Approximation and Estimation of Functions
Using Series Expansions in Basis Functions
(Chapter 5)

Our objective is to represent a function of interest, \( f(x) \), over some domain \( D \), as a linear combination of “simpler” functions, \( q_0(x), q_1(x), \ldots \):

\[
  f(x) = \sum_{k=0}^{\infty} c_k q_k(x).
\]

There are various ways of constructing the \( q_k \) functions.

If they are developed through a linear operator on a function space, they are called *eigenfunctions*, and the corresponding \( c_k \) are called eigenvalues.
We choose a set \( \{q_k\} \) that spans some class of functions over the given domain \( D \).

A set of orthogonal basis functions is often the best choice because they have nice properties that facilitate computations and a large body of theory about their properties is available.
Series Expansions in Basis Functions

If the function to be estimated, $f(x)$, is continuous and integrable over a domain $D$, the orthonormality property allows us to determine the coefficients $c_k$ in the expansion:

$$c_k = \langle f, q_k \rangle.$$

The coefficients $\{c_k\}$ are called the Fourier coefficients of $f$ with respect to the orthonormal functions $\{q_k\}$.
Approximation with a Series Expansion

In applications, we approximate the function using a truncated orthogonal series, that is, at any point \( x \),

\[
    f(x) \approx \sum_{k=1}^{j} c_k q_k(x).
\]
Approximation with a Series Expansion

The error due to finite truncation at $j$ terms of the infinite series is the residual function $f - \sum_{k=1}^{j} c_k q_k$.

The mean squared error over the domain $D$ is the scaled, squared $L_2$ norm of the residual,

$$\frac{1}{d} \left\| f - \sum_{k=0}^{j} c_k q_k \right\|^2,$$

where $d$ is some measure of the domain $D$.

A very important property of Fourier coefficients is that they yield the minimum mean squared error for a given set of basis functions $\{q_i\}$; that is, for any other constants, $\{a_i\}$, and any $j$,

$$\left\| f - \sum_{k=0}^{j} c_k q_k \right\|^2 \leq \left\| f - \sum_{k=0}^{j} a_k q_k \right\|^2.$$
In applications of statistical data analysis, after forming the approximation, we then estimate the coefficients by identifying an appropriate probability density that is a factor of the function of interest, $f$.

Expected values can be estimated using observed or simulated values of the random variable and the approximation of the probability density function.

The basis functions are generally chosen to be easy to use in computations.

Common examples include the Fourier trigonometric functions $\sin(kt)$ and $\cos(kt)$ for $k = 1, 2, \ldots$, orthogonal polynomials such as Legendre, Hermite, and so on, splines, and wavelets.
Orthogonal Polynomials

The most useful type of basis function depends on the nature of the function being estimated.

The orthogonal polynomials are useful for a very wide range of functions.

In applications, we approximate the function using a truncated orthogonal series. At any point $x$

$$\tilde{f}(x) = \sum_{k=1}^{j} c_k q_k(x)$$

and

$$f(x) \approx \tilde{f}(x).$$
Standard Systems of Orthogonal Polynomials

The different systems are characterized by the one-dimensional intervals over which they are defined and by their weight functions.

The Legendre, Chebyshev, and Jacobi polynomials are defined over \([-1, 1]\) and hence can be scaled into any finite interval.

The weight function of the Jacobi polynomials is more general, but the Legendre and Chebyshev polynomials are simpler and so are often used.

The Laguerre polynomials are defined over the half line \([0, \infty)\).

The Hermite polynomials are defined over the reals, \((-\infty, \infty)\).
Approximation; Determining the Coefficients

As an example of the use of orthogonal polynomials to approximate a given function, consider the expansion of $f(x) = e^{-x}$ over the interval $[-1, 1]$.

We just compute the Fourier coefficients using the inner product.
Estimation of the Coefficients in an Orthogonal Expansion

We first decompose the function of interest to have a factor that is a probability density function, \( p \),

\[
    f(x) = g(x)p(x).
\]

This is the PDF decomposition. We use it extensively in computational statistics.

