Chapter 4. Some Elementary Statistical Inferences

Note. The ETSU *Graduate Catalog* describes Mathematical Statistics 1 (STAT 4047/5047) as including probability distributions, random variables, distributions, and the Central Limit Theorem. Chapters 1, 2, 3, and 5 set up the mathematical part of Mathematical Statistics; Chapter 4 introduces the statistics part. Hogg, McKean, and Craig mention in the Preface that "the instructor would have the option of interchanging the order of Chapters 4 and 5," and we follow that plan here.

Section 4.1. Sampling and Statistics

Note. In this section, we present the vocabulary of statistics and illustrate these ideas with examples.

Note. In statistical problems, we consider random variable X and desire to find the probability density (or mass) function. We may have no information about the function, or we may know the type of distribution (such as knowing we have a Poisson distribution but not knowing λ). We take a sample $X_1, X_2, \ldots, X_n$ from the population; the actual values in the sample (called "realizations") are often denoted $x_1, x_2, \ldots, x_n$.

Definition 4.1.1. If the random variables $X_1, X_2, \ldots, X_n$ are independent and identically distributed ("iid") then these random variables constitute a *random sample* of size n from the common distribution.

Definition 4.1.2. Let $X_1, X_2, \ldots, X_n$ denote a sample on a random variable X. Let $T = T(X_1, X_2, \ldots, X_n)$ be a function of the sample. Then T is a *statistic*. Once the sample is drawn with the realizations $x_1, x_2, \ldots, x_n$, $t = T(x_1, x_2, \ldots, x_n)$ is the *realization* of the sample.

Definition. If $X_1, X_2, \ldots, X_n$ is a random sample on a random variable X with a probability density function $f(x; \theta)$ (or mass function $p(x; \theta)$ in the discrete case) where $\theta \in \Omega$ for some given set Ω, then the statistic T is an *estimator* of θ (or a "point estimator") of θ. The realization t or T is an *estimate* of θ

Definition 4.1.3. Let $X_1, X_2, \ldots, X_n$ denote a sample of a random variable X with probability density function $f(x; \theta)$, where $\theta \in \Omega$. Let $T = T(X_1, X_2, \ldots, X_n)$ be a statistic. Then T is an *unbiased* estimator of θ is $E(T) = \theta$.

Definition. With $x_1, x_2, \ldots, x_n$ as the realization of a sample of random variable X with probability function $f(x; \theta)$, where $\theta \in \Omega$, the function

$$L(\theta) = L(\theta; x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \theta)$$

is the *likelihood function* of the random sample. A value of θ for which $L(\theta)$ is maximum, if the value if unique, is the *maximum likelihood estimator* of θ, denoted $\hat{\theta}$.

Note. Think of $f(x_n;\theta)$ as the probability of X_i taking on the value x_i given θ (so this conditional probability is a function of θ). Then the probability of the realized sample values $x_1, x_2, \ldots, x_n$ is the likelihood function $L(\theta) = f(x_i;\theta)$. So the maximum likelihood estimator $\hat{\theta}$ is the value of θ that maximizes the probability of the observed sample. In practice, it may be easier to maximize $\ell(\theta) = \log(L(\theta))$. We start by setting $\ell'(\theta) = \dfrac{d\,\ell(\theta)}{d\theta} = 0$. If θ is a vector of parameters then we need to consider the system of equations $\dfrac{\partial\,\ell(\theta)}{\partial\theta} = 0$ where "θ" ranges over the components of the vector. Notice that Hogg, McKean, and Craig use the partial notation, even when θ is not a vector.

