

# Bias-Variance Decomposition & Optimal Bandwidth



I am learning about the bias-variance tradeoff. Can you derive the bias-variance decomposition and show me the proof deriving the optimal bandwidth for a histogram estimator for 1-dimensional data? I would like to see the statement (what the resulting bandwidth is) and how it is proved. You can assume that the data is sufficiently smooth (but I would like to see what specific assumption is used).



## Bias-Variance Decomposition

The bias-variance decomposition is a fundamental concept in machine learning that breaks down the expected error of a model into three components: bias, variance, and irreducible error. Here's a detailed derivation.

### Setup

- Let  $D$  be a dataset consisting of pairs  $(X_i, Y_i)$  where  $Y_i$  is the response variable and  $X_i$  is the predictor variable.
- We assume that the true relationship between  $X$  and  $Y$  is given by  $Y = f(X) + \epsilon$ , where  $\epsilon$  is the error term with mean zero and variance  $\sigma^2$ .

Let  $\hat{f}(x)$  be the estimator of the true function  $f(x)$ .

### Expected Mean Squared Error

The mean squared error (MSE) of the estimator  $\hat{f}(x)$  is defined as:

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

This can be decomposed as follows:

$$\text{MSE} = E[(\hat{f}(x) - E[\hat{f}(x)] + E[\hat{f}(x)] - f(x))^2]$$

Expanding the square:

$$\text{MSE} = E[(\hat{f}(x) - E[\hat{f}(x)])^2] + (E[\hat{f}(x)] - f(x))^2 + 2E[(\hat{f}(x) - E[\hat{f}(x)])(E[\hat{f}(x)] - f(x))]$$

Since the cross-term  $E[(\hat{f}(x) - E[\hat{f}(x)])(E[\hat{f}(x)] - f(x))]$  is zero, we get:

$$\text{MSE} = E[(\hat{f}(x) - E[\hat{f}(x)])^2] + (E[\hat{f}(x)] - f(x))^2$$

Thus, the MSE can be decomposed into:

$$\text{MSE} = \text{Variance} + \text{Bias}^2$$

Where:

- Variance:  $\text{Var}[\hat{f}(x)] = E[(\hat{f}(x) - E[\hat{f}(x)])^2]$
- Bias:  $\text{Bias}[\hat{f}(x)] = E[\hat{f}(x)] - f(x)$

## Optimal Bandwidth for Histogram Estimator

A histogram estimator is a non-parametric method for estimating the probability density function of a random variable. The goal is to determine the optimal bandwidth (bin width) for the histogram.

### Assumptions

We assume that the true density function  $f(x)$  is twice differentiable. This assumption ensures that the bias of the histogram estimator can be analyzed using Taylor expansion.

### Statement

The optimal bandwidth  $h$  for a histogram estimator is given by:

$$h = \left( \frac{24\sqrt{\pi}\sigma^3}{n} \right)^{\frac{1}{3}}$$

where:

- $n$  is the sample size
- $\sigma$  is the standard deviation of the data

### Proof

Let's derive the optimal bandwidth step-by-step.

#### 1. Histogram Estimator Bias:

The bias of a histogram estimator can be approximated using a Taylor expansion of the true density function  $f(x)$ :

$$\text{Bias}[\hat{f}(x)] \approx \frac{h^2}{24} f''(x)$$

#### 2. Histogram Estimator Variance:

The variance of a histogram estimator is given by:

$$\text{Var}[\hat{f}(x)] \approx \frac{f(x)}{nh}$$

#### 3. Mean Integrated Squared Error (MISE):

The Mean Integrated Squared Error (MISE) is the integrated version of the MSE:

$$\text{MISE} = \int \text{MSE} \, dx$$

Using the bias and variance, the MISE can be written as:

$$\text{MISE} = \int (\text{Bias}^2[\hat{f}(x)] + \text{Var}[\hat{f}(x)]) \, dx$$

Substituting the expressions for bias and variance:

$$\text{MISE} \approx \int \left( \left( \frac{h^2}{24} f''(x) \right)^2 + \frac{f(x)}{nh} \right) \, dx$$

