

As in the previous section, we assume a sample  $\mathcal{S} = \{x_1, x_2, \dots, x_n\}$  with  $n$  sufficiently large that it is reasonable to do more than just calculate the sample mean and standard deviation. In sharp contrast to the discrete-data situation in the previous section, however, we now consider continuous-data histograms where the data values  $x_1, x_2, \dots, x_n$  are assumed to be real-valued and generally distinct.

### 4.3.1 CONTINUOUS-DATA HISTOGRAMS

Given a real-valued sample  $\mathcal{S} = \{x_1, x_2, \dots, x_n\}$ , without loss of generality we can assume the existence of real-valued lower and upper bounds  $a, b$  with the property that

$$a \leq x_i < b \quad i = 1, 2, \dots, n.$$

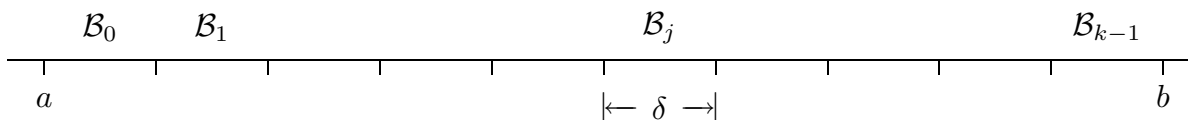
This defines an interval of possible values for some random variable  $X$  as  $\mathcal{X} = [a, b) = \{x \mid a \leq x < b\}$ , that can be partitioned into  $k$  equal-width *bins* ( $k$  is a positive integer) as

$$[a, b) = \bigcup_{j=0}^{k-1} \mathcal{B}_j = \mathcal{B}_0 \cup \mathcal{B}_1 \cup \dots \cup \mathcal{B}_{k-1},$$

where the bins are  $\mathcal{B}_0 = [a, a + \delta)$ ,  $\mathcal{B}_1 = [a + \delta, a + 2\delta)$ ,  $\dots$  and the width of each bin is

$$\delta = \frac{b - a}{k}$$

as illustrated on the axis below.



**Definition 4.3.1** Given the sample  $\mathcal{S} = \{x_1, x_2, \dots, x_n\}$  and the related parameters  $a, b$ , and either  $k$  or  $\delta$ , for each  $x \in [a, b)$  there is a unique bin  $\mathcal{B}_j$  with  $x \in \mathcal{B}_j$ . The estimated *density* of the random variable  $X$  is then

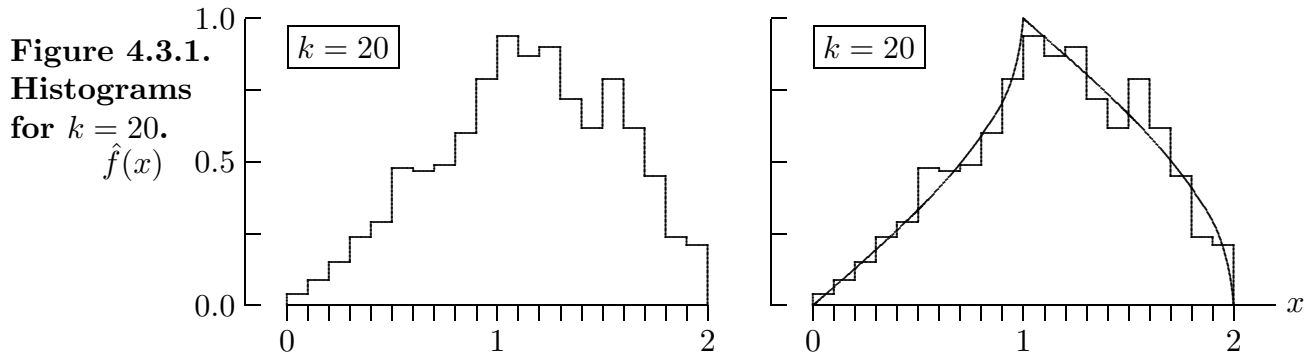
$$\hat{f}(x) = \frac{\text{the number of } x_i \in \mathcal{S} \text{ for which } x_i \in \mathcal{B}_j}{n \delta} \quad a \leq x < b.$$

A *continuous-data histogram* is a “bar” plot of  $\hat{f}(x)$  versus  $x$ .\*

As the following example illustrates,  $\hat{f}(\cdot)$  is a piecewise constant function (constant over each bin) with discontinuities at the histogram bin boundaries. Since simulations tend to produce large data sets, we adopt the graphics convention illustrated in Example 4.3.1 of drawing  $\hat{f}(\cdot)$  as a sequence of piecewise constant horizontal segments connected by vertical lines. This decision is consistent with maximizing the “data-to-ink” ratio (see Tufte, 2001).

\* Compare Definition 4.3.1 with Definition 4.2.1. The bin index is  $j = \lfloor (x - a)/\delta \rfloor$ . The *density* is the relative frequency of the data in bin  $\mathcal{B}_j$  *normalized* via a division by  $\delta$ .

**Example 4.3.1** As an extension of Example 4.1.1, a modified version of program `buf-fon` was used to generate a random variate sample of  $n = 1000$  observations of the  $x$ -coordinate of the righthand endpoint of a unit-length needle dropped at random. To form a continuous-data histogram of the sample, the values  $a = 0.0$  and  $b = 2.0$  are, in this case, obvious choices for lower and upper bounds.\* The number of histogram bins was selected, somewhat arbitrarily, as  $k = 20$  so that  $\delta = (b - a)/k = 0.1$ . The resulting histogram is illustrated on the left side of Figure 4.3.1.