We have

\[
    c_k = \langle f, q_k \rangle
    = \int_D q_k(x)g(x)p(x)dx
    = E(q_k(X)g(X)),
\]

where \( X \) is a random variable whose probability density function is \( p \).
Estimation of the Coefficients

If we can obtain a random sample, $x_1, \ldots, x_n$, from the distribution with density $p$, the $c_k$ can be unbiasedly estimated by

$$\hat{c}_k = \frac{1}{n} \sum_{i=1}^{n} q_k(x_i) g(x_i).$$

The series estimator of the function for all $x$ therefore is

$$\hat{f}(x) = \frac{1}{n} \sum_{k=0}^{j} \sum_{i=1}^{n} q_k(x_i) g(x_i) q_k(x)$$

for some truncation point $j$.

The random sample, $x_1, \ldots, x_n$, may be an observed dataset, or it may be the output of a random number generator.
Splines

An important type of basis function is a spline.

Designed to be 0 over certain intervals, and to have smooth joins at the interval endpoints.

Interpolating splines.

Smoothing splines.
Another approach to function estimation and approximation is to use a filter or kernel function to provide local weighting of the observed data.

This approach ensures that at a given point the observations close to that point influence the estimate at the point more strongly than more distant observations.

A standard method in this approach is to convolve the observations with a unimodal function that decreases rapidly away from a central point.

This function is the filter or the kernel.

A kernel function has two arguments representing the two points in the convolution, but we typically use a single argument that represents the distance between the two points.
**Kernels**

Some examples of univariate kernel functions are

**uniform:** \( K_u(t) = 0.5 \), for \( |t| \leq 1 \),

**quadratic or Epanechnikov:** \( K_q(t) = 0.75(1 - t^2) \), for \( |t| \leq 1 \),

**normal:** \( K_n(t) = \frac{1}{\sqrt{2\pi}}e^{-t^2/2} \), for all \( t \),

Sometimes the compact kernels are better.
Kernel Methods

In kernel methods, the locality of influence is controlled by a window around the point of interest.

The choice of the size of the window is the most important issue in the use of kernel methods.

In practice, for a given choice of the size of the window, the argument of the kernel function is transformed to reflect the size.

The transformation is accomplished using a positive definite matrix, $V$, whose determinant measures the volume (size) of the window.
Estimation of a Function Using Kernel Methods

To estimate the function \( f \) at the point \( x \), we first do a PDF decomposition

\[
f(x) = g(x)p(x).
\]

For a given set of data, \( x_1, \ldots, x_n \), and a given scaling transformation matrix \( V \), the kernel estimator of the function at the point \( x \) is

\[
\hat{f}(x) = (n|V|)^{-1} \sum_{i=1}^{n} g(x_i)K \left( V^{-1}(x - x_i) \right).
\]
Kernel Methods; the Univariate Form

In the univariate case, the size of the window is just the width $h$.

The argument of the kernel is transformed to $s/h$, so the function that is convolved with the function of interest is $K(s/h)/h$.

The univariate kernel estimator is

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} g(x_i) K \left( \frac{x - x_i}{h} \right).$$
Pointwise Properties of Function Estimators

The statistical properties of an estimator of a function at a given point are analogous to the usual statistical properties of an estimator of a scalar parameter.

The statistical properties involve expectations or other properties of random variables.

The expectations are usually taken with respect to the (unknown) distribution of the underlying random variable.

Occasionally, we may explicitly indicate the distribution by writing, for example, $E_p(\cdot)$, where $p$ is the density of the random variable with respect to which the expectation is taken.
Bias

The bias of the estimator of a function value at the point $x$ is

$$\mathbb{E}(\hat{f}(x)) - f(x).$$

If this bias is zero, we would say that the estimator is unbiased at the point $x$.

If the estimator is unbiased at every point $x$ in the domain of $f$, we say that the estimator is pointwise unbiased.

