Elements of Nonparametric Statistics

Nicholas Henderson

2019-12-25

Contents

Preface

This is the very first part of the book.

Chapter 1

Introduction

1.1 What is Nonparametric Statistics?

What is Parametric Statistics?

- Parametric models refer to probability distributions that can be fully described by a fixed number of parameters that do not grow with the sample size.
- Typical examples include
  - Gaussian
  - Poisson
  - Exponential
  - Beta
- Could also refer to a regression setting where the mean function is described by a fixed number of parameters.

What is Nonparametric Statistics?

- Difficult to give a concise, all-encompassing definition, but nonparametric statistics generally refers to statistical methods where there is not a clear parametric component.
- The uses of nonparametric methods in several common statistical contexts are described in Sections 1.3 - 1.7.

1.2 Outline of Course

This course is roughly divided into the following 5 categories.

1. **Nonparametric Testing**
   - Rank-based Tests
   - Permutation Tests
2. **Estimation of Basic Nonparametric Quantities**
   - The Empirical Distribution Function
   - Density Estimation
3. **Nonparametric Confidence Intervals**
   - Bootstrap
   - Jacknife
4. **Nonparametric Regression Part I (Smoothing Methods)**
   - Kernel Methods
   - Splines
   - Local Regression
5. **Nonparametric Regression Part II (Machine Learning Methods)**
   - Decision Trees/CART
   - Ensemble Methods

1.3 Example 1: Nonparametric vs. Parametric Two-Sample Testing

Suppose we have data from two groups. For example, outcomes from two different treatments.

- **Group 1 outcomes**: $X_1, \ldots, X_n$ an i.i.d (independent and identically distributed) sample from distribution function F_X. That is,

$$F_X(t) = P(X_i \leq t)$$

- **Group 2 outcomes**: $Y_1, \ldots, Y_m$ an i.i.d. sample from distribution function F_Y.

- To test the impact of a new treatment, we usually want to test whether or not F_X differs from F_Y in some way. This can be stated in hypothesis testing language as

$$\begin{aligned} H_0 &: F_X = F_Y (\text{ populations are the same}) \\ H_A &: F_X \neq F_Y (\text{ populations are different}) \end{aligned}$$

Parametric Tests

- A common parametric test for () is the t-test. The t-test assumes that

$$F_X = \text{Normal}(\mu_x, \sigma^2) \quad \text{and} \quad F_Y = \text{Normal}(\mu_y, \sigma^2) \tag{1.1}$$

- Under this parametric assumption, the hypothesis test () reduces to

$$H_0 : \mu_x = \mu_y \quad \text{vs.} \quad H_A : \mu_x \neq \mu_y \tag{1.2}$$

- The standard t-statistic (with a pooled estimate of σ^2) is the following

$$T = \frac{\bar{X} - \bar{Y}}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}}, \tag{1.3}$$

  where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\bar{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i$ are the group-specific sample means and s_p^2 is the pooled estimate of σ^2

$$s_p^2 = \frac{1}{m+n-2}\Big\{ \sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{m}(Y_i - \bar{Y})^2 \Big\} \tag{1.4}$$

- The t-test is based on the **null distribution** of T - the distribution of T under the null hypothesis.

*Under the assumption of normality, the null distribution of T is a t distribution with $n + m - 2$ degrees of freedom.

- Put graph here
- Notice that this null distribution depends on the parametric assumption that both $F_X = \text{Normal}(\mu_x, \sigma^2)$ and $F_Y = \text{Normal}(\mu_y, \sigma^2)$. (Mention CLT argument here)
- Moreover, we used the parametric assumption in the formulation of the hypothesis test itself because we assumed that any difference between F_X and F_Y would be fully described by difference in μ_x and μ_y.