Illustrated on the right is the same histogram with a continuous curve superimposed. As discussed later in Chapter 7, this curve represents the *probability density function* of the righthand endpoint, which is the limit to which the histogram will converge as the sample size ( $n$ ) approaches infinity *and* simultaneously  $k$  approaches infinity or, equivalently,  $\delta$  approaches zero.

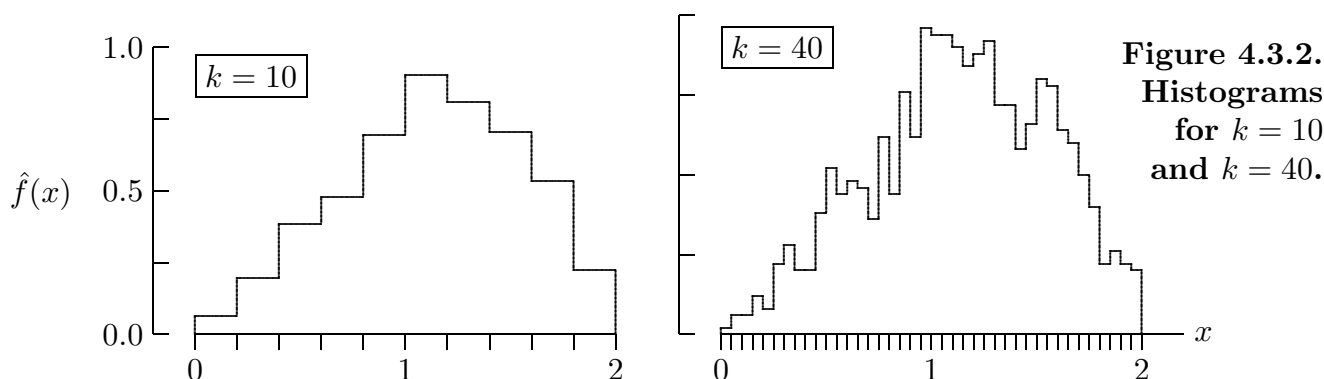
### Histogram Parameter Guidelines

There is an experimental component to choosing the continuous-data histogram parameters  $a$ ,  $b$ , and either  $k$  or  $\delta$ . These guidelines are certainly not rigid rules.

- For data produced by a (valid) simulation there should be few, if any, outliers. That is, the bounds  $a$ ,  $b$  should be chosen so that few, if any, data points in the sample are excluded. Of course, as in the discrete-data case, prior knowledge of reasonable values for  $a$ ,  $b$  may not always be easily obtained.
- If  $k$  is too large ( $\delta$  is too small) then the histogram will be too “noisy” with the potential of exhibiting false features caused by natural sampling variability; if  $k$  is too small ( $\delta$  is too large) then the histogram will be too “smooth” with the potential of masking significant features — see Example 4.3.2.
- The histogram parameters should always be chosen with the aesthetics of the resulting figure in mind, e.g., for  $a = 0$  and  $b = 2$ , if  $n$  is sufficiently large a choice of  $k = 20$  ( $\delta = 0.1$ ) would be a better choice than  $k = 19$  ( $\delta \cong 0.10526$ ).
- Typically  $\lfloor \log_2(n) \rfloor \leq k \leq \lfloor \sqrt{n} \rfloor$  with a bias toward  $k \cong \lfloor (5/3) \sqrt[3]{n} \rfloor$  (Wand, 1997).
- Sturges’s rule (Law and Kelton, 2000, page 336) suggests  $k \cong \lfloor 1 + \log_2 n \rfloor$ .

\* Because the needle has unit length,  $0.0 < x_i < 2.0$  for  $i = 1, 2, \dots, n$ .

**Example 4.3.2** As a continuation of Example 4.3.1, two additional histograms are illustrated in Figure 4.3.2 corresponding to  $k = 10$  ( $\delta = 0.2$ ) on the left and to  $k = 40$  ( $\delta = 0.05$ ) on the right. The histogram on the right is clearly too noisy consistent with a choice of  $k$  that violates the second histogram parameter guideline (i.e.,  $k$  is too large relative to  $n$ ). Although this characterization is less clear, the histogram on the left may be too smooth because  $k$  is too small. For this sample, the best choice of  $k$  seems to be somewhere between 10 and 20 (using the last of the histogram guidelines,  $9 \leq k \leq 31$  and  $k \cong \lfloor (5/3) \sqrt[3]{1000} \rfloor = 16$ ). Since simulations typically produce large numbers of observations, and therefore many histogram bins are required, we avoid the common practice of dropping the vertical lines to the horizontal axis which visually partitions the bins. Including these lines unnecessarily clutters the histogram and obscures the shape of the histogram, particularly for large  $k$ .

