandwidth.* Recall that the naive estimator has the same form as a kernel estimator; however, in that case, the function $K_0(\cdot)$ is not continuous so that $\widehat{p}_N(\cdot)$ is not continuous.

Note that a kernel estimator $\widehat{p}(\cdot)$ is a genuine density function in the sense that it is non-negative (because $K(\cdot)$ is non-negative) and it integrates to 1:

$$\begin{aligned}
\int_{-\infty}^{\infty} \widehat{p}(y)dy &= \int_{-\infty}^{\infty} \frac{1}{nh}\sum_{j=1}^{n} K(\frac{y-Y_j}{h})dy \\
&= \frac{1}{nh}\sum_{j=1}^{n}\int_{-\infty}^{\infty} K(\frac{y-Y_j}{h})dy \\
&= \frac{1}{n}\sum_{j=1}^{n}\int_{-\infty}^{\infty} K(u)du \quad \text{using the change-of-variable} \quad u = (y-Y_j)/h \\
&= 1.
\end{aligned}$$

The kernel $K(\cdot)$ is generally chosen to be a differentiable function and to be symmetric about 0 so that $K(-u) = K(u)$ for all $-\infty < u < \infty$ and, hence,

$$\int_{-\infty}^{\infty} uK(u)du = 0.$$

Commonly-used kernels include

- Gaussian:

$$K(u) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right), \quad -\infty < u < \infty$$

- biweight:

$$K(u) = \begin{cases} \frac{15}{16}(1-u^2)^2 & \text{if } |u| < 1 \\ 0 & \text{otherwise} \end{cases}$$

- Epanechnikov:

$$K(u) = \begin{cases} \frac{3}{4}(1-u^2/5)/\sqrt{5} & \text{if } |u| < \sqrt{5} \\ 0 & \text{otherwise} \end{cases}$$

- triangular:

$$K(u) = \begin{cases} 1-|u| & \text{if } |u| < 1 \\ 0 & \text{otherwise} \end{cases}$$

Often, kernels are scaled so that they have standard deviation 1 when viewed as density functions; then h is the standard deviation of the scaled kernel $K(u/h)/h$ used in the kernel density estimator. For example, for the triangular kernel,

$$\int_{-\infty}^{\infty} u^2K(u)du = \int_{-1}^{1} u^2(1-|u|)du = 2\int_0^1 u^2(1-u)du = \frac{1}{6}.$$

Therefore, the scaled triangular kernel is given by

$$\sqrt{6}(1-\sqrt{6}|u|), \quad |u| < \frac{1}{\sqrt{6}}.$$

This type of rescaling makes the values of h used in different kernels roughly comparable. For instance, using $h = 5$ with the Gaussian kernel is about the same as using $h = 5$ for the scaled triangular kernel. We will assume that all kernels that we use are scaled in this way; that is, we assume that for any kernel $K(\cdot)$,

$$\int_{-\infty}^{\infty} u^2 K(u) du = 1.$$

Different kernels lead to different density estimates, although they tend to be very similar. Unless otherwise specified, we will always use the Gaussian kernel.

Example 4.6 To calculate a kernel density estimate in R we can use the function `density`. A number of different kernels are available, including the four discussed in this section. The argument `kernel` is used to select the kernel; only the first letter of the name is needed: `"g"` (Gaussian), `"b"` (biweight), `"e"` (Epanechnikov), and `"t"` (triangular), among others.

The argument `bw` is used to specify the value of the smoothing parameter to use. Thus, the command

```
> density(speed, bw=10, kernel="g")
```

calculates the kernel density estimate for the data in the variable `speed`, using the Gaussian kernel and $h = 10$. A Gaussian kernel is the default so `kernel="g"` does not need to be specified.

Using the result of the function `density` as the argument of the function `plot` produces a plot of the density estimate. For example,

```
> plot(density(speed, bw=10))
```

Other, optional, arguments can be included in `plot` if the default values produce undesirable results; for example, it is possible to specify the title for the plot (the argument `main`) and the label for the x-axis (the argument `xlab`). If multiple estimates are being compared, it is useful to use the same scale on the x-axis; the argument `xlim` can be used to specify this.

Figure 4.4 contains plots of density estimates for the speed-of-light data for the four kernels discussed in this section, using $h = 10$; Figure 4.5 contains similar plots using $h = 20$.