Example 4.1.1. Suppose the common probability density function of the random sample $X_1, X_2, \ldots, X_n$ is the $\Gamma(1,\theta)$ density function $f(x) = \theta^{-1}\exp(-x/\theta)$ with support $0 < x < \infty$. This gamma distribution is called the *exponential distribution.* The logarithm of the likelihood functions is

$$\ell(\theta) = \log\left(\prod_{i=1}^{n}\theta^{-1}e^{-x_i/\theta}\right) = \log\left(\theta^{-n}\prod_{i=1}^{n}e^{-x_i/\theta}\right)$$

$$= -n\log\theta + \sum_{i=1}^{n}(-x_i/\theta) = -n\log\theta - \theta^{-1}\left(\sum_{i=1}^{n}x_i\right).$$

Now

$$\frac{d}{d\theta}[\ell(\theta)] = \frac{-n}{\theta} - (0\theta^{-2})\sum_{i=1}^{n}x_i = -n\theta^{-1} + \theta^{-2}\sum_{i=1}^{n}x_i.$$

Setting $\dfrac{d}{d\theta}[\ell(\theta)] = 0$ gives the critical value $-n\theta{-}1 + \theta^{-2}\sum_{i=1}^{n}x_i = 0$ or $-n\theta + \sum_{i=1}^{n}x_i = 0$ or $\theta = \frac{1}{n}\sum_{i=1}^{n}x_i = \overline{x}$. Now

$$\frac{d^2}{d\theta^2}[\ell(\theta)] = -n(-\theta^{-2}) + (-2\theta^{-3})\sum_{i=1}^{n}x_i = n\theta^{-2} - 2\theta^{-3}(n\overline{x}) = n\theta^{-3}(\theta - 2\overline{x})$$

and

$$\ell''(\theta)|_{\theta=\overline{w}} = n(\overline{x})^{-1}((\overline{x}) - 2\overline{x}) = -n(\overline{x})^{-2} < 0,$$

so that ℓ has a local maximum at $\theta = \overline{x}$ by the Second Derivative Test. Since $\theta = \overline{x}$ is the only critical value of ℓ, then it must be an absolute maximum of ℓ (notice $\theta > 0$ since it is a parameter of a Γ-distribution). So the statistic $\hat{\theta} = \overline{X}$ is the maximum likelihood estimator of θ. Now $E(X) = (1)(\theta) = \theta$ (see Note 3.3.C), so $\hat{\theta}$ is an unbiased estimator of θ.

Example 4.1.2. Consider a Bernoulli experiment and let random variable X be 1 ot 0, depending on the outcome of the experiment as a success or a failure, respectively. Let θ, $0 < \theta < 1$, denote the probability of a success. The probability mass function is (by definition of the Bernoulli distribution) $p(x;\theta) = \theta^x(1-\theta)^{1-x}$ where $x \in \{0,1\}$. If $X_1, X_2, \ldots, X_n$ is a random sample on X with realization $x_1, x_2, \ldots, x_n$ then the likelihood function is

$$L(\theta) = \prod_{i=1}^{n} p(x_i;\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{(\sum_{i=1}^n x_i)}(1-\theta)^{(n-\sum_{i=1}^n x_i)} \text{ where } x_i \in \{0,1\}.$$

So

$$\ell(\theta) = \log(L(\theta)) = \left(\sum_{i=1}^{n} x_i\right)\log\theta + \left(x - \sum_{i=1}^{n} x_i\right)\log(1-\theta) \text{ where } x_i \in \{0,1\}.$$

Now

$$\frac{d}{d\theta}[\ell(\theta)] = \frac{\sum_{i=1}^n x_i}{\theta} - \frac{n - \sum_{i=1}^n x_i}{1-\theta}.$$

Setting $\dfrac{d}{d\theta}[\ell(\theta)] = 0$ gives $(1-\theta)\sum_{i=1}^n x_i - \theta\left(n - \sum_{i=1}^n x_i\right) = 0$ or $\sum_{i=1}^n x_i - \theta n = 0$ or $\theta - \frac{1}{n}\sum_{i=1}^n x_i = \overline{x}$. Now

$$\frac{d^2}{d\theta^2}[\ell(\theta)] = \frac{-\sum_{i=1}^n x_i}{\theta^2} - \frac{n - \sum_{i=1}^n x_i}{(1-\theta)2}$$

$$= -\left(\sum_{i=1}^{n} x_i\right) \frac{(1-\theta^2)+\theta^2}{\theta^2(1-\theta^2)} = -\sum_{i=1}^{n} x_i \frac{1}{\theta^2(1-\theta^2)}$$