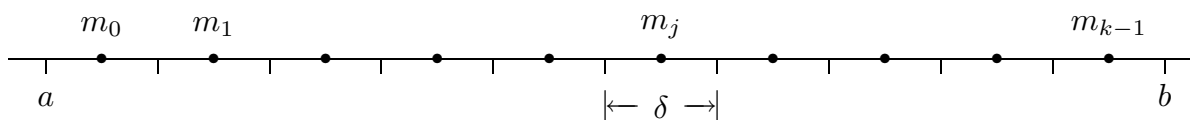


### Histogram Integrals

**Definition 4.3.2** As an extension of Definition 4.3.1, for each  $j = 0, 1, \dots, k-1$  define  $p_j$  as the *relative frequency* of points in  $\mathcal{S} = \{x_1, x_2, \dots, x_n\}$  that fall into bin  $\mathcal{B}_j$ . Since the bins form a partition of  $[a, b]$  each point in  $\mathcal{S}$  is counted exactly once (assuming no outliers) and so  $p_0 + p_1 + \dots + p_{k-1} = 1$ . In addition, define the *bin midpoints*

$$m_j = a + \left(j + \frac{1}{2}\right) \delta \quad j = 0, 1, \dots, k-1$$

as illustrated on the axis below.



It follows from Definitions 4.3.1 and 4.3.2 that

$$p_j = \delta \hat{f}(m_j) \quad j = 0, 1, \dots, k-1$$

and that  $\hat{f}(\cdot)$  is a non-negative function with unit area. That is

$$\int_a^b \hat{f}(x) dx = \sum_{j=0}^{k-1} \int_{\mathcal{B}_j} \hat{f}(x) dx = \sum_{j=0}^{k-1} \hat{f}(m_j) \int_{\mathcal{B}_j} dx = \sum_{j=0}^{k-1} \left(\frac{p_j}{\delta}\right) \delta = \sum_{j=0}^{k-1} p_j = 1.$$

In addition to proving that  $\int_a^b \hat{f}(x) dx = 1$ , the previous derivation can be extended to the two integrals

$$\int_a^b x \hat{f}(x) dx \quad \text{and} \quad \int_a^b x^2 \hat{f}(x) dx.$$

That is, for the first of these two integrals,

$$\int_a^b x \hat{f}(x) dx = \sum_{j=0}^{k-1} \int_{\mathcal{B}_j} x \hat{f}(x) dx = \sum_{j=0}^{k-1} \hat{f}(m_j) \int_{\mathcal{B}_j} x dx = \sum_{j=0}^{k-1} \left(\frac{p_j}{\delta}\right) \int_{\mathcal{B}_j} x dx$$

and in analogous fashion, for the second integral

$$\int_a^b x^2 \hat{f}(x) dx = \dots = \sum_{j=0}^{k-1} \left(\frac{p_j}{\delta}\right) \int_{\mathcal{B}_j} x^2 dx.$$

In this way, because  $\hat{f}(\cdot)$  is piecewise constant, the two integrals over  $[a, b)$  are reduced to simple polynomial integration over each histogram bin. In particular, for  $j = 0, 1, \dots, k-1$ ,

$$\int_{\mathcal{B}_j} x dx = \int_{m_j - \delta/2}^{m_j + \delta/2} x dx = \frac{(m_j + \delta/2)^2 - (m_j - \delta/2)^2}{2} = \dots = m_j \delta,$$

so that the first integral reduces to

$$\int_a^b x \hat{f}(x) dx = \sum_{j=0}^{k-1} \left(\frac{p_j}{\delta}\right) \int_{\mathcal{B}_j} x dx = \sum_{j=0}^{k-1} m_j p_j.$$

Similarly

$$\int_{\mathcal{B}_j} x^2 dx = \frac{(m_j + \delta/2)^3 - (m_j - \delta/2)^3}{3} = \dots = m_j^2 \delta + \frac{\delta^3}{12}$$

so that the second integral reduces to

$$\int_a^b x^2 \hat{f}(x) dx = \sum_{j=0}^{k-1} \left(\frac{p_j}{\delta}\right) \int_{\mathcal{B}_j} x^2 dx = \left(\sum_{j=0}^{k-1} m_j^2 p_j\right) + \frac{\delta^2}{12}$$

Therefore, the two integrals  $\int_a^b x \hat{f}(x) dx$  and  $\int_a^b x^2 \hat{f}(x) dx$  can be evaluated *exactly* by finite summation. This is significant because the continuous-data histogram sample mean and sample standard deviation are defined in terms of these two integrals.

### Histogram Mean and Standard Deviation

**Definition 4.3.3** Analogous to the discrete-data equations in Definition 4.2.2 (replacing  $\sum$ 's with  $\int$ 's), the *continuous-data histogram mean* and *standard deviation* are defined as

$$\bar{x} = \int_a^b x \hat{f}(x) dx \quad \text{and} \quad s = \sqrt{\int_a^b (x - \bar{x})^2 \hat{f}(x) dx}.$$

The continuous-data histogram variance is  $s^2$ .

The integral equations in Definition 4.3.3 provide the motivation for defining the population mean and standard deviation of continuous random variables in Chapter 7. From the integral equations derived previously it follows that a continuous-data histogram mean can be evaluated exactly by finite summation with the equation

$$\bar{x} = \sum_{j=0}^{k-1} m_j p_j.$$

Moreover,

$$s^2 = \int_a^b (x - \bar{x})^2 \hat{f}(x) dx = \cdots = \left( \int_a^b x^2 \hat{f}(x) dx \right) - \bar{x}^2$$

and, similarly,

$$\sum_{j=0}^{k-1} (m_j - \bar{x})^2 p_j = \cdots = \left( \sum_{j=0}^{k-1} m_j^2 p_j \right) - \bar{x}^2.$$

From these last two equations it follows that a continuous-data histogram standard deviation can be evaluated exactly by finite summation with either of the two equations\*

$$s = \sqrt{\left( \sum_{j=0}^{k-1} (m_j - \bar{x})^2 p_j \right) + \frac{\delta^2}{12}} \quad \text{or} \quad s = \sqrt{\left( \sum_{j=0}^{k-1} m_j^2 p_j \right) - \bar{x}^2 + \frac{\delta^2}{12}}.$$

In general the continuous-data *histogram* mean and standard deviation will differ slightly from the *sample* mean and standard deviation, even if there are no outliers. This difference is caused by the *quantization error* associated with the arbitrary binning of continuous data. In any case this difference should be slight — if the difference is not slight then the histogram parameters  $a$ ,  $b$ , and either  $k$  or  $\delta$  should be adjusted. Although the histogram mean and standard deviation are inferior to the sample mean and standard deviation, there are circumstances where a data analyst is presented with binned data and does not have access to the associated raw data.