Obviously, in order for $\hat{f}(\cdot)$ to be pointwise unbiased, it must be defined over the full domain of $f$. 
Variance

The variance of the estimator at the point $x$ is

$$V(\hat{f}(x)) = \mathbb{E} \left( (\hat{f}(x) - \mathbb{E}(\hat{f}(x)))^2 \right).$$

Estimators with small variance are generally more desirable, and an optimal estimator is often taken as the one with smallest variance among a class of unbiased estimators.
Mean Squared Error

The mean squared error, MSE, at the point $x$ is

$$\text{MSE}(\hat{f}(x)) = \mathbb{E}((\hat{f}(x) - f(x))^2).$$

The mean squared error is the sum of the variance and the square of the bias:

$$\text{MSE}(\hat{f}(x)) = \mathbb{E}((\hat{f}(x))^2 - 2\hat{f}(x)f(x) + (f(x))^2) = \text{Var}(\hat{f}(x)) + (\mathbb{E}(\hat{f}(x)) - f(x))^2.$$

Sometimes, the variance of an unbiased estimator is much greater than that of an estimator that is only slightly biased, so it is often appropriate to compare the mean squared error of the two estimators.

In some cases, as we will see, unbiased estimators do not exist, so rather than seek an unbiased estimator with a small variance, we seek an estimator with a small MSE.
Mean Absolute Error

The mean absolute error, MAE, at the point $x$ is similar to the MSE:

$$\text{MAE}(\hat{f}(x)) = \mathbb{E}(|\hat{f}(x) - f(x)|).$$

It is more difficult to do mathematical analysis of the MAE than it is for the MSE.

Furthermore, the MAE does not have a simple decomposition into other meaningful quantities similar to the MSE.
Consistency

Consistency of an estimator refers to the convergence of the expected value of the estimator to what is being estimated as the sample size increases without bound.

A point estimator $T_n$, based on a sample of size $n$, is consistent for $\theta$ if

$$\mathbb{E}(T_n) \to \theta \quad \text{as } n \to \infty.$$ 

The convergence is stochastic, of course, so there are various types of convergence that can be required for consistency.

The most common kind of convergence considered is weak convergence, or convergence in probability.
Consistency

In addition to the type of stochastic convergence, we may consider the convergence of various measures of the estimator.

In general, if $m$ is a function (usually a vector-valued function that is an elementwise norm), we may define consistency of an estimator $T_n$ in terms of $m$ if

$$E(m(T_n - \theta)) \to 0.$$
Consistency: Type of Convergence

For an estimator, we are often interested in weak convergence in mean square or weak convergence in quadratic mean, so the common definition of consistency of $T_n$ is

$$
\mathbb{E}((T_n - \theta)^T (T_n - \theta)) \to 0,
$$

where the type of convergence is convergence in probability.

Consistency defined by convergence in mean square is also called $L_2$ consistency.
Rate of Convergence

If convergence does occur, we are interested in the rate of convergence.

We define rate of convergence in terms of a function of \( n \), say \( r(n) \), such that

\[
E(m(T_n - \theta)) = O(r(n)).
\]

A common form of \( r(n) \) is \( n^\alpha \), where \( \alpha < 0 \).

For example, in the simple case of a univariate population with a finite mean \( \mu \) and finite second moment, use of the sample mean \( \bar{x} \) as the estimator \( T_n \), and use of \( m(z) = z^2 \), we have

\[
E(m(\bar{x} - \mu)) = E((\bar{x} - \mu)^2) = \text{MSE}(\bar{x}) = O(n^{-1}).
\]
Consistency at a Point

In the estimation of a function, we say that the estimator \( \hat{f} \) of the function \( f \) is \textit{pointwise consistent} if

\[
E(\hat{f}(x)) \to f(x)
\]

for every \( x \) the domain of \( f \). Just as in the estimation of a parameter, there are various kinds of pointwise consistency in the estimation of a function.

If the convergence is in probability, for example, we say that the estimator is weakly pointwise consistent.