and

$$\ell''(\theta)|_{\theta=\overline{x}} = -(n\overline{x}) \frac{1}{(\overline{x})^2(1-(\overline{x})^2)} < 0$$

since $0 < \theta = \overline{x} < 1$. So ℓ has a local maximum at $\theta = \overline{x}$ by the Second Derivative Test. Since $\theta = \overline{x}$ is the only critical value of ℓ, then it must give an absolute maximum likelihood estimator of θ. Now $E(X) = (1)(\theta) = \theta$ (see Note 3.1.A; here we have $n = 1$ and $p = \theta$), so $\hat{\theta}$ is an unbiased estimator of θ.

Example 4.1.3. Let X have a $N(\mu, \sigma^2)$ distribution with probability density function

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \text{ for } -\infty < x < \infty.$$

If $X_1, X_2, \ldots, X_n$ is a random sample on X with realization $x_1, x_2, \ldots, x_n$ then the logarithm of the likelihood function in terms of μ and σ (and so a function of the vector (μ, σ)) is

$$\ell(\mu,\sigma) = \log\left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-1}{2}\left(\frac{x_1-\mu}{\sigma}\right)^2\right) \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-1}{2}\left(\frac{x_2-\mu}{\sigma}\right)^2\right)\right.$$

$$\left.\cdots \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-1}{2}\left(\frac{x_n-\mu}{\sigma}\right)^2\right)\right) = -n\log(\sqrt{2\pi}\sigma) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2.$$

Now we consider the two first partials

$$\frac{\partial}{\partial\mu}[\ell(\mu,\sigma)] = \frac{-1}{2}\sum_{i=1}^{n} 2\left(\frac{x_i-\mu}{\sigma}\right)\left(\frac{-1}{\sigma}\right) = \frac{1}{\sigma^2}(n\overline{x} - n\mu) = \frac{n}{\sigma^2}(\overline{x}-\mu),$$

$$\frac{\partial}{\partial\sigma}[\ell(\mu,\sigma)] = -n\left(\frac{1}{\sqrt{2\pi}\sigma}\right)(\sqrt{2\pi}) - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2(-2\sigma^{-3}) = \frac{-n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Setting both first partials equal to 0 gives

$$\frac{n}{\sigma^2}(\overline{x}-\mu) = 0 \text{ and } \frac{-n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

$$\text{or } n(\overline{x}-\mu) = 0 \text{ and } -n\sigma^2 + \sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

$$\text{or } \mu = \overline{x} \text{ and } -n\sigma^2 + \sum_{i=1}^{n}(x_i = \overline{x})^2 = 0$$

$$\text{or } \mu = \overline{x} \text{ and } \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2.$$

Since $\dfrac{\partial \ell}{\partial \mu} = \dfrac{n}{\sigma^2}(\overline{x}-\mu)$, then $\dfrac{\partial^2 \ell}{\partial \mu^2} = \dfrac{-n}{\sigma^2}$ and $\dfrac{\partial^2 \ell}{\partial \mu\,\partial \sigma} = \dfrac{-2n}{\sigma^3}(\overline{x}-\mu)$. Since $\dfrac{\partial \ell}{\partial \sigma} = \dfrac{-n}{\sigma} + \dfrac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2$ then $\dfrac{\partial^2 \ell}{\partial \sigma^2} = \dfrac{n}{\sigma^2} - \dfrac{3}{\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2$. Notice that

$$\frac{\partial^2 \ell}{\partial \sigma\,\partial \mu} = \frac{1}{\sigma^3}\sum_{i=1}^{n}(-2(x_i-\mu)) = \frac{-2n}{\sigma^3}\sum_{i=1}^{n}\left(\frac{x_i}{n}-\frac{\mu}{n}\right) = \frac{-2n}{\sigma^3}(\overline{x}-\mu) = \frac{\partial^2 \ell}{\partial \mu\,\partial \sigma},$$

as expected. So at the critical value $(\mu,\sigma) = \left(\overline{x}, \left(\dfrac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x}^2\right)^{1/2}\right)$ we have