---

\* There is some disagreement in the literature relative to these two equations. Many authors use these equations, with the  $\delta^2/12$  term ignored, to *define* the continuous-data histogram standard deviation.

**Example 4.3.3** For the 1000-point sample in Example 4.3.1, using  $a = 0.0$ ,  $b = 2.0$ ,  $k = 20$ , the difference between the sample and histogram statistics is slight:

	raw data	histogram	histogram with $\delta = 0$
$\bar{x}$	1.135	1.134	1.134
$s$	0.424	0.426	0.425

Moreover, by comparing the last two columns in this table we see that, in this case, there is essentially no impact of the  $\delta^2/12 = (0.1)^2/12 = 1/1200$  term in the computation of the histogram standard deviation.

### 4.3.2 COMPUTATIONAL CONSIDERATIONS

**Algorithm 4.3.1** Given the parameters  $a$ ,  $b$ ,  $k$ , and the real-valued data  $x_1, x_2, \dots$ , this algorithm computes a continuous-data histogram.

```

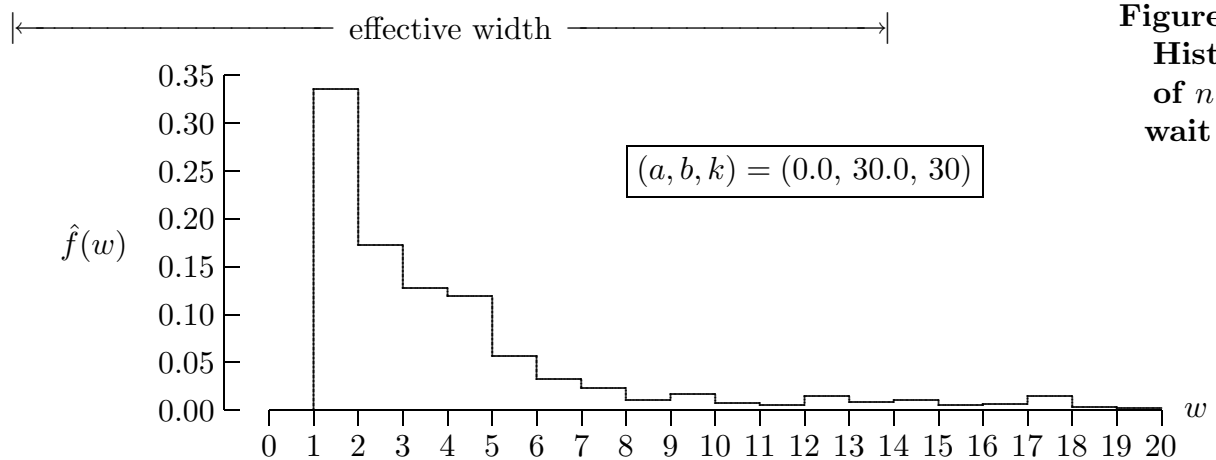
long count[k];
 $\delta = (b - a) / k$ ;
n = 0;
for (j = 0; j < k; j++)
    count[j] = 0;                                /* initialize bin counters */
outliers.lo = 0;                                /* initialize outlier counter on  $(-\infty, a)$  */
outliers.hi = 0;                                /* initialize outlier counter on  $[b, \infty)$  */
while ( more data ) {
    x = GetData();
    n++;
    if ((a <= x) and (x < b)) {
        j = (long) (x - a) /  $\delta$ ;
        count[j]++;                               /* increment appropriate bin counter */
    }
    else if (a > x)
        outliers.lo++;
    else
        outliers.hi++;
}
return n, count[], outliers                       /*  $p_j$  is (count[j] / n) */

```

The previously derived summation equations for  $\bar{x}$  and  $s$  can then be used to compute the histogram mean and standard deviation.

If the sample is written to a disk file (using sufficient floating-point precision), then one can experiment with different values for the continuous-data histogram parameters in an *interactive* graphics environment. Consistent with this, program `cdh` illustrates the construction of a continuous-data histogram for data read from a text file. For an alternative approach see Exercise 4.3.7.

**Example 4.3.4** As in Example 4.1.7, a modified version of program `ssq2` was used to generate a sample consisting of the waits  $w_1, w_2, \dots, w_n$  experienced by the first  $n = 1000$  jobs. The simulation was initialized to simulate steady-state and the `rng` initial seed was 12345. Program `cdh` was used to process this sample with the continuous-data histogram parameters set to  $(a, b, k) = (0.0, 30.0, 30)$ , as illustrated in Figure 4.3.3.