We could also define other kinds of pointwise consistency in function estimation along the lines of other types of consistency.
Global Properties of Estimators of Functions

Often we are interested in some measure of the statistical properties of an estimator of a function over the full domain of the function.

The obvious way of defining statistical properties of an estimator of a function is to integrate the pointwise properties discussed in the previous section.

Statistical properties of a function, such as the bias of the function, are often defined in terms of a norm of the function.
Global Properties of Estimators of Functions

For comparing $\hat{f}(x)$ and $f(x)$, the $L_p$ norm of the error is

$$\left( \int_D |\hat{f}(x) - f(x)|^p \, dx \right)^{1/p},$$

where $D$ is the domain of $f$.

The integral may not exist, of course.

Clearly, the estimator $\hat{f}$ must also be defined over the same domain.
Global Properties of Estimators of Functions

Three useful measures are the $L_1$ norm, also called the integrated absolute error, or IAE,

$$\text{IAE}(\hat{f}) = \int_D \left| \hat{f}(x) - f(x) \right| \, dx,$$

the square of the $L_2$ norm, also called the integrated squared error, or ISE,

$$\text{ISE}(\hat{f}) = \int_D (\hat{f}(x) - f(x))^2 \, dx,$$

and the $L_\infty$ norm, the sup absolute error, or SAE,

$$\text{SAE}(\hat{f}) = \sup \left| \hat{f}(x) - f(x) \right|.$$
Global Properties of Estimators of Functions

The $L_1$ measure is invariant under monotone transformations of the coordinate axes, but the measure based on the $L_2$ norm is not.

The $L_\infty$ norm, or SAE, is the most often used measure in general function approximation.

In statistical applications, this measure applied to two cumulative distribution functions is the *Kolmogorov distance*. The measure is not so useful in comparing densities and is not often used in density estimation.
Global Properties of Estimators of Functions

Other measures of the difference in $\hat{f}$ and $f$ over the full range of $x$ are the Kullback-Leibler measure,

$$\int_D \hat{f}(x) \log \left( \frac{\hat{f}(x)}{f(x)} \right) \, dx,$$

and the Hellinger distance,

$$\left( \int_D (\hat{f}^{1/p}(x) - f^{1/p}(x))^p \, dx \right)^{1/p}.$$

For $p = 2$, the Hellinger distance is also called the Matusita distance.
Integrated Bias and Variance

We now want to develop global concepts of bias and variance for estimators of functions.

Bias and variance are statistical properties that involve expectations of random variables.

The obvious global measures of bias and variance are just the pointwise measures integrated over the domain.

In the case of the bias, of course, we must integrate the absolute value, otherwise points of negative bias could cancel out points of positive bias.

The estimator \( \hat{f} \) is pointwise unbiased if

\[
E(\hat{f}(x)) = f(x) \quad \text{for all } x \in \mathbb{R}^d.
\]
Because we are interested in the bias over the domain of the function, we define the \textit{integrated absolute bias} as
\[
\text{IAB}(\hat{f}) = \int_D |E(\hat{f}(x)) - f(x)| \, dx
\]
and the \textit{integrated squared bias} as
\[
\text{ISB}(\hat{f}) = \int_D (E(\hat{f}(x)) - f(x))^2 \, dx.
\]

If the estimator is unbiased, both the integrated absolute bias and integrated squared bias are 0.

This, of course, would mean that the estimator is pointwise unbiased almost everywhere.
Integrated Bias

Although it is not uncommon to have unbiased estimators of scalar parameters or even of vector parameters with a countable number of elements, it is not likely that an estimator of a function could be unbiased at almost all points in a dense domain.

(“Almost” means all except possibly a set with a probability measure of 0.)
Integrated Variance

The *integrated variance* is defined in a similar manner:

\[
IV(\hat{f}) = \int_D V(\hat{f}(x)) \, dx \\
= \int_D E((\hat{f}(x) - E(\hat{f}(x)))^2) \, dx.
\]
Integrated Mean Squared Error

Global unbiasedness is generally not to be expected.

An important measure for comparing estimators of functions is, therefore, based on the mean squared error.