- Two-sample nonparametric tests are meant to be "distribution-free". This means that the null distribution of the test statistic does not depend on any parametric assumptions about the two populations F_X and F_Y.
- Also, the hypotheses tests themselves do not rely on any parametric assumptions.
- For example,

1.4 Example 2: Nonparametric Estimation

- Suppose we have n observations $(X_1, \ldots, X_n)$ which are assumed to be i.i.d. (independent and identically distributed). The distribution function of X_i is F_X.
- Suppose we are interested in estimating F_X.
- In a **parametric** approach to estimating F_X, we would assume the distribution of X_i belongs to some parametric family of distributions. For example, $X_i \sim \text{Normal}(\mu, \sigma^2)$, $X_i \sim \text{Exponential}(\lambda)$, or $X_i \sim \text{Beta}(\alpha, \beta)$.
- If we assume that $X_i \sim \text{Normal}(\mu, \sigma^2)$, we only need to estimate 2 parameters to fully describe the distribution of X_i, and the number of parameters does not depend on the sample size.

1.5 Example 3: Confidence Intervals

1.6 Example 4: Nonparametric Regression with a Single Covariate

1.7 Example 5: Nonparametric Regression

Chapter 2

Working with R

Before we can start exploring data in R, there are some key concepts to understand first:

1. What are R and RStudio?
2. How do I code in R?
3. What are R packages?

If you are already familiar with these concepts, feel free to skip to Section **??** below introducing some of the datasets we will explore in depth in this book. Much of this chapter is based on two sources which you should feel free to use as references if you are looking for additional details:

Part I

Nonparametric Testing

Chapter 3

Rank and Sign Statistics

3.1 Introduction

Start with t-test example, difference in means is sufficient for superiority

Give example of type of tests we are interested in.

Why ranks and why nonparametric testing?

(reduce influence of outliers)

3.2 Ranks

3.2.1 Definition

- Suppose we have n observations $X_1, \ldots, X_n$. The **rank** of the i^{th} observation R_i is defined as

$$R_i = R(X_i) = \sum_{j=1}^{n} I(X_j \leq X_i) \tag{3.1}$$

  where

$$I(X_j \leq X_i) = \begin{cases} 1 & \text{if } X_j \leq X_i \\ 0 & \text{if } X_j > X_i \end{cases} \tag{3.2}$$

- The largest observation has a rank of n.
- The smallest observation has a rank of 1 (if there are no ties).

```
x <- c(3, 7, 1, 12, 6)  ## 5 observations
rank(x)
```

```
## [1] 2 4 1 5 3
```

- In the definition of ranks shown in (3.1), tied observations receive their maximum possible rank.
- For example, suppose that $(X_1, X_2, X_3, X_4) = (0, 1, 1, 2)$. In this case, one could argue whether both observations 2 and 3 should be ranked 2^{nd} or 3^{rd} while observations 1 and 4 should unambiguously receive ranks of 1 and 4 respectively.
- Under definition @rank(eq:rankdef), both observations 2 and 3 receive a rank of 3.
- In **R**, such handling of ties is done using the **ties.method = "max"** argument

```
x <- c(0, 1, 1, 2)
rank(x, ties.method="max")
```

```
## [1] 1 3 3 4
```

- The default in **R** is to replace the ranks of tied observations with their "average" rank

```
x <- c(0, 1, 1, 2)
rank(x)
```

```
## [1] 1.0 2.5 2.5 4.0
```

3.2.2 Properties of Ranks

Suppose $(X_1, \ldots, X_n)$ is random sample from a continuous distribution F (so that the probability of ties is zero). Then, the following properties hold for the associated ranks $R_1, \ldots, R_n$.

- Each R_i follows a discrete uniform distribution

$$P(R_i = j) = 1/n, \quad \text{for any } j = 1, \ldots, n. \tag{3.3}$$

- The expectation of R_i is

$$E(R_i) = \sum_{j=1}^{n} jP(R_i = j) = \frac{1}{n}\sum_{j=1}^{n} j = \frac{(n+1)}{2} \tag{3.4}$$

- The variance of R_i is

$$\text{Var}(R_i) = E(R_i^2) - E(R_i)^2 = \frac{1}{n}\sum_{j=1}^{n} j^2 - \left(\frac{n+1}{2}\right)^2 = \frac{n^2-1}{12} \tag{3.5}$$

- The random variables $R_1, \ldots, R_n$ are **not** independent (why?). However, the vector $\mathbf{R}_n = (R_1, \ldots, R_n)$ is uniformly distributed on the set of $n!$ permutations of $(1, 2, \ldots, n)$.