**Figure 4.3.3.**  
Histogram  
of  $n = 1000$   
wait times.

The histogram mean is 4.57 and the histogram standard deviation is 4.65. The associated two-standard deviation “effective width” of the sample is  $4.57 \pm 9.30$ , as illustrated. Consistent with previous examples this interval includes most of the points in the sample.\*

### Point Estimation

The issue of sampling variability and how it relates to the uncertainty of any probability estimate derived from a Monte Carlo simulation was considered in the previous section. This issue is sufficiently important to warrant reconsideration here in the context of continuous-data histograms. In particular, recall that for Example 4.2.8 there was a sample  $\mathcal{S} = \{p_1, p_2, \dots, p_n\}$  of  $n = 1000$  point estimates of the probability of winning at the game of craps. One figure in Example 4.2.8 corresponds to probability estimates based on 25 plays of the game per estimate; the other figure is based on 100 plays per estimate. For both figures, the samples were displayed as discrete-data histograms and we were primarily interested in studying how the width of the histograms decreased with an increase in the number of games per estimate.

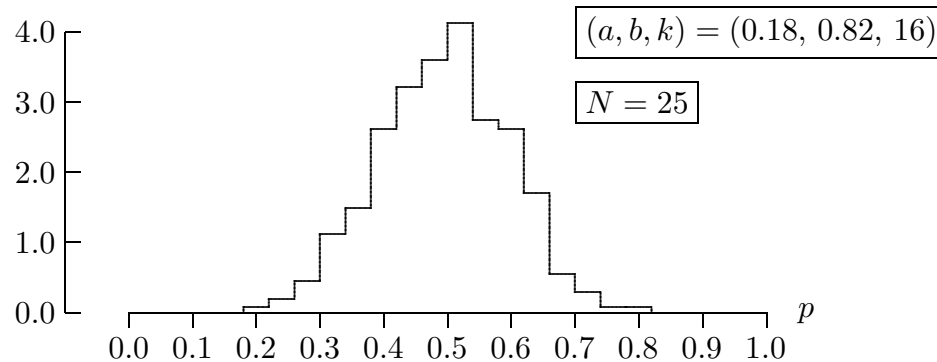
If we were to study the issue of how the histogram width depends on the number of games in more detail, it would be natural to quadruple the number of games per estimate to 400, 1600, 6400, etc. As the number of games increases, the associated discrete-data histograms will look more continuous as the histogram spikes get closer together. Given that, it is natural to ignore the inherently discrete nature of the probability estimates and, instead, treat the sample as though it were continuous data. That is what is done in the following example.

---

\* The interval  $4.57 \pm 9.30$  includes approximately 93% of the sample. There are 21 points in the sample with a value larger than 20.0, the largest of which is 28.2.

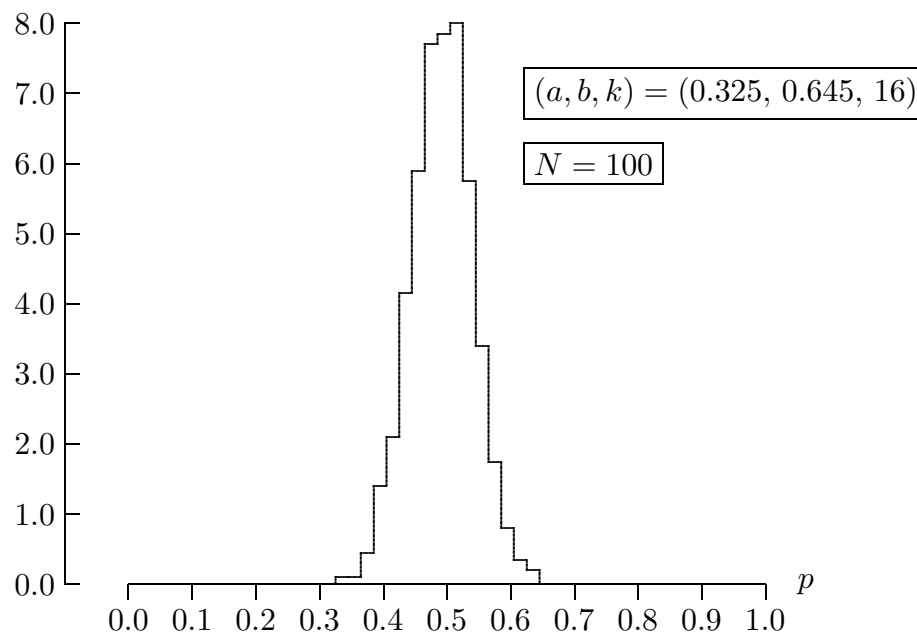
**Example 4.3.5** The sample of  $n = 1000$  probability estimates based on  $N = 25$  plays of the game of craps from Example 4.2.8 was processed as a continuous-data histogram with parameters  $(a, b, k) = (0.18, 0.82, 16)$ . This choice of parameters is matched to the “resolution” of the estimates (which is  $\delta = 0.04$ ) with the center of each histogram bin corresponding to exactly one possible value of an estimate.\*

**Figure 4.3.4.**  
Histogram of  
 $n = 1000$   
estimates  $\hat{f}(p)$   
of winning at  
craps from  
 $N = 25$  plays.



In a similar way the sample of  $n = 1000$  probability estimates based on  $N = 100$  plays of the game was processed as a continuous-data histogram with parameters  $(a, b, k) = (0.325, 0.645, 16)$ . This choice of parameters is matched to half the resolution of the estimates ( $\delta = 0.02$ ) with the center of each histogram bin corresponding to the midpoint of exactly two possible values of an estimate, as illustrated in Figure 4.3.5.