$$\left.\frac{\partial^2 \ell}{\partial \mu^2}\right|_{\sigma^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2} = \left.\frac{-n}{\sigma^2}\right|_{\sigma^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2} = \frac{-n}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2} < 0$$

and

$$\left.\left(\left(\frac{\partial^2 \ell}{\partial \mu^2}\right)\left(\frac{\partial^2 \ell}{\partial \sigma^2}\right) - \left(\frac{\partial^2 \ell}{\partial \mu\,\partial \sigma}\right)^2\right)\right|_{(\mu,\sigma)=\left(\overline{x},\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\right)^{1/2}\right)}$$

$$= \left.\left(\left(\frac{-n}{\sigma^2}\right)\left(\frac{n}{\sigma^2} - \frac{3}{\sigma^4}\sum_{i=1}^{n}(x_i=\mu)^2\right) - \left(\frac{-2n}{\sigma^3}(\overline{x}-\mu)\right)^2\right)\right|_{(\mu,\sigma)=\left(\overline{x},\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\right)^{1/2}\right)}$$

$$= \left.\left(\frac{-n^2}{\sigma^4} + \frac{3n}{\sigma^6}\sum_{i=1}^{n}(x_i-\mu)^2\right)\right|_{(\mu,\sigma)=\left(\overline{x},\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\right)^{1/2}\right)}$$

$$= \frac{-n^2}{\left(\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2\right)^2} + \frac{3n}{\left(\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2\right)^3} \sum_{i=1}^n (x_i - \overline{x})^2$$

$$= \frac{-n^4}{\left(\sum_{i=1}^n (x_i - \overline{x})^2\right)^2} + \frac{3n^4}{\left(\sum_{i=1}^n (x_i - \overline{x})^2\right)^2} = \frac{2n^4}{\left(\sum_{i=1}^n (x_i - \overline{x})^2\right)^2} > 0.$$

So by the Second Derivative Test for Local Extreme Values (see Theorem 11 in my online notes for Calculus 3 [MATH 2110] on Section 14.7. Extreme Values and Saddle Points), ℓ has a local maximum at $(\mu, \sigma) = \left(\overline{x}, \left(\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2\right)^{1/2}\right)$. Since ℓ has only one critical value, then it must give an absolute maximum of ℓ. So we have the statistics $\hat{\mu} = \overline{X}$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$. By Theorem 2.8.A, $E[\overline{X}] = \mu$ so that $\overline{X}$ is an unbiased estimator of μ. However, with

$$S^2 = \frac{\left(\sum_{i=1}^n X_i^2\right) - n\overline{X}^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$$

we have by Theorem 2.8.A that $E[S^2] = \sigma^2$. Since E is linear (by Theorem 1.8.2), then

$$E[\hat{\sigma}^2] = E\left[\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2\right] = E\left[\frac{n}{n-1}S^2\right] = \frac{n}{n-1}ES^2] = \frac{n}{n-1}\sigma^2 \neq \sigma^2.$$

So $\hat{\sigma}^2$ is a biased estimator of σ^2. However, for n "large" we have $\frac{n}{n-1}\sigma^2 \approx \sigma^2$. Hogg, McKean, and Craig state: "In practice, however, S^2 is the preferred estimator for σ^2." See page 229. Presumably this is because S^2 is *unbiased* (i.e., $E[S^2] = \sigma^2$).

Note. We now consider a random sample $X_1, X_2, \ldots, X_n$ on a random variable X with a cumulative distribution function $F(x)$. We use realized samples to create histograms, which estimate the probability density (mass) function $f(x)$ ($p(x)$). We do not assume any form of the distribution (that is, we do not assume a particular function or "parametric" form). For this reason, the histogram is called a *nonparametric* estimator. In Chapter 10, we consider "Nonparametric and Robust Statistics."