3.3 Two-Sample Tests

Wilcoxon statistic and Wilcoxon signed-rank statistic (what is the difference between these two?)

3.3.1 The Wilcoxon Rank Sum Test

3.3.1.1 Purpose

- The Wilcoxon Rank Sum (WRS) test (sometimes referred to as the Wilcoxon-Mann-Whitney test) is a two-sample test.
- The WRS test is used to test whether or not observations from one group tend to be larger (or smaller) than observations from the other group.
- Suppose we have observations from two groups: $X_1, \ldots, X_n \sim F_X$ and $Y_1, \ldots, Y_m \sim F_Y$. The WRS test will test the hypothesis

  H_0 : $F_X = F_Y$ versus
  H_A : Observations from F_X tend to be larger than observations from F_Y

- What is meant by "tend to be larger"?
- Two common ways of stating the alternative hypothesis for the WRS include

  $$\begin{aligned} H_0 &: \quad F_X = F_Y \quad \text{versus} \\ H_A &: \quad F_X \text{ is stochastically larger than } F_Y \end{aligned}$$

  or

  $$\begin{aligned} H_0 &: \quad F_X = F_Y \quad \text{versus} \\ H_A &: \quad F_X(t) = F_Y(t - \Delta), \Delta > 0. \end{aligned} \tag{3.6}$$

- A distribution function F_X is said to be stochastically larger than F_Y if $F_X(t) \geq F_Y(t)$ for all t with $F_X(t) > F_Y(t)$ for at least one value of t.
- Note that the "shift alternative" implies stochastic dominance.
- Why do we need to specify an alternative?

3.3.1.2 Definition of Test Statistic

- The WRS test statistic is based on computing the sum of ranks (ranks based on the pooled sample) in one group.

- If observations from this group tend to be larger, their average rank should exceed the average rank from the other group.
-
- Give exercise, compute p-values for Wilcoxon test where we have two populations. both are Normally distributed with mean zero but different variances.

3.4 One Sample Tests

3.4.1 The Sign Test

- Suppose we have observations $W_1, \ldots, W_n$ which arise from the following model
$$W_i = \theta + \varepsilon_i,$$
where ε_i are iid random variables each with distribution function F that is assumed to have a median of zero.

3.4.2 The Signed-Rank Wilcoxon Test

3.5 Comparisons with Parametric Tests

3.6 Thinking about Rank statistics more generally

Chapter 4

Rank Tests for Multiple Groups

In Subsection

Chapter 5

Permutation Tests (Next Chapter Should be U-statistics)

Permutation tests are ...

Chapter 6

U-Statistics

6.1 Examples

- A wide range of well-known statistics can also be represented as U-statistics.
- Sample Mean, Variance, Signed-Rank Statistic, Gini Mean Difference

6.2 Mann-Whitney Statistic

6.3 Kendall's tau

6.4 Distance Covariance

Part II

Nonparametric Estimation

Chapter 7

The Empirical Distribution Function

7.1 Empirical Distribution Functions

7.1.1 Definition and Basic Properties

- Every random variable has a cumulative distribution function (cdf).
- The cdf of a random variable X is defined as

$$F(t) = P(X \leq t) \tag{7.1}$$

-
- The empirical distribution function or empirical cumulative distribution function (ecdf) estimates $F(t)$ by just finding the proportion of observations which are less than or equal to t.
- For i.i.d. random variables $X_1, \ldots, X_n$, the empirical distribution function is defined as

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq t)$$

- Note that the empirical distribution function can be computed for any type of data without making any assumptions about the distribution from which the data arose.
- The only assumption we are making is that $X_1, \ldots, X_n$ constitute an i.i.d. sample from some common distribution function F.