**Figure 4.3.5.**  
Histogram of  
 $n = 1000$   
estimates  $\hat{f}(p)$   
of winning at  
craps from  
 $N = 100$  plays.



\* Continuous-data histograms are *density* estimates, not *probability* estimates. Thus values of  $\hat{f}(p)$  can be greater than 1.0 are possible, as in this example, but the histogram value can't stay at that height too long since  $\int_0^1 \hat{f}(p) dp = 1$ .

As in Example 4.2.8, we see that increasing the number of replications per estimate by a factor of *four* causes the uncertainty in any one probability estimate to decrease by a factor of *two*. The advantage to using a continuous-data histogram representation in this example is that experimentation with more games per estimate can be naturally accommodated. As the number of games per estimate is increased the histogram will become taller and narrower, always centered near the true value of  $244/495 \cong 0.4929$ , and always consistent with the invariant unit-area requirement  $\int_0^1 \hat{f}(p) dp = 1$ .

### Random Events Yield Exponentially Distributed Inter-Events

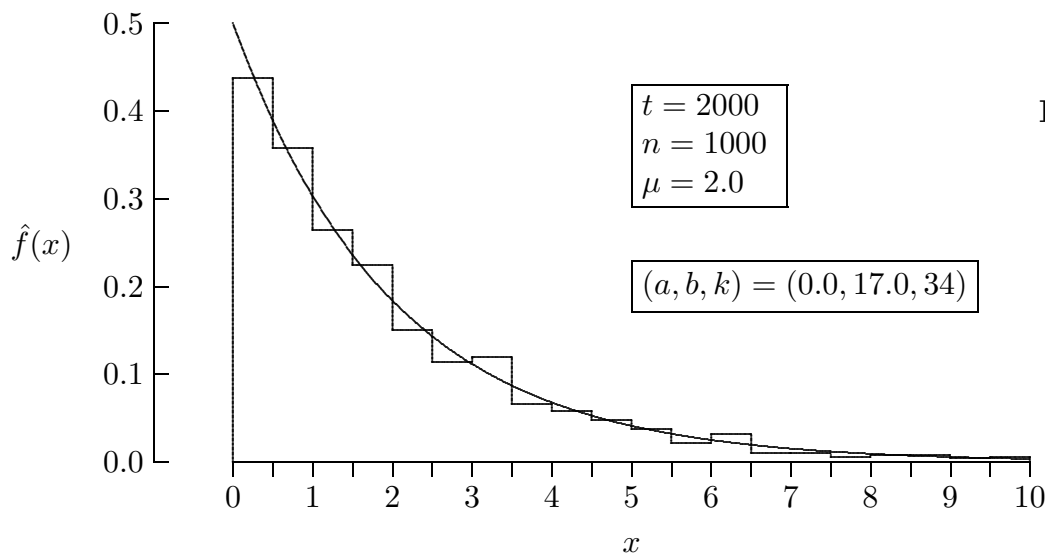
**Example 4.3.6** As another continuous-data histogram example, pick a positive parameter  $t > 0$  and suppose that  $n$  calls to the function `Uniform(0, t)` are used to generate a random variate sample of  $n$  events occurring *at random* in the interval  $(0, t)$ . If these  $n$  event times are then sorted into increasing order, the result is a sequence of events times  $u_1, u_2, \dots, u_n$  ordered so that  $0 < u_1 < u_2 < \dots < u_n < t$ . With  $u_0 = 0$  define

$$x_i = u_i - u_{i-1} > 0 \quad i = 1, 2, \dots, n,$$

as the inter-event times. Let  $\mu = t/n$  and recognize that

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{u_n - u_0}{n} \cong \frac{t}{n} = \mu$$

so that the sample mean of the inter-event times is approximately  $\mu$ . One might expect that a histogram of the inter-event times will be approximately “bell-shaped” and centered at  $\mu$ . As illustrated in Figure 4.3.6, however, that is *not* the case — the histogram has an *exponential* shape.\* In particular, the smallest inter-event times are the most likely.



**Figure 4.3.6.**  
Histogram  
of  $n = 1000$   
inter-event  
times.

\* The histogram actually has an even longer tail than that illustrated in Figure 4.3.6. There are eight data points in the sample (out of 1000) with a value larger than 10.

The continuous curve superimposed illustrates that in this case  $\hat{f}(x) \cong f(x)$  where

$$f(x) = \frac{1}{\mu} \exp(-x/\mu) \quad x > 0.$$

Indeed, in the limit as  $n \rightarrow \infty$  and  $\delta \rightarrow 0$  (with  $\mu$  held constant)  $\hat{f}(x) \rightarrow f(x)$  for all  $x > 0$ .<sup>\*</sup> We will return to this important example in Chapter 7.

### 4.3.3 EMPIRICAL CUMULATIVE DISTRIBUTION FUNCTIONS

The fact that parameters, such as the number of bins  $k$ , must be chosen by the modeler is a distinct drawback for continuous-data histograms. Two different data analysts could have the extreme misfortune of choosing different binning schemes for the same data set and produce histograms with somewhat different shapes. This is particularly true if the sampling variability inherent in the data set conspires with the two different binning schemes to accentuate the difference between the two histograms.

An alternative approach to plotting continuous data that avoids arbitrary parameters from the modeler is known as the *empirical cumulative distribution function*.