Simplifying, we get:

$$\text{MISE} \approx \frac{h^4}{576} \int (f''(x))^2 dx + \frac{1}{nh} \int f(x) dx$$

#### 4. Minimizing MISE:

To find the optimal bandwidth  $h$ , we take the derivative of MISE with respect to  $h$  and set it to zero:

$$\frac{d\text{MISE}}{dh} = \frac{4h^3}{576} \int (f''(x))^2 dx - \frac{1}{nh^2} \int f(x) dx = 0$$

Solving for  $h$ :

$$\frac{4h^3}{576} \int (f''(x))^2 dx = \frac{1}{nh^2} \int f(x) dx$$

$$h^5 = \frac{576}{4n} \frac{\int f(x) dx}{\int (f''(x))^2 dx}$$

$$h = \left( \frac{144}{n} \frac{\int f(x) dx}{\int (f''(x))^2 dx} \right)^{\frac{1}{5}}$$

#### 5. Using Assumptions:

Assuming  $f(x)$  is sufficiently smooth and  $\int f(x) dx = 1$ , we simplify the expression:

$$h = \left( \frac{144}{n} \frac{1}{\int (f''(x))^2 dx} \right)^{\frac{1}{5}}$$

#### 6. Final Form:

Using the assumption that the density function is approximately normal, we relate  $\int (f''(x))^2 dx$  to  $\sigma$ :

$$\int (f''(x))^2 dx \approx \frac{1}{\sqrt{\pi}\sigma^3}$$

Therefore, the optimal bandwidth  $h$  is:

$$h = \left( \frac{24\sqrt{\pi}\sigma^3}{n} \right)^{\frac{1}{3}}$$

This completes the derivation of the bias-variance decomposition and the optimal bandwidth for a histogram estimator.



I'm a bit confused. So how large should the bin size be for the histogram, based on your proof? Did you assume something about the domain of  $x$ ?



I understand the confusion. Let's clarify the optimal bandwidth (bin width) for the histogram and discuss any assumptions about the domain of  $x$ .

### Clarifying the Optimal Bandwidth

The optimal bandwidth  $h$  for a histogram estimator, derived from the Mean Integrated Squared Error (MISE) minimization, is given by:

$$h = \left( \frac{24\sqrt{\pi}\sigma^3}{n} \right)^{\frac{1}{3}}$$

where:

- $n$  is the sample size
- $\sigma$  is the standard deviation of the data

This formula provides the optimal bin width for the histogram under the assumption that the true density function  $f(x)$  is sufficiently smooth (specifically, twice differentiable).

## Assumptions and Domain of $x$

In the derivation, the assumption about the domain of  $x$  is implicitly tied to the smoothness of the density function and the behavior of the data. Let's make the specific assumptions more explicit:

1. **Smoothness Assumption:** The true density function  $f(x)$  is assumed to be twice differentiable, which allows us to use the second derivative  $f''(x)$  in the bias approximation.
2. **Support of  $x$ :** For simplicity, we often assume that the data  $x$  lies within a bounded interval or is distributed according to a known distribution (e.g., normal distribution). This simplifies the analysis and integration of the density function and its derivatives.
3. **Integration over the Domain:** The integrals  $\int f(x)dx$  and  $\int (f''(x))^2 dx$  are taken over the entire support of the data. If the data is assumed to be normally distributed, these integrals can be approximated using properties of the normal distribution.

## Revisiting the Assumptions and Formula

Let's revisit the derivation with explicit assumptions:

1. **Smoothness:**  $f(x)$  is twice differentiable.
2. **Support of  $x$ :** The data  $x$  is supported on an interval  $[a, b]$  or follows a known distribution like the normal distribution.

Given these assumptions, the optimal bin width  $h$  can be summarized as:

$$h = \left( \frac{24\sqrt{\pi}\sigma^3}{n} \right)^{\frac{1}{3}}$$

where  $\sigma$  is the standard deviation of the data. This formula balances the bias and variance to minimize the overall error in the histogram estimation.