Example. Let X be s discrete random variable with probability mass function $p(x)$. Let $X_1, X_2, \ldots, X_n$ be a random sample on X. We consider two settings; first we suppose that the sample space of X is finite, say $\mathcal{D} = \{a_1, a_2, \ldots, a_m\}$. For $j = 1, 2, \ldots, m$ define

$$I_j(X_i) = \begin{cases} 1 & \text{if } X_i = a_j \\ 0 & \text{if } X_i \neq a_j. \end{cases}$$

Then the sample average gives the statistic $\hat{p}(a_j) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} I_j(X_i)$. Then $\{\hat{p}(a_1), \hat{p}(a_2), \ldots, \hat{p}(a_m)\}$ are the nonparametric estimates of the probability mass function $p(x)$. Then $I_j(X_i)$ has a Bernoulli distribution (with $X_i = 1$ as a "success") and $p(a_j)$ as the probability of a success. Now

$$E[\hat{p}(a_j)] = \frac{1}{n}\sum_{i=1}^{n} E[I_j(X_i)[= \frac{1}{n} p(a_i j) = p(a_j).$$

So $\hat{p}(a_j)$ is an unbiased estimator of $p(a_j)$. Second, suppose that the sample space of X is infinite (but countable), say $\mathcal{D} = \{a_1, a_2, \ldots\}$. In what follows we consider histograms and so we need a finite number of categories. So we define the "groupings"

$$\{a_1\}, \{a_2\}, \ldots, \{a_m\}, \overline{a}_{m+1} = \{a_{m+1}, a_{m+2}, \ldots\}.$$

Then the estimates $\{\hat{p}(a_1), \hat{p}(a_2), \ldots, \hat{p}(a_m), \hat{p}(\overline{a}_{m+1})\}$ give estimates of $p(x)$. The "rule of thumb" is to include enough categories so that the frequency of category a_m exceeds twice the combined frequencies of $a_{m+1}, a_{m+2}, \ldots$.

Definition. A histogram of $\hat{p}(a_j)$ versus a_j is a *boxplot*. If the values of a_j represent qualitative categories (that is, a_j is a categorical variable) then the histogram is a *barchart*. If the values in sample space $\mathcal{D}$ are ordinal (i.e., the natural ordering $a_1, a_2, \ldots$ is numerically meaningful) then the histogram is given as a bar chart of abutting categories with height $\hat{p}(a_j)$ that are plotted in the natural order of the a_j's.

Example 4.1.5. The following data gives the hair color of $n = 22{,}361$ Scottish school children from the early 1990s. So discrete random variable a_j is a categorical variable with values Fair, Red, Medium, Dark, and Brown.

	Fair	**Red**	**Medium**	**Dark**	**Black**
Count	5789	1319	9418	5678	157
$\hat{p}(a_j)$	0.259	0.059	0.421	0.254	0.007

A barchart for this information is given in Figure 4.1.1.

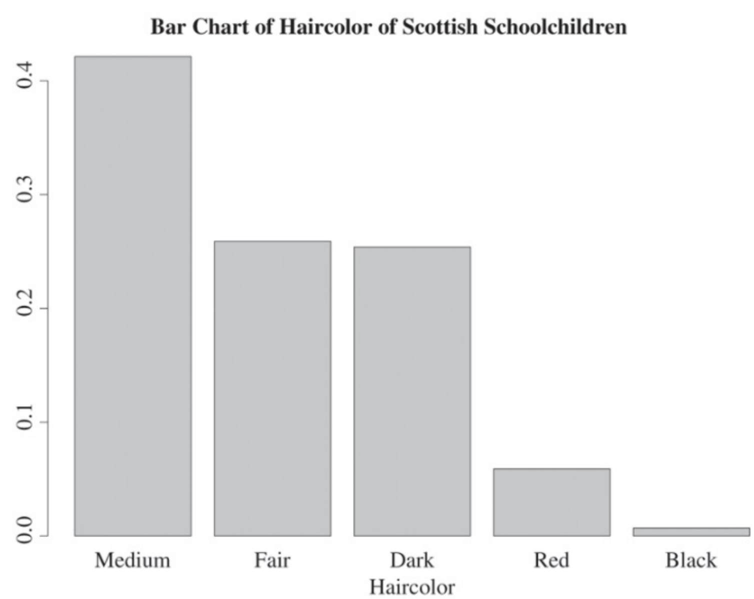

Figure 4.1.1

Example 4.1.6. The following 30 data points are simulated values generated from a Poisson distribution with mean $\lambda = 2$.