Note the differences between estimates based on different kernels but the same value of h tend to be small relative to the differences between estimates based on the same kernel but different values of h. The main difference between estimates based on different kernels is that "smoother" kernels, such as the Gaussian, which has derivatives of all orders, tend to produce "smoother" estimates. Larger values of h tend to produce smoother estimates for any choice of kernel. □

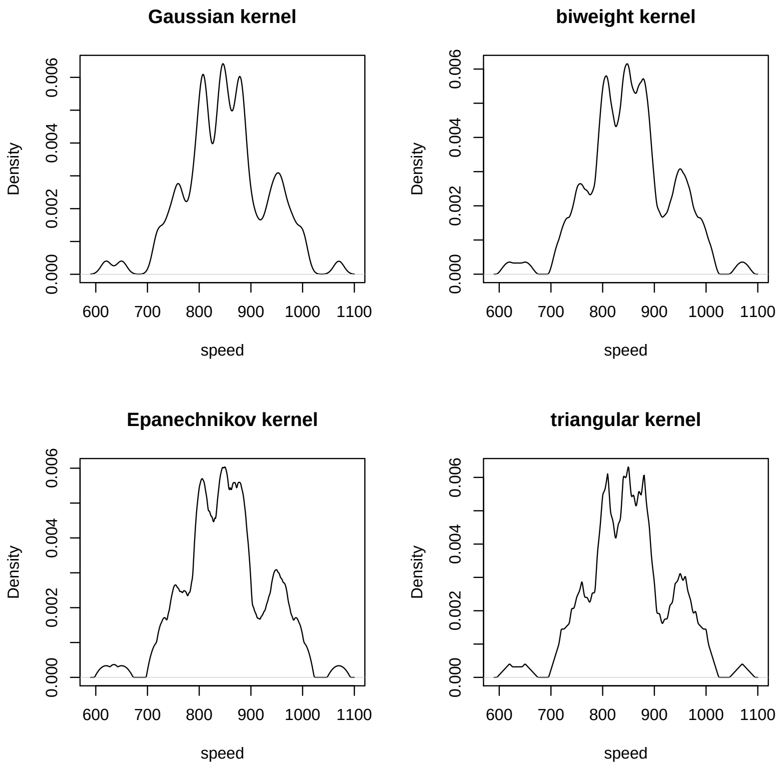

Figure 4.4: Density Estimates of the Michelson Speed of Light Data using $h = 10$

4.5 Exercises

4.1. Find the order of each of the following expressions as $n \to \infty$.

(a)

$$\frac{2}{n} + \frac{3\log(n)}{n^2}$$

(b)

$$\frac{4n^2+1}{(n+1)^2}$$

(c)

$$O(n^2)O(\frac{1}{n}) + 4$$

(d)

$$O(\frac{1}{n^2}) + O(\frac{1}{\sqrt{n}}) + O(1)$$

4.2. (a) Let X denote a random variable with a Poisson distribution with mean λ; that is, X has a discrete distribution with frequency function