7.1.2 Confidence intervals for Fhat

7.2 The Empirical Distribution Function in R

7.3 The empirical distribution functino and statistical functionals

Chapter 8

Density Estimation

8.1 Introduction

- In this section, we focus on methods for estimating a **probability density function** (pdf) $f(x)$.

- For a continuous random variable X, areas under the probability density function are probabilities

$$P(a < X < b) = \int_a^b f(x)dx$$

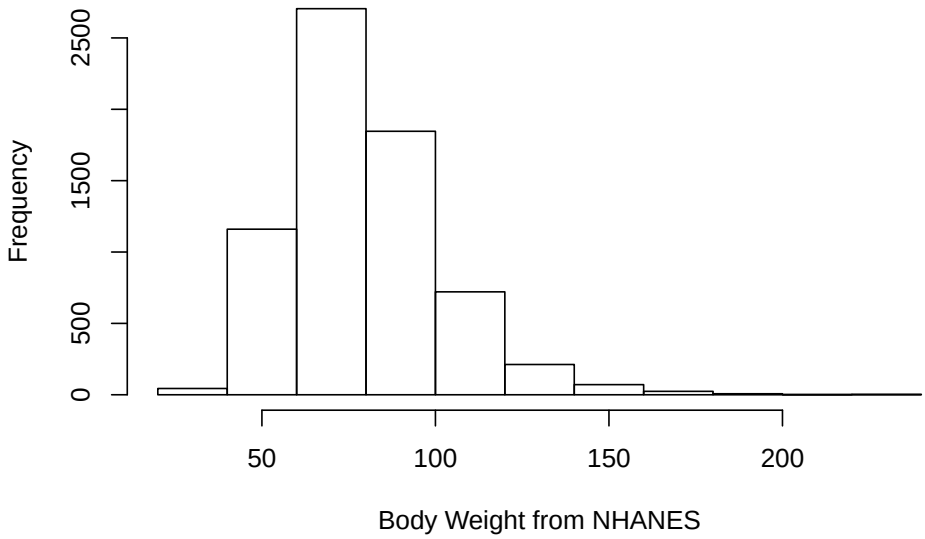
  and $f(x)$ is related to the distribution function via $f(x) = F'(x)$.

- With parametric approaches to density estimation, you only need to estimate a couple of parameters as these parameters completely determine the form of $f(x)$.

- For example, with a Gaussian distribution you only need to estimate μ and σ^2 to draw the appropriate bell curve.

- In a nonparametric approach, you assume that your observations $X_1, \ldots, X_n$ are an independent sample from a distribution having pdf $f(x)$, but otherwise you make few assumptions about the particular form of $f(x)$.

8.2 Histograms



8.2.1 Definition

Histograms are one of the oldest ways to estimate a pdf.

To construct a histogram, you first need to define a series of "bins": $B_1, \ldots, B_D$. Each bin is a left-closed interval which is often assumed to have the form $B_k = [x_0 + (k-1)h, x_0 + kh)$:

$$\begin{aligned} B_1 &= [x_0, x_0 + h) \\ B_2 &= [x_0 + h, x_0 + 2h) \\ &\vdots \\ B_D &= [x_0 + (D-1)h, x_0 + Dh) \end{aligned}$$

- x_0 - the origin
- h_n - bin width

For each bin, you first need to count the number of observations which fall into that bin

$$\begin{aligned} n_k &= \#\text{ of observations falling into the } k^{th} \text{ bin} \\ &= \sum_{i=1}^{n} I(x_0 + (k-1)h_n \leq X_i < x_0 + kh_n) \end{aligned} \tag{8.1}$$

The histogram estimate of the density at a point x in the k^{th} bin is then defined

as

$$\hat{f}(x) = \frac{n_k}{nh_n} \tag{8.2}$$

Note: What is often shown in histogram plots are the actual bin counts n_k rather than the values of $\hat{f}(x)$.