2	1	1	1	1	5	1	1	3	0	2	1	1	3	4
2	1	2	2	6	5	2	3	2	4	1	3	1	3	0

The nonparametric estimate of the probability mass function is:

j	0	1	2	3	4	5	≥ 6
$\hat{p}(j)$	0.067	0.367	0.233	0.167	0.067	0.067	0.033

The histogram for this data is given in Figure 4.1.2.

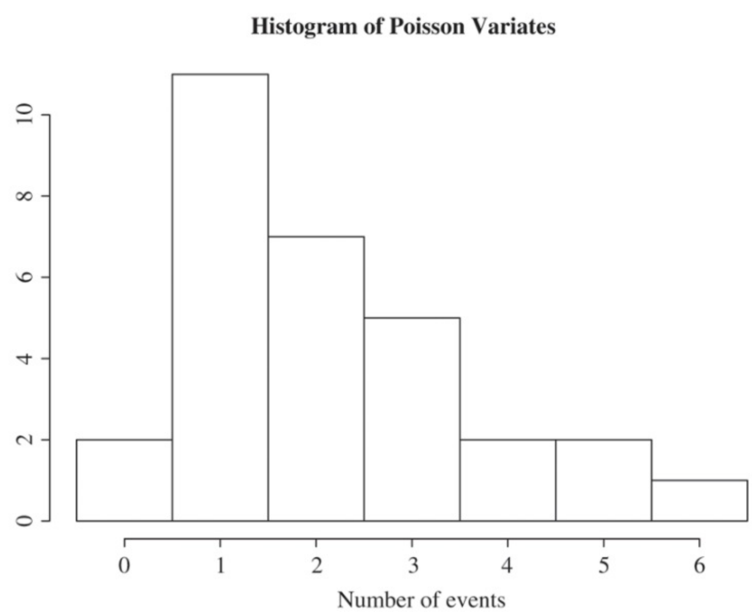

Figure 4.1.2

Note. Now suppose $X_1, X_2, \ldots, X_n$ is a random sample from a continuous random variable X with continuous probability density function $f(t)$. For some fixed x and a given $h > 0$, we consider the interval $(x - h, x + h)$. By the Mean Value Theorem for Integrals (see my online notes for Calculus 1 [MATH 1910] on 5.4.

The Fundamental Theorem of Calculus; see Theorem 5.3), there is some ξ with $\xi \in (x-h, x+h)$ such that

$$P(x-h < X < x+h) = \int_{x-h}^{x+h} f(t)\,dt = 2hf(\xi)$$

and this is approximately $2hf(x)$ (with the approximation "good" for h "small"). No we expect $P(X-h < X < x+h)$ to be the proportion of the random sample in $(s-h, x+h)$. So we take the (nonparametric) estimate of $f(x)$ at the given x value as

$$\hat{f}(x) = \frac{\#\{x-h < X < x+h\}}{2hn} = \frac{|\{s-x < X < x+h\}|}{2hn}.$$

Similar to the discrete case, we define

$$I_i(x) = \begin{cases} 1 & \text{if } x-h < X_i < x+h \\ 0 & \text{otherwise} \end{cases} \quad \text{for } i = 1, 2, \ldots, n,$$

and the nonparametric estimate of $f(x)$ is

$$\hat{f}(x) = \frac{1}{2hn}\sum_{i=1}^{n} I_i(x).$$

Now $E[I_i(x)] = 2hf(\xi)$ where ξ is as introduced above. Since the samples are identically distributed then

$$\begin{aligned} E[\hat{f}(x)] &= E\left[\frac{1}{2hn}\sum_{i=1}^{n} I_i(x)\right] \\ &= \frac{1}{2hn}\sum_{i=1}^{n} E[I_i(x)] \text{ since } E \text{ is linear by Theorem 1.8.2} \\ &= \frac{1}{2hn}(n)(2hf(\xi)) = f(\xi), \end{aligned}$$

and (since f is continuous) $\lim_{h\to 0^+} f(\xi) = f(x)$. So $\hat{f}(x)$ is *approximately* an unbiased estimator of $f(x)$. The indicator function I_i is called a *rectangular kernel* with *bandwidth* $2h$.