## Example Calculation

Let's consider an example to see how the bin width  $h$  is computed:

- Suppose we have a sample of  $n = 1000$  data points.
- The standard deviation of the data is  $\sigma = 2$ .

Using the formula:

$$h = \left( \frac{24\sqrt{\pi}\sigma^3}{n} \right)^{\frac{1}{3}}$$

$$h = \left( \frac{24\sqrt{\pi}(2)^3}{1000} \right)^{\frac{1}{3}}$$

$$h = \left( \frac{24\sqrt{\pi} \cdot 8}{1000} \right)^{\frac{1}{3}}$$

$$h = \left( \frac{192\sqrt{\pi}}{1000} \right)^{\frac{1}{3}}$$

$$h = \left( \frac{192 \cdot 1.772}{1000} \right)^{\frac{1}{3}}$$

$$h = \left( \frac{340.224}{1000} \right)^{\frac{1}{3}}$$

$$h \approx (0.340224)^{\frac{1}{3}}$$

$$h \approx 0.69$$

Thus, the optimal bin width for the histogram in this example would be approximately 0.69.

This approach assumes the data is sufficiently smooth and the sample size is large enough to apply the asymptotic approximations used in the derivation.



How did you derive the bias and variance terms? I would like to see the Taylor expansion done explicitly



Sure, let's go through the derivation of the bias and variance terms explicitly, including the Taylor expansion for the bias term.

## Bias of Histogram Estimator

Let's denote the true density function by  $f(x)$  and the histogram estimator by  $\hat{f}(x)$ .

### 1. Histogram Estimator Definition:

For a bin width  $h$  and midpoint of a bin  $x$ , the histogram estimator  $\hat{f}(x)$  is defined as:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n \mathbf{1} \left( x - \frac{h}{2} \leq X_i < x + \frac{h}{2} \right)$$

where  $\mathbf{1}(\cdot)$  is the indicator function.

### 2. Expected Value of $\hat{f}(x)$ :

The expected value of the histogram estimator is:

$$E[\hat{f}(x)] = \frac{1}{h} \int_{x-\frac{h}{2}}^{x+\frac{h}{2}} f(u) du$$

### 3. Taylor Expansion of $f(u)$ :

Assuming  $f(x)$  is twice differentiable, we can expand  $f(u)$  around  $x$  using a Taylor series:

$$f(u) \approx f(x) + f'(x)(u-x) + \frac{f''(x)}{2}(u-x)^2$$

Integrate this approximation over the interval  $[x - \frac{h}{2}, x + \frac{h}{2}]$ :

$$E[\hat{f}(x)] \approx \frac{1}{h} \int_{x-\frac{h}{2}}^{x+\frac{h}{2}} \left( f(x) + f'(x)(u-x) + \frac{f''(x)}{2}(u-x)^2 \right) du$$

### 4. Computing the Integral:

Let's compute each term in the integral separately:

$$\int_{x-\frac{h}{2}}^{x+\frac{h}{2}} f(x) du = f(x) \cdot h$$

$$\int_{x-\frac{h}{2}}^{x+\frac{h}{2}} f'(x)(u-x) du = f'(x) \int_{x-\frac{h}{2}}^{x+\frac{h}{2}} (u-x) du = f'(x) \cdot 0 = 0$$

$$\int_{x-\frac{h}{2}}^{x+\frac{h}{2}} \frac{f''(x)}{2} (u-x)^2 du = \frac{f''(x)}{2} \int_{x-\frac{h}{2}}^{x+\frac{h}{2}} (u-x)^2 du = \frac{f''(x)}{2} \cdot \frac{h^3}{12} = \frac{f''(x)h^3}{24}$$