**Definition 4.3.4** Given the sample  $\mathcal{S} = \{x_1, x_2, \dots, x_n\}$ , the estimated *cumulative distribution function* of the random variable  $X$  is

$$\hat{F}(x) = \frac{\text{the number of } x_i \in \mathcal{S} \text{ for which } x_i \leq x}{n}.$$

The *empirical cumulative distribution function* is a plot of  $\hat{F}(x)$  versus  $x$ .

When  $x_1, x_2, \dots, x_n$  are distinct, the plot of  $\hat{F}(x)$  versus  $x$  is a step function with an upward step of  $1/n$  at each data value. On the other hand, if  $d$  values are tied at a particular  $x$ -value, the height of the riser on that particular step is  $d/n$ .

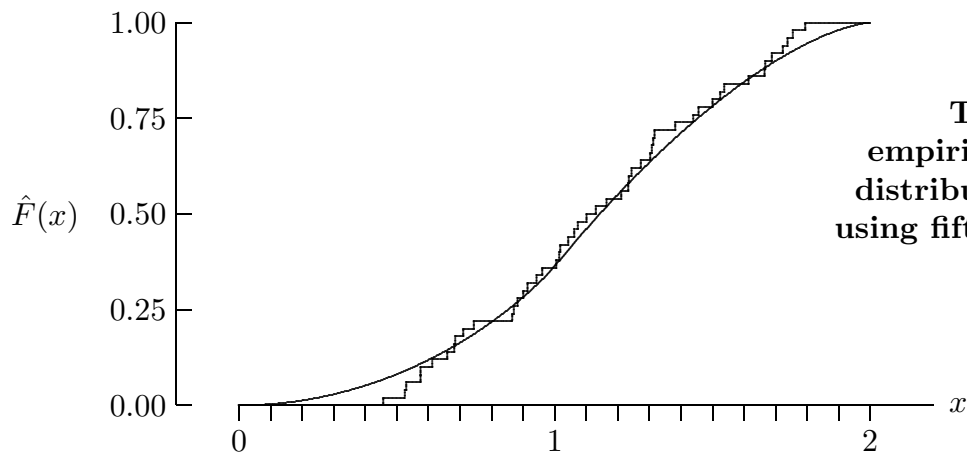
The empirical cumulative distribution function requires no parameters from the modeler, which means that one data set always produces the same empirical cumulative distribution function.

We now compare the computational complexity and memory requirements of the two graphical procedures. The continuous-data histogram algorithm performs a single-pass through the data values, with time complexity  $O(n)$  and requires  $k$  memory locations. Plotting the empirical cumulative distribution function requires a sort, with time complexity  $O(n \log n)$  at best, and all data values must be stored simultaneously, requiring  $n$  memory locations.

---

<sup>\*</sup> See, for example, Rigdon and Basu (2000, pages 50–52) for the details concerning the relationship between the uniform and exponential distributions in this case.

**Example 4.3.7** Consider again the modified version of program `buffon` that was used to generate a random variate sample of  $n = 50$  observations of the  $x$ -coordinate of the righthand endpoint of a unit-length needle dropped at random using an initial seed of 123456789. The empirical cumulative distribution function is plotted in Figure 4.3.6 using our choice among the four plotting formats displayed for discrete data in Figure 4.2.8. There is an upward step of  $1/50$  at each of the values generated. The minimum and maximum values generated by the program are 0.45688 and 1.79410, respectively, which correspond to the horizontal position of the first and last step of the plot of the empirical cumulative distribution function. In this example it is possible to compute the theoretical cumulative distribution function using the axiomatic approach to probability. This is the smooth curve superimposed in Figure 4.3.7. The difference between the step function and the smooth curve is due to random sampling variability.



**Figure 4.3.7.**  
Theoretical and  
empirical cumulative  
distribution functions  
using fifty replications.

How does one compare the advantages and disadvantages of the continuous-data histogram and empirical cumulative distribution function? The histogram is clearly superior at detecting the *shape* of the distribution of the random quantity of interest. The arbitrary parameters associated with binning are its only downside. Selecting the continuous-data histogram parameters is more of an art than a science, which drives us to an alternative. The empirical cumulative distribution function is nonparametric, and thus less susceptible to the effects of sampling variability since there is no binning. Unfortunately, its shape is less distinct than the continuous-data histogram. It is often used to compare a hypothesized or fitted distribution to a data set using a statistical “goodness-of-fit” test.\*

Increased CPU speeds makes generating large data sets possible in simulation, putting a strain on the memory and speed associated with plotting an empirical cumulative distribution function. Fortunately this is typically only done once during a simulation run.

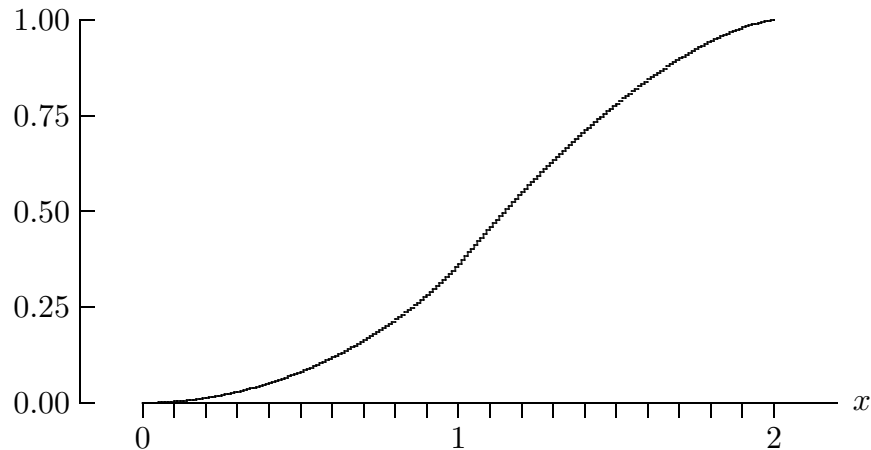
We end this section with an example which combines some of the best features of continuous-data histograms and empirical cumulative distribution functions.