$$\frac{\lambda^x \exp(-\lambda)}{x!}, \quad x = 0, 1, 2, \ldots.$$

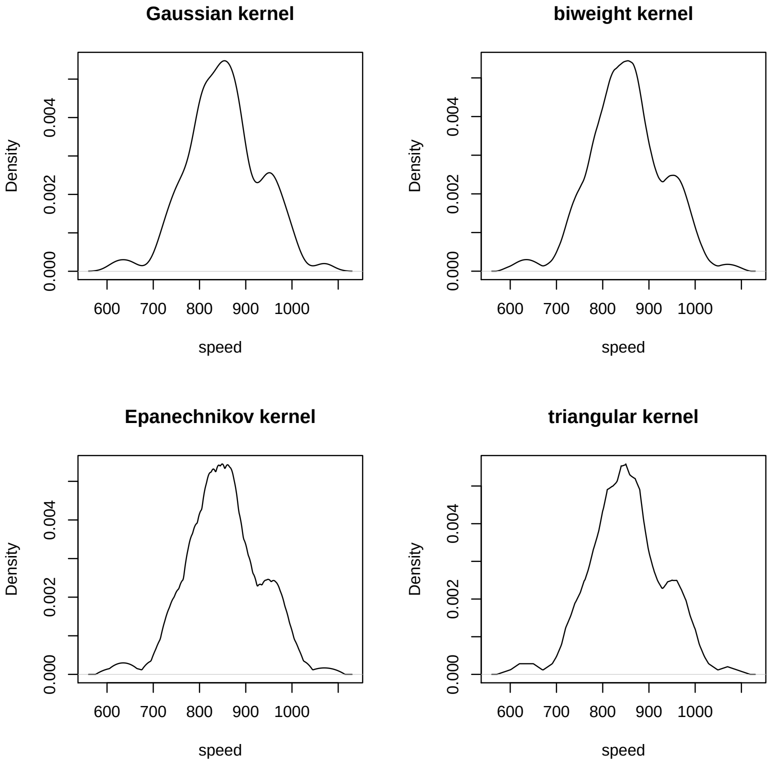

Figure 4.5: Density Estimates of the Michelson Speed of Light Data using $h = 20$

Find the order of

$$\frac{\Pr(X = n+1)}{\Pr(X = n)}$$

as $n \to \infty$.

(b) Let Y denote a random variable with a Poisson distribution with mean $n\lambda$; for instance, Y might be the sum of n independent Poisson random variables each with mean λ. Find the order of

$$\frac{\Pr(Y = n+1)}{\Pr(Y = n)}$$

as $n \to \infty$.

4.3. Let $X_1, X_2, \ldots, X_{100}$ denote independent random variables, each with a continuous distribution with density function

$$p_X(x) = \frac{3}{2}(x-1)^2, \quad 0 < x < 2$$

and let $Y_1, Y_2, \ldots, Y_{100}$ denote independent random variables, each with a continuous distribution with density function

$$p_Y(y) = \frac{3}{8}y^2, \quad 0 < y < 2.$$

Suppose that a histogram is constructed for each set of data, using a bin width of $h = 0.1$ in both cases (that is, the bins are $(0, 0.1], (0.1, 0.2], \ldots$) and let $\widehat{p}_{XH}(\cdot)$ and $\widehat{p}_{YH}(\cdot)$ denote the corresponding histogram density estimators.

(a) Use the AIMSE to determine which estimator do you expect to be more accurate? That is, do expect $\widehat{p}_{HX}(\cdot)$ to be a better estimator of $p_X(\cdot)$ than $\widehat{p}_{HY}(\cdot)$ is of $p_Y(\cdot)$? Or do you expect the reverse to be true?

(b) Now suppose that we are primarily interested in estimating the density at the argument 1; that is, we want to estimate either $p_X(1)$ or $p_Y(1)$. Using the leading terms in the expansions for the MSE of $\widehat{p}_{HX}(x)$ and $\widehat{p}_{HY}(y)$, do you expect $\widehat{p}_{HX}(1)$ to be a better estimator of $p_X(1)$ than $\widehat{p}_{HY}(1)$ is of $p_Y(1)$? Or do you expect the reverse to be true? Why?

4.4. Consider the failure data for a software system, available in the dataset “software”. Construct a histogram for the data, choosing the bin width using the Freedman-Diaconis method. Compare the number of bins corresponding to the Freedman-Diaconis bin width to the number of bins you chose in Exercise 3.5.

4.5. Consider the failure data analyzed in the previous exercise. Construct four kernel density estimates for these data, using values of the smoothing parameter of $1, 4, 8$, and 12, respectively, and a Gaussian kernel.

Which of the estimates appears to give the best summarization of the data? In answering this question, keep in mind the fact that the data are necessarily non-negative.

4.6. Consider the scores on the Peabody Picture Vocabulary Test contained in the dataset “Peabody”.

(a) Construct a kernel density estimate for these data, using a Gaussian kernel. Consider four possible values for the smoothing parameter h, $2, 3, 4, 5$. Choose the best value of the smoothing parameter h subjectively and plot the estimate corresponding to your choice. Which value of h did you choose?

(b) For the density plot constructed in part (a), include a plot of the normal density function with mean μ and standard deviation σ, with the values of μ and σ taken to be the estimates based on the Peabody data.

Based on this plot, do the peabody scores appear for the children in the study appear to be normally distributed?