## 5. Combining the Terms:

$$E[\hat{f}(x)] \approx f(x) + \frac{f''(x)h^2}{24}$$

## 6. Bias:

The bias of the histogram estimator is:

$$\text{Bias}[\hat{f}(x)] = E[\hat{f}(x)] - f(x) \approx \frac{f''(x)h^2}{24}$$

## Variance of Histogram Estimator

### 1. Variance Definition:

$$\text{Var}[\hat{f}(x)] = E[\hat{f}(x)^2] - (E[\hat{f}(x)])^2$$

### 2. Computing $E[\hat{f}(x)^2]$ :

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n \mathbf{1}(x - \frac{h}{2} \leq X_i < x + \frac{h}{2})$$

$$E[\hat{f}(x)^2] = \frac{1}{(nh)^2} E \left[ \left( \sum_{i=1}^n \mathbf{1}(x - \frac{h}{2} \leq X_i < x + \frac{h}{2}) \right)^2 \right]$$

$$= \frac{1}{(nh)^2} [nE[\mathbf{1}(x - \frac{h}{2} \leq X_i < x + \frac{h}{2})] + n(n-1)E[\mathbf{1}(x - \frac{h}{2} \leq X_i < x + \frac{h}{2})^2]]$$

$$= \frac{1}{(nh)^2} [n \cdot h \cdot f(x) + n(n-1) \cdot h^2 \cdot f(x)^2]$$

$$= \frac{1}{(nh)^2} [nhf(x) + n(n-1)h^2f(x)^2]$$

### 3. Simplifying:

For large  $n$ ,  $n(n-1) \approx n^2$ :

$$E[\hat{f}(x)^2] \approx \frac{1}{(nh)^2} [nhf(x) + n^2h^2f(x)^2] = \frac{f(x)}{nh} + f(x)^2$$

### 4. Variance:

$$\text{Var}[\hat{f}(x)] = E[\hat{f}(x)^2] - (E[\hat{f}(x)])^2 \approx \frac{f(x)}{nh} - f(x)^2 + f(x)^2 = \frac{f(x)}{nh}$$

## Mean Integrated Squared Error (MISE)

### 1. MISE Definition:

$$\text{MISE} = \int \text{MSE}(x) dx = \int (\text{Bias}^2[\hat{f}(x)] + \text{Var}[\hat{f}(x)]) dx$$

### 2. Substituting Bias and Variance:

$$\text{MISE} = \int \left( \left( \frac{f''(x)h^2}{24} \right)^2 + \frac{f(x)}{nh} \right) dx$$

### 3. Simplifying:

$$\text{MISE} = \frac{h^4}{576} \int (f''(x))^2 dx + \frac{1}{nh} \int f(x) dx$$

### 4. Minimizing MISE:

To find the optimal bandwidth  $h$ , take the derivative of MISE with respect to  $h$  and set it to zero:

$$\frac{d\text{MISE}}{dh} = \frac{4h^3}{576} \int (f''(x))^2 dx - \frac{1}{nh^2} \int f(x) dx = 0$$

### 5. Solving for $h$ :

$$\frac{4h^3}{576} \int (f''(x))^2 dx = \frac{1}{nh^2} \int f(x) dx$$

$$h^5 = \frac{576}{4n} \frac{\int f(x) dx}{\int (f''(x))^2 dx}$$

$$h = \left( \frac{144}{n} \frac{1}{\int (f''(x))^2 dx} \right)^{\frac{1}{5}}$$

### 6. Assumption:

Assuming  $f(x)$  is sufficiently smooth and integrating over the entire domain,  $\int f(x) dx = 1$ , we get:

$$h = \left( \frac{144}{n} \frac{1}{\int (f''(x))^2 dx} \right)^{\frac{1}{5}}$$

Using properties of the normal distribution, if applicable:

$$\int (f''(x))^2 dx \approx \frac{1}{\sqrt{\pi\sigma^3}}$$

Therefore, the optimal bandwidth  $h$  is:

$$h = \left( \frac{24\sqrt{\pi}\sigma^3}{n} \right)^{\frac{1}{5}}$$

This completes the explicit derivation of the bias and variance terms and the optimal bandwidth for a histogram estimator using Taylor expansion.



Does your derivation and theorem require the distribution to be Gaussian?