---

\* The Kolmogorov–Smirnov, Anderson–Darling, and Cramer–von Mises are three well-known statistical goodness-of-fit tests for continuous data.

**Example 4.3.8** If the sample size from the previous example ( $n = 50$ ) were dramatically increased to  $n = 1\,000\,000\,000$ , one would expect the empirical cumulative distribution function to become very smooth and approximate the theoretical curve. We plot an empirical cumulative distribution function in order to take advantage of its nonparametric nature. Unfortunately, plotting the empirical cumulative distribution function requires that we store and sort the one billion righthand needle endpoints. Since each data value lies on  $0 \leq x \leq 2$ , we can create a close approximation to the empirical cumulative distribution function by defining, for instance, 200 equal-width cells on  $0 \leq x \leq 2$ , i.e.,  $[0, 0.01), [0.01, 0.02), \dots, [1.99, 2.00)$ .<sup>\*</sup> Counts associated with these cells are accumulated, and plot of the cumulative proportions associated with these cells should be virtually identical to a plot from the raw data. The cells take advantage of the fact that the counts can be updated as the data is generated, eliminating the need for storage and sorting. The plot of the cumulative proportions shown in Figure 4.3.8 is much smoother than the plot for  $n = 50$  in Figure 4.3.7 because of the huge number of replications. The difference between the plot in Figure 4.3.8 and the true cumulative distribution function could only be apparent to someone with a powerful microscope!

**Figure 4.3.8.** Approximate empirical cumulative distribution function with one billion replications.



#### 4.3.4 EXERCISES

**Exercise 4.3.1** (a) Use program `cdh` to construct a continuous-data histogram like the one on the left in Example 4.3.1, but corresponding to a needle of length  $r = 2$ . (b) Based on this histogram what is the probability that the needle will cross at least one line. (c) What is the corresponding axiomatic probability that a needle of length  $r = 2$  will cross at least one line?

---

<sup>\*</sup> The number of cells chosen in this example (200) is arbitrary. The formulas given in Section 4.3.1 for choosing the number of histogram cells do not apply here. The choice depends on the physical size of the plot, the desired smoothness, and the sample size.

**Exercise 4.3.2** Repeat the experiment in Example 4.3.6 with  $t = 5000$  and  $n = 2000$ . Do not use a bubble sort.

**Exercise 4.3.3** Fill in the  $= \dots =$ 's in the derivation of the two equations

$$\int_a^b x \hat{f}(x) dx = \dots = \sum_{j=0}^{k-1} m_j p_j \quad \text{and} \quad \int_a^b x^2 \hat{f}(x) dx = \dots = \left( \sum_{j=0}^{k-1} m_j^2 p_j \right) + \frac{\delta^2}{12}.$$

**Exercise 4.3.4** Generate a random variate sample  $x_1, x_2, \dots, x_n$  of size  $n = 10\,000$  as follows:

```
for (i = 1; i <= n; i++)
    x_i = Random() + Random();
```

(a) Use program `cdh` to construct a 20-bin continuous-data histogram. (b) Can you find an equation that seems to fit the histogram density well?

**Exercise 4.3.5** (a) As a continuation of Exercise 1.2.6, construct a continuous-data histogram of the service times. (b) Compare the *histogram* mean and standard deviation with the corresponding *sample* mean and standard deviation and justify your choice of the histogram parameters  $a$ ,  $b$ , and either  $k$  or  $\delta$ .

**Exercise 4.3.6** As an extension of Definition 4.3.1, the *cumulative* histogram density is defined as

$$\hat{F}(x) = \int_a^x \hat{f}(t) dt \quad a \leq x < b.$$

Derive a finite summation equation for  $\hat{F}(x)$ .

**Exercise 4.3.7<sup>a</sup>** To have more general applicability, program `cdh` needs to be restructured to support file redirection, like programs `uvs` and `ddh`. That part is easy. The ultimate objective here, however, should be a large-sample “auto-`cdh`” program that buffers, say, the first 1000 sample values and then automatically computes good values for the histogram parameters  $a$ ,  $b$ , and either  $k$  or  $\delta$  with a dynamic data structure allowance for additional bins to be added at the tails of the histogram and thereby avoid outliers if extreme values occur. (a) Construct such a program. (b) Discuss the logic you used for computing good values for the histogram parameters.

**Exercise 4.3.8<sup>a</sup>** Show that the theoretical cumulative distribution function superimposed over the empirical cumulative distribution function in Figure 4.3.7 is

$$F(x) = \begin{cases} \frac{2(x \arcsin(x) + \sqrt{1-x^2} - 1)}{\pi} & 0 < x < 1 \\ \frac{2(1-x) \arcsin(x-1) - 2\sqrt{x(2-x)} + \pi x}{\pi} & 1 < x < 2. \end{cases}$$