The \textit{integrated mean squared error} is

\[
\text{IMSE}(\hat{f}) = \int_D \mathbb{E}((\hat{f}(x) - f(x))^2) \, dx
\]

\[
= \text{IV}(\hat{f}) + \text{ISB}(\hat{f}).
\]
Mean Integrated Squared Error

If the expectation integration can be interchanged with the outer integration in the expression above, we have

\[
\text{IMSE}(\hat{f}) = \mathbb{E}\left( \int_D (\hat{f}(x) - f(x))^2 \, dx \right)
= \text{MISE}(\hat{f}),
\]

the \textit{mean integrated squared error}.

We will assume that this interchange leaves the integrals unchanged, so we will use MISE and IMSE interchangeably.
**Integrated Mean Absolute Error**

Similarly, for the *integrated mean absolute error*, we have

\[
\text{IMAE}(\hat{f}) = \int_D \mathbb{E}(|\hat{f}(x) - f(x)|) \, dx
\]

\[
= \mathbb{E} \left( \int_D |\hat{f}(x) - f(x)| \, dx \right)
\]

\[
= \text{MIAE}(\hat{f}),
\]

the *mean integrated absolute error*. 
Mean SAE

The *mean sup absolute error*, or MSAE, is

$$\text{MSAE}(\hat{f}) = \int_D \mathbb{E}(\sup|\hat{f}(x) - f(x)|) \, dx.$$ 

This measure is not very useful unless the variation in the function $f$ is relatively small.

For example, if $f$ is a density function, $\hat{f}$ can be a “good” estimator, yet the MSAE may be quite large.

On the other hand, if $f$ is a cumulative distribution function (monotonically ranging from 0 to 1), the MSAE may be a good measure of how well the estimator performs.

As mentioned earlier, the SAE is the *Kolmogorov distance*.

The Kolmogorov distance (and, hence, the SAE and the MSAE) does poorly in measuring differences in the tails of the distribution.
Large-Sample Statistical Properties

The pointwise consistency properties are extended to the full function in the obvious way.

Consistency of the function estimator is defined in terms of

$$\int_D \mathbb{E}(m(\hat{f}(x) - f(x))) \, dx \to 0.$$
Large-Sample Statistical Properties

The estimator of the function is said to be mean square consistent or $L_2$ consistent if the MISE converges to 0; that is,

$$\int_D E((\hat{f}(x) - f(x))^2) \, dx \to 0.$$

If the convergence is weak, that is, if it is convergence in probability, we say that the function estimator is weakly consistent; if the convergence is strong, that is, if it is convergence almost surely or with probability 1, we say the function estimator is strongly consistent.
Large-Sample Statistical Properties

The estimator of the function is said to be $L_1$ consistent if the mean integrated absolute error (MIAE) converges to 0; that is,

$$\int_D E(|\hat{f}(x) - f(x)|) \, dx \to 0.$$  

As with the other kinds of consistency, the nature of the convergence in the definition may be expressed in the qualifiers “weak” or “strong”.

As we have mentioned above, the integrated absolute error is invariant under monotone transformations of the coordinate axes, but the $L_2$ measures are not.

As with most work in $L_1$, however, derivation of various properties of IAE or MIAE is more difficult than for analogous properties with respect to $L_2$ criteria.
Convergence

If the MISE converges to 0, we are interested in the rate of convergence.

To determine this, we seek an expression of MISE as a function of \( n \). We do this by a Taylor series expansion.

In general, if \( \hat{\theta} \) is an estimator of \( \theta \), the Taylor series for \( \text{ISE}(\hat{\theta}) \) about the true value is

\[
\text{ISE}(\hat{\theta}) = \sum_{k=0}^{\infty} \frac{1}{k!} (\hat{\theta} - \theta)^k \text{ISE}^{(k)}(\theta),
\]

where \( \text{ISE}^{(k)}(\theta) \) represents the \( k \)th derivative of \( \text{ISE} \) evaluated at \( \theta \).