- To see the motivation for the histogram estimate, notice that if we choose a relatively small value $h > 0$

$$P(x < X_i < x + h) = \int_x^{x+h} f(x)dx \approx hf(x) \tag{8.3}$$

- The expected value of $\hat{f}(x)$ is

$$E\{\hat{f}(x)\} ==\approx f(x) \tag{8.4}$$

8.2.2 Histograms in R

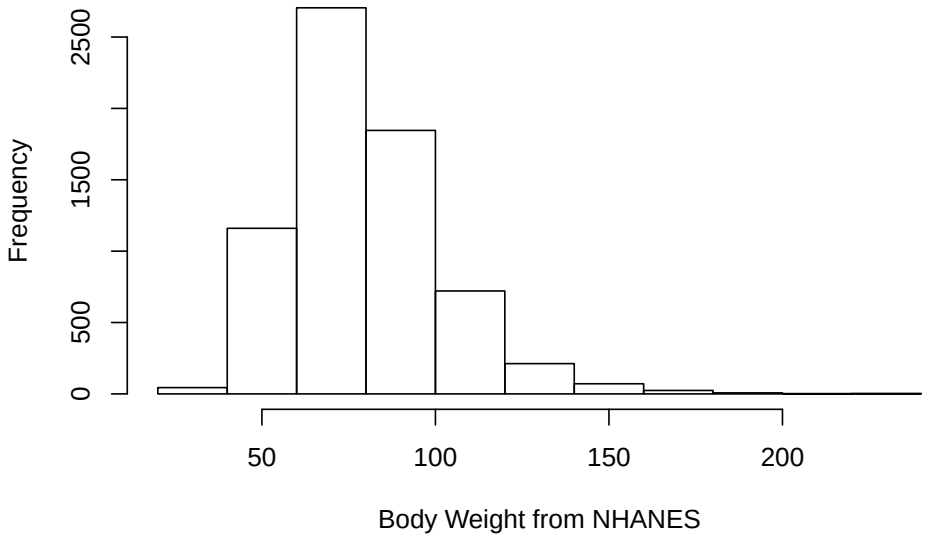
In **R**, use the `hist` function

```
hist(x, breaks, probability, plot, ...)
```

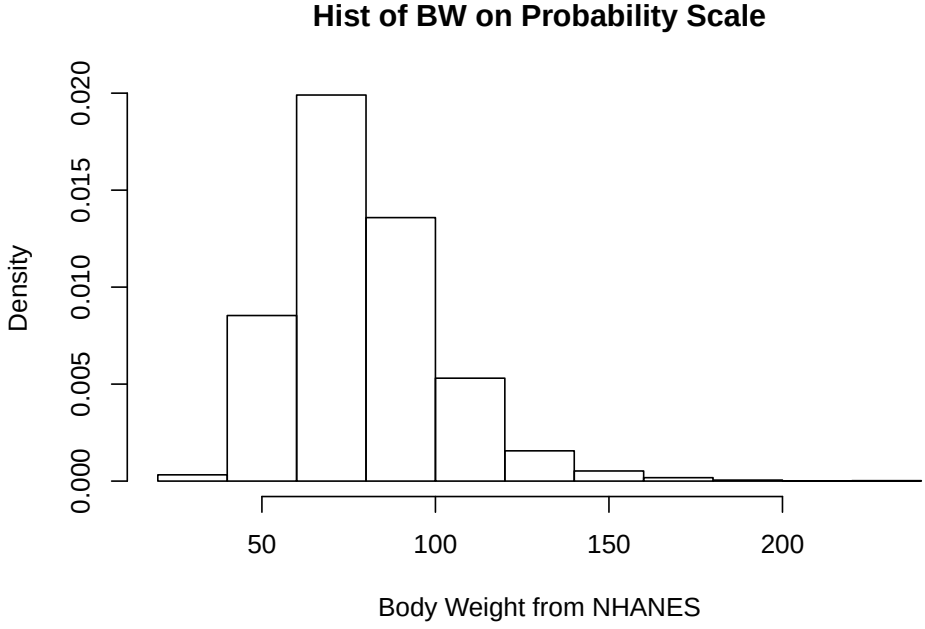
- The **breaks** argument
  - Default is "Sturges". This is a method for finding the binwidth.
  - Can be a name giving the name of an algorithm for computing bin-width (e.g., "Scott" and "FD").
  - Can also be a single number. This gives the number of bins used.
  - Could also be a ..
- The **probability** argument
- The **plot** argument