Note. If $x_1, x_2, \ldots, x_n$ are the realized values of a random sample on a continuous random variable X with probability density function $f(x)$. A histogram approximation of f can be created by choosing $m \in \mathbb{N}$, $h > 0$, and $a < \min_{1 \le i \le n}\{x_i\}$ so that the disjoint intervals

$$(a-h, a+h], (a+h, a+3h], (a+3h, 1+5h], \ldots, (a+(2m-3)h, a+(2m-1)h]$$

cover the range of the realized sample interval $[\min_{1 \le i \le n}\{x_i\}, \max_{1 \le i \le n}\{x_i\}]$. We then define the classes

$$A_j(a+(2j-3)h, a+(2j-1)h] \text{ for } j = 1, 2, \ldots, m.$$

Define $\hat{f}_h(x)$ as

$$\hat{f}_h(x) = \frac{\#\{x_i \in A_j\}}{2hn} = \frac{|\{x_i \in A_j\}|}{2hn} \text{ for } x \in A_j.$$

Notice that

$$\int_{-\infty}^{\infty} \hat{f}_h(x)\,dx = \int_{1-h}^{a+(2m-1)h} \hat{f}_h(x)\,dx = \sum_{j=1}^{m} \int_{A_j} \frac{\#\{x_i \in A_j\}}{2hn}\,dx$$

$$= \frac{1}{2hn} \sum_{j=1}^{m} \#\{x_i \in A_j\}\,(2h) = \frac{1}{n}(n) = 1,$$

as desired. Now *how* to best choose m and h is not addressed in Hogg, McKean, and Craig. They state that "most statistical packages...are current on recent research for selection of classes." See page 232.

Example 4.1.7. In Example 4.1.3, the following 24 data points are presented for the concentration of sulfur dioxide in a damaged Bavarian forest:

33.4	38.6	41.7	43.9	44.4	45.3	46.1	47.6	50.0	52.4	52.7	53.9
54.3	55.1	56.4	56.5	60.7	61.8	62.2	63.4	65.5	66.6	70.0	71.5

The mean of this data is 53.91 and the (sample) standard deviation is 10.07. Software package R produces the histogram given in Figure 4.1.3 along with the solid curve that approximates the histogram. The dashed curve is a normal distribution with mean and standard deviation equal to those of the sample. Notice that the normal approximation appears to be a poorer fit to the data.

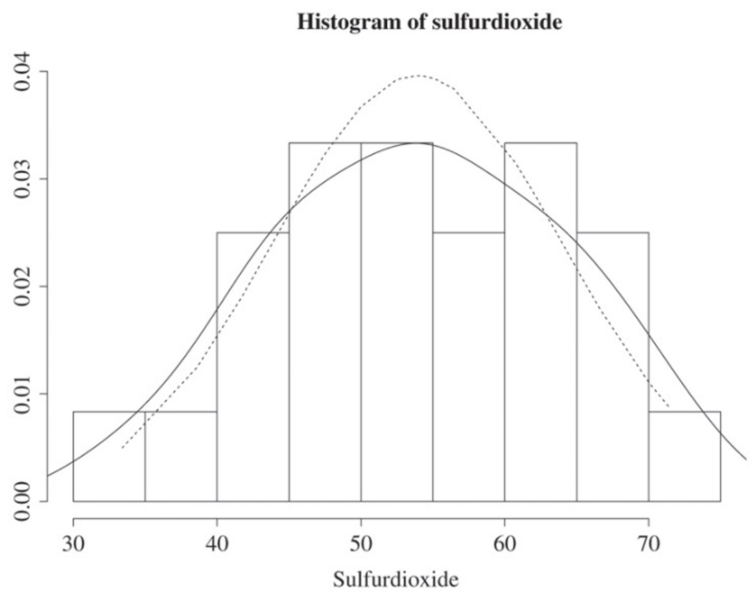

Figure 4.1.3