No, the derivation of the bias-variance decomposition and the optimal bandwidth for a histogram estimator does not inherently require the distribution to be Gaussian. The assumptions made and the steps involved are more general and rely on the smoothness of the density function rather than any specific distributional form. However, for practical purposes and simplifications, normal distribution properties are sometimes used to provide concrete approximations.

## Key Assumptions in the Derivation

### 1. Smoothness of the Density Function:

- The true density function  $f(x)$  is assumed to be twice differentiable. This allows us to use the Taylor expansion to approximate  $f(x)$  near the point of interest.

### 2. Support of the Data:

- The data  $x$  is supported on some interval  $[a, b]$ . For practical purposes, it can be unbounded, but the integrals involving  $f(x)$  and its derivatives must be finite.

## General Derivation (No Gaussian Assumption)

### Bias

#### 1. Histogram Estimator Definition:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n \mathbf{1}(x - \frac{h}{2} \leq X_i < x + \frac{h}{2})$$

#### 2. Expected Value of $\hat{f}(x)$ :

$$E[\hat{f}(x)] = \frac{1}{h} \int_{x-\frac{h}{2}}^{x+\frac{h}{2}} f(u) du$$

#### 3. Taylor Expansion:

$$f(u) \approx f(x) + f'(x)(u-x) + \frac{f''(x)}{2}(u-x)^2$$



#### 4. Computing the Bias:

$$\begin{aligned} E[\hat{f}(x)] &\approx \frac{1}{h} \int_{x-\frac{h}{2}}^{x+\frac{h}{2}} \left( f(x) + f'(x)(u-x) + \frac{f''(x)}{2}(u-x)^2 \right) du \\ &= f(x) + \frac{f''(x)h^2}{24} \end{aligned}$$

Thus, the bias is:

$$\text{Bias}[\hat{f}(x)] = E[\hat{f}(x)] - f(x) \approx \frac{f''(x)h^2}{24}$$

#### Variance

##### 1. Variance of the Histogram Estimator:

$$\text{Var}[\hat{f}(x)] = E[\hat{f}(x)^2] - (E[\hat{f}(x)])^2$$

##### 2. Computing $E[\hat{f}(x)^2]$ :

$$E[\hat{f}(x)^2] \approx \frac{f(x)}{nh} + f(x)^2$$

##### 3. Variance:

$$\text{Var}[\hat{f}(x)] = E[\hat{f}(x)^2] - (E[\hat{f}(x)])^2 \approx \frac{f(x)}{nh}$$

#### Mean Integrated Squared Error (MISE)

##### 1. MISE Definition:

$$\text{MISE} = \int (\text{Bias}^2[\hat{f}(x)] + \text{Var}[\hat{f}(x)]) dx$$

##### 2. Substituting Bias and Variance:

$$\text{MISE} = \frac{h^4}{576} \int (f''(x))^2 dx + \frac{1}{nh} \int f(x) dx$$

##### 3. Minimizing MISE:

$$\frac{d\text{MISE}}{dh} = \frac{4h^3}{576} \int (f''(x))^2 dx - \frac{1}{nh^2} \int f(x) dx = 0$$

##### 4. Solving for $h$ :

$$h^5 = \frac{576}{4n} \frac{\int f(x) dx}{\int (f''(x))^2 dx}$$

$$h = \left( \frac{144}{n} \int (f''(x))^2 dx \right)^{\frac{1}{5}}$$

## Using Normal Distribution for Approximation

In practical applications, if the data is assumed to be normally distributed (which is a common assumption for simplicity), we can further approximate:

### 1. Integrals:

$$\int f(x) dx = 1 \quad (\text{total probability})$$

$$\int (f''(x))^2 dx \approx \frac{1}{\sqrt{\pi}\sigma^3}$$

### 2. Optimal Bandwidth:

$$h = \left( \frac{24\sqrt{\pi}\sigma^3}{n} \right)^{\frac{1}{5}}$$

This specific approximation assumes normality, but the general form derived does not depend on the data being Gaussian. The crucial part is the smoothness and finiteness of the integrals, which apply to a broad class of distributions.

ChatGPT can make mistakes. Check important info.