Note: The default for R, is to use right-closed intervals $(a, b]$. This can be changed using the **right** argument of the **hist** function.

```
## Use a real dataset here
bodywt.hist <- hist(nhgh$wt, main="", xlab="Body Weight from NHANES")
```




```
## Use a real dataset here
bodywt.hist2 <- hist(nhgh$wt, main="Hist of BW on Probability Scale", xlab="Body Weight
```



In addition to generating a histogram plot, the histogram function also returns useful stuff.

```
names(bodywt.hist)
```

```
## [1] "breaks"   "counts"   "density"  "mids"     "xname"    "equidist"
```

- **breaks**
  –
- **counts**
- **mids**
- **density**

```
bodywt.hist$breaks
```

```
##  [1]  20  40  60  80 100 120 140 160 180 200 220 240
```

```
bodywt.hist$counts
```

```
##  [1]   44 1160 2705 1846  721  212   71   24    7    2    3
```

```
binwidth <- bodywt.hist$breaks[2] - bodywt.hist$breaks[1]
bodywt.hist$density
```

```
##  [1] 3.237675e-04 8.535688e-03 1.990434e-02 1.358352e-02 5.305372e-03
##  [6] 1.559971e-03 5.224430e-04 1.766004e-04 5.150846e-05 1.471670e-05
## [11] 2.207506e-05
```

```
bodywt.hist$counts/(length(nhgh$wt)*binwidth)
```

```
##  [1] 3.237675e-04 8.535688e-03 1.990434e-02 1.358352e-02 5.305372e-03
##  [6] 1.559971e-03 5.224430e-04 1.766004e-04 5.150846e-05 1.471670e-05
## [11] 2.207506e-05
```

8.2.3 Performance of the Histogram Estimate

8.2.3.1 Bias/Variance Decomposition

- It is common to evaluate the performance of a density estimator through its **mean-squared error** (MSE).
- In general, MSE is a function of bias and variance

$$MSE = Bias^2 + Variance \tag{8.5}$$

- We will first look at the mean-squared error of $\hat{f}(x)$ at a single point x

$$\begin{aligned} \text{MSE}\{\hat{f}(x)\} &= E\Big(\{\hat{f}(x) - f(x)\}^2\Big) \\ &= \underbrace{\Big(E\{\hat{f}(x)\} - f(x)\Big)^2}_{\text{Bias Squared}} + \underbrace{\text{Var}\{\hat{f}(x)\}}_{Variance} \end{aligned}$$

- In general, as the bin-width h_n increases, the histogram estimate has less variation but becomes more biased.

8.2.3.2 Bias and Variance of the Histogram Estimate

- Recall that, for a histogram estimate, we have D_n bins where the k^{th} bin takes the form
$$B_k = [x_0 + (k-1)h_n, x_0 + kh_n)$$
- For a point $x \in B_k$, that "belongs" to the k^{th} bin, the histogram density estimate is
$$\hat{f}(x) = \frac{n_k}{nh_n}, \quad \text{where } n_k = \text{ number of observations falling into bin } B_k \tag{8.6}$$
- To better examine what happens as n changes, we will define the function $A_{h_n}(x)$ as the function which returns the index of the interval to which x belongs.
- For example, if we had three bins $B_1 = [0, 1/3)$, $B_2 = [1/3, 2/3)$, $B_3 = [2/3, 1)$ and $x = 1/2$, then $A_{h_n}(x) = 2$.
- So, we can also write the histogram density estimate as
$$\hat{f}(x) = \frac{n_{A_{h_n}(x)}}{nh_n} \tag{8.7}$$