Taking the expectation yields the MISE. The limit of the MISE as \( n \to \infty \) is the asymptotic mean integrated squared error, AMISE.

One of the most important properties of an estimator is the order of the AMISE.
Large-Sample Statistical Properties

In the case of an unbiased estimator, the first two terms in the Taylor series expansion are zero, and the AMISE is

\[ V(\hat{\theta}) \text{ISE}''(\theta) \]

to terms of second order.
Other Global Properties of Estimators of Functions

There are often other properties that we would like an estimator of a function to possess.
Weights or Densities

We may want the estimator to weight given functions in some particular way.

For example, if we know how the function to be estimated, $f$, weights a given function $r$, we may require that the estimate $\hat{f}$ weight the function $r$ in the same way; that is,

$$\int_{D} r(x)\hat{f}(x)dx = \int_{D} r(x)f(x)dx.$$
Range of the Function

We may want to restrict the minimum and maximum values of the estimator.

For example, because many functions of interest are nonnegative, we may want to require that the estimator be nonnegative.
Smoothness

We may want to restrict the variation in the function. This can be thought of as the “roughness” of the function. A reasonable measure of the variation is

\[
\int_D \left( f(x) - \int_D f(x) \, dx \right)^2 \, dx.
\]

If the integral \( \int_D f(x) \, dx \) is constrained to be some constant (such as 1 in the case that \( f(x) \) is a probability density), then the variation can be measured by the square of the L_2 norm,

\[
S(f) = \int_D (f(x))^2 \, dx.
\]
Smoothness

We may want to restrict the derivatives of the estimator or the smoothness of the estimator.

Another intuitive measure of the roughness of a twice-differentiable and integrable univariate function $f$ is the integral of the square of the second derivative:

$$\mathcal{R}(f) = \int_D (f''(x))^2 dx.$$ 

Often in function estimation, we may seek an estimator $\hat{f}$ such that its roughness (by some definition) is small.
Estimation of Probability Density Functions

Estimation of a probability density function is similar to the estimation of any other function.

Recall:

pointwise

... expectation, variance, etc.

global

... max or sup (bias, etc.)

... integrated (ISB, IMSE, etc)
Bona Fide Density Estimator

It seems reasonable that we require the density estimate to have the characteristic properties of a density:

• \( \hat{p}(y) \geq 0 \) for all \( y \);

• \( \int_{\mathbb{R}^d} \hat{p}(y) \, dy = 1 \).

A probability density estimator that is nonnegative and integrates to 1 is called a *bona fide* estimator.
Maximum Likelihood Estimation

Given a random sample, \( y_1, \ldots, y_n \), from a population with density \( p \). The likelihood functional is

\[
L(p; y_1, \ldots, y_n) = \prod_{i=1}^{n} p(y_i).
\]

The density \( p \) itself is a variable.

The *maximum likelihood method* of estimation obviously cannot be used directly because this functional is unbounded in \( p \).
Restricted Maximum Likelihood Estimation

Consider a finite dimensional class, such as the class of step functions that are bona fide density estimators. We assume that the sizes of the regions over which the step function is constant are greater than 0.

For a step function with $m$ regions having constant values, $c_1, \ldots, c_m$, the likelihood is

$$L(c_1, \ldots, c_m; y_1, \ldots, y_n) = \prod_{i=1}^{n} p(y_i) = \prod_{k=1}^{m} c_k^{n_k},$$

where $n_k$ is the number of data points in the $k^{th}$ region. For the step function to be a bona fide estimator, all $c_k$ must be nonnegative and finite.

A maximum therefore exists in the class of step functions that are bona fide estimators.
Restricted Maximum Likelihood Estimation, Continued

The restricted maximum likelihood estimator is

\[ \hat{p}(y) = \frac{n_k}{nv_k}, \quad \text{for } y \in \text{region } k, \]

\[ = 0, \quad \text{otherwise.} \]
Histogram Estimators

Let us assume finite support $D$, and construct a fixed partition of $D$ into a grid of $m$ nonoverlapping bins $T_k$. (We can arbitrarily assign bin boundaries to one or the other bin.) Let $v_k$ be the volume of the $k^{th}$ bin (in one dimension, $v_k$ is a length and in this simple case is often denoted $h_k$; in two dimensions, $v_k$ is an area, and so on).