- Note that $n_{A_{h_n}(x)}$ is a binomial random variable with n trials and success probability $p_{h_n}(x)$ (why?)
$$n_{A_{h_n}(x)} \sim \text{Binomial}\{n, p_{h_n}(x)\}$$
- The success probability $p_{h_n}(x)$ is defined as
$$p_{h_n}(x) = P\Big\{X_i \text{ falls into bin } A_{h_n}(x)\Big\} = \int_{x_0+(A_{h_n}(x)-1)h_n}^{x_0+A_{h_n}(x)h_n} f(t)dt. \tag{8.8}$$

- Using what is known about the Binomial distribution (i.e., $E(n_{A_{h_n}(x)}) = np_{h_n}(x)$ and $\text{Var}(n_{A_{h_n}(x)}) = np_{h_n}(x)\{1 - p_{h_n}(x)\}$), we can express the bias and variance of $\hat{f}(x)$ as
$$\begin{aligned} \text{Bias}\{\hat{f}(x)\} &= E\{\hat{f}(x)\} - f(x) \\ &= \frac{1}{nh_n} E(n_{A_{h_n}(x)}) - f(x) \\ &= \frac{p_{h_n}(x)}{h_n} - f(x) \end{aligned}$$
and
$$\text{Var}\{\hat{f}(x)\} = \frac{1}{n^2 h_n^2} \text{Var}(n_{A_{h_n}(x)}) = \frac{p_{h_n}(x)\{1 - p_{h_n}(x)\}}{nh_n^2} \tag{8.9}$$

- Using the approximation $f(t) \approx f(x) + f'(x)(t-x)$ for t close to x, we have that

$$\frac{p_{h_n}(x)}{h_n} = \frac{1}{h_n}\int_{x_0+(A_{h_n}(x)-1)h_n}^{x_0+A_{h_n}(x)h_n} f(t)dt \approx f(x)+f'(x)\{x-x_0-(A_{h_n}(x)-1)h_n\} \tag{8.10}$$

- So, the bias of the histogram density estimate $\hat{f}(x)$ is

$$\text{Bias}\{\hat{f}(x)\} \approx f'(x)\{x-(x_0+(A_{h_n}(x)-1)h_n)\} \tag{8.11}$$

- [[Double-check this bias formula and check with Scott]]
- Choosing a very small bin width will result in a small bias because the left endpoint of the bin $x_0 + (A_{h_n}(x)-1)h_n$ will always be very close to x.

- Now, turning to the variance of the histogram estimate

$$\text{Var}\{\hat{f}(x)\} = \frac{p_{h_n}(x)}{nh_n^2}\{1-p_{h_n}(x)\} \approx \frac{f(x)+f'(x)\{x-x_0-(A_{h_n}(x)-1)h_n\}}{nh_n}\{1-p_{h_n}(x)\} \approx \frac{f(x)}{nh_n} \tag{8.12}$$

- For a more detailed description of the above approximation see Scott.
- Note that large bin widths will reduce variance.

8.2.3.3 Point-wise Mean Squared Error

- Recalling (), the approximate mean-squared error of the histogram density estimate at a particular point x is given by

$$\begin{aligned} \text{MSE}\{\hat{f}(x)\} &= E\Big(\{\hat{f}(x)-f(x)\}^2\Big) \\ &= \Big(\text{Bias}\{\hat{f}(x)\}\Big)^2 + \text{Var}\{\hat{f}(x)\} \\ &\approx [f'(x)]^2\{x-(x_0+(A_{h_n}(x)-1)h_n)\}^2 + \frac{f(x)}{nh_n} \end{aligned} \tag{8.13}$$

- For any approach to bin width selection, we should have $h_n \longrightarrow 0$ and $nh_n \longrightarrow \infty$.
- This MSE approximation depends on a particular choice of x.
- Difficult to use () as a criterion for selecting the bandwidth because this could vary depending on your choice of x.