The number of such bins we choose, and consequently their volumes, depends on the sample size $n$, so we sometimes indicate that dependence in the notation: $v_{n,k}$. For the sample $y_1, \ldots, y_n$, the histogram estimator of the probability density function is defined as

$$
\hat{p}_H(y) = \sum_{k=1}^{m} \frac{1}{v_k} \frac{\sum_{i=1}^{n} I_{T_k}(y_i)}{n} I_{T_k}(y), \quad \text{for } y \in D,
$$

$$
= 0, \quad \text{otherwise}.
$$

The histogram is the restricted maximum likelihood estimator discussed before.
Histogram Estimators

Letting $n_k$ be the number of sample values falling into $T_k$,

$$n_k = \sum_{i=1}^{n} I_{T_k}(y_i),$$

we have the simpler expression for the histogram over $D$,

$$\hat{p}_H(y) = \sum_{k=1}^{m} \frac{n_k}{nv_k} I_{T_k}(y).$$

This is a bona fide estimator:

$$\hat{p}_H(y) \geq 0$$

and

$$\int_{\mathbb{R}^d} \hat{p}_H(y) dy = \sum_{k=1}^{m} \frac{n_k}{nv_k}$$

$$= 1.$$
Some Properties of the Histogram Estimator

The histogram estimator, being a step function, is discontinuous at cell boundaries, and it is zero outside of a finite range.

An important advantage of the histogram estimator is its simplicity, both for computations and for analysis. In addition to its simplicity, as we have seen, it has two other desirable global properties:

- It is a bona fide density estimator.

- It is the unique maximum likelihood estimator confined to the subspace of functions of the form

\[ g(t) = c_k, \text{ for } t \in T_k, \]
\[ = 0, \text{ otherwise,} \]

and where \( g(t) \geq 0 \) and \( \int_{\bigcup_{k=1}^{m} T_k} g(t) \, dt = 1. \)
Pointwise and Binwise Properties

Properties of the histogram vary from bin to bin.

We see that the number in the $k^{th}$ bin, $n_k$, is a binomial random variable with parameters $n$ and $p_k$, where

$$p_k = \int_{T_k} p(t) \, dt$$

is the probability content of the $k^{th}$ bin.

The **expectation** of the histogram estimator at the point $y$ in bin $T_k$ is

$$E(\hat{p}_H(y)) = \frac{p_k}{v_k},$$

and the **variance** of the histogram at the point $y$ within the $k^{th}$ bin is

$$V(\hat{p}_H(y)) = V(n_k)/(nv_k)^2$$

$$= \frac{p_k(1 - p_k)}{nv_k^2}.$$
Pointwise Bias

The bias of the histogram at the point $y$ within the $k^{th}$ bin is

$$\frac{p_k}{v_k} - p(y).$$

Note that the bias at $y$ is a function of $y$, and it is different from bin to bin, even if the bins are of constant size.

The bias tends to decrease as the bin size decreases.
Pointwise Variance

The variance of the histogram at the point $y$ within the $k^{th}$ bin is

$$\text{Var}(\hat{p}_H(y)) = \frac{p_k(1 - p_k)}{nv_k^2}.$$

Notice that the variance decreases as the bin size increases. Note also that the variance is different from bin to bin. We can bound the variance:

$$\text{Var}(\hat{p}_H(y)) \leq \frac{p_k}{nv_k^2}.$$

By the mean-value theorem, we have $p_k = v_k p(\xi_k)$ for some $\xi_k \in T_k$, so we can write

$$\text{Var}(\hat{p}_H(y)) \leq \frac{p(\xi_k)}{nv_k}.$$
Tradeoffs

Notice the tradeoff between bias and variance:

as $h$ increases,

the variance decreases, but

the bound on the bias increases.