8.2.3.4 Integrated Mean Squared Error and Optimal Histogram Bin Width

- Using mean integrated squared error (MISE) allows us to find an optimal bin width that does not depend on a particular choice of x.

- The MISE is defined as

$$\begin{aligned} MISE\{\hat{f}(x)\} &= E\Big\{\int_{-\infty}^{\infty}\{\hat{f}(x)-f(x)\}^2dx\Big\} \\ &= \int_{-\infty}^{\infty}\text{MSE}\{\hat{f}(x)\}dx \end{aligned} \tag{8.14}$$

Using our previously derived approximation for the MSE, we have

$$\begin{aligned} MISE\{\hat{f}(x)\} &\approx \int x[f'(x)]^2 - x_0\int[f'(x)]^2dx + (A_{h_n}(x)-1)h_n)\}^2 + \frac{1}{nh_n}\int f(x)dx \\ &= \end{aligned} \tag{8.15}$$

- To select the optimal bin width, we minimize the MISE as a function of h_n.
- Minimizing (), as a function of h_n yields the following formula for the optimal bin width

$$h_n^{opt} = \Big(\frac{6}{n\int_{-\infty}^{\infty}[f'(x)]^2dx}\Big)^{1/3} = Cn^{-1/3}$$

- Notice that $h_n^{opt} \longrightarrow 0$ and $nh_n^{opt} \longrightarrow \infty$ as $n \longrightarrow \infty$.
- Notice also that the optimal bin width depends on the unknown quantity $\int_{-\infty}^{\infty}[f'(x)]^2dx$.

8.2.4 Choosing the Histogram Bin Width

- We will mention three rules for selecting the bin width of a histogram.
  - Scott rule: (based on the optimal bin width formula)
  - Friedman and Diaconis rule (also based on the optimal bin width formula)
  - Sturges rule: (based on …)

- Both Scott and the FD rule are based on the optimal bin width formula ().
- The main problem with this formula is the presence of $\int_{-\infty}^{\infty}[f'(x)]^2dx$.
- **Solution:** See what this quantity looks like if we assume that $f(x)$ corresponds to a $N(\mu,\sigma^2)$ density.
- With this assumption,

$$h_n^{opt} = 3.5\sigma n^{-1/3} \tag{8.16}$$

- Scott rule: use $\hat{\sigma} = 2$

8.3 Kernel Density Estimation

8.3.1 Histograms and a "Naive" Density Estimate

8.3.2 Kernels, Bandwidth, and Smooth Density Estimation

8.3.3 Bias and Variance of Kernel Density Estimates

8.3.4 Bandwidth Selection

8.4 Kernel Density Estimation in Practice

Part III

Uncertainty Measures

Chapter 9

The Bootstrap

9.1 Introduction

The jackknife and bootstrap are nonparametric procedures for finding standard errors and constructing confidence intervals.

Why use the jackknife or bootstrap?

1. To compute …
2. To find …
3. When you have no idea how to compute reasonable standard errors.

Chapter 10

The Jacknife

Definition: Confidence Interval

10.1 Bootstrapping

Part IV

Nonparametric Regression: Part I

Chapter 11

Kernel Regression

11.1 Introduction

11.1.1 An Example

11.1.2 Linear Smoothers and Naive Nonparametric Estimates

11.2 The Nadaraya-Watson estimator

11.3 Local Linear and Polynomial Regression

Chapter 12

Splines and Penalized Regression

12.1 Introduction

12.2 Spline Basis Functions

12.3 Smoothing Splines/Penalized Regression

12.3.1 Selection of Smoothing Parameter

Part V

Nonparametric Regression: Part II

Chapter 13

Decision Trees and CART