Discrete-Event Simulation:
A First Course

Section 4.3: Continuous-Data Histograms

Section 4.3: Continuous-Data Histograms

- Consider a real-valued sample $\mathcal{S} = \{x_1, x_2, \ldots, x_n\}$
- Data values are generally distinct
- Assume lower and upper bounds a, b

$$a \le x_i < b \qquad\qquad i = 1, 2, \ldots, n$$

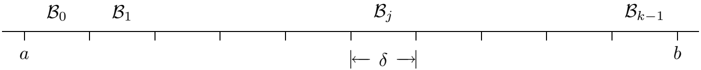
- Defines interval of possible values for random variable X

$$\mathcal{X} = [a, b) = \{x \mid a \le x < b\}$$

Binning

- Partition the interval $\mathcal{X} = [a, b)$ into k equal-width bins

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



- The bins are $\mathcal{B}_0 = [a, a + \delta)$, $\mathcal{B}_1 = [a + \delta, a + 2\delta)$ ...
- Width of each bin is $\delta = (b - a)/k$

Continuous Data Histogram

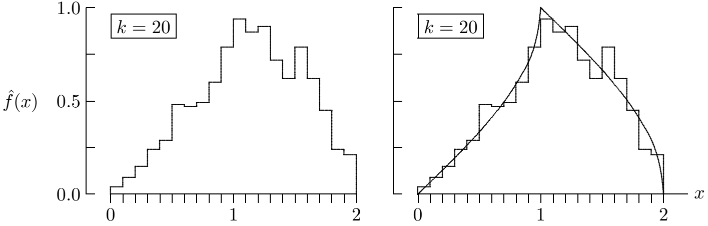
- For each $x \in [a, b)$, there is a unique bin $\mathcal{B}_j$ with $x \in \mathcal{B}_j$
- Estimated *density* of random variable X is

$$\hat{f}(x) = \frac{\text{the number of } x_i \in \mathcal{S} \text{ for which } x_i \in \mathcal{B}_j}{n\,\delta}$$

- Continuous-data histogram: a "bar" plot of $\hat{f}(x)$ versus x
- *Density*: relative frequency normalized via division by δ
- $\hat{f}(x)$ is piecewise constant

Example 4.3.1: buffon

- $n = 1000$ observations of the needle from `buffon`
- Let $a = 0.0$, $b = 2.0$, and $k = 20$ so that $\delta = (b - a)/k = 0.1$



- As $n \to \infty$ and $k \to \infty$ (i.e., $\delta \to 0$), the histogram will converge to the *probability density function*

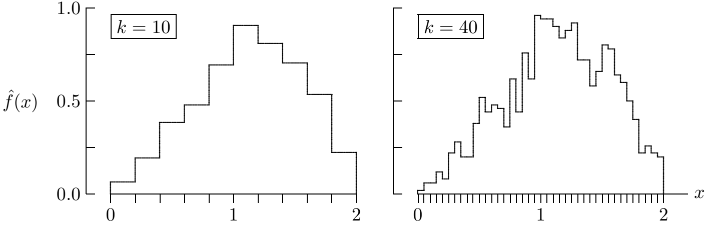
Histogram Parameter Guidelines

- Choose a, b so that few, if any, data points are outliers
- If k is too large (δ is too small), histogram will be "noisy"
- If k is too small (δ is too large), histogram will be too "smooth"
- Keep figure aesthetics in mind
- Typically $\lfloor \log_2(n) \rfloor \le k \le \lfloor \sqrt{n} \rfloor$ with a bias toward

$$k \cong \lfloor (5/3)\sqrt[3]{n} \rfloor$$

Example 4.3.2: Smooth, Noisy Histograms

- $k = 10$ ($\delta = 0.2$) gives perhaps too smooth a histogram
- $k = 40$ ($\delta = 0.05$) gives too noisy a histogram

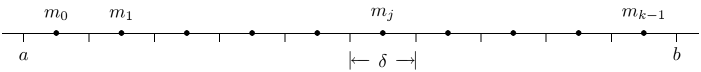


- Guidelines: $9 \leq k \leq 31$ with bias toward $k \cong \lfloor (5/3)\sqrt[3]{1000} \rfloor = 16$
- Note no vertical lines to horizontal axis

Relative Frequency

- Define p_j to be the *relative frequency* of points in bin $\mathcal{B}_j$
- Define the *bin midpoints*

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



- Then $p_j = \delta \hat{f}(m_j)$
- Note that $p_0 + p_1 + \cdots + p_{k-1} = 1$ and $\hat{f}(\cdot)$ has unit area

$$\int_a^b \hat{f}(x)dx = \cdots = \sum_{j=0}^{k-1} p_j = 1$$

Histogram Integrals

- Consider the two integrals

$$\int_a^b x\hat{f}(x)\,dx \qquad \int_a^b x^2\hat{f}(x)\,dx$$

- Because $\hat{f}(\cdot)$ is piecewise constant, integrals become summations

$$\int_a^b x\hat{f}(x)dx = \cdots = \sum_{j=0}^{k-1} m_j p_j$$

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

- Continuous-data histogram mean, standard deviation are defined in terms of these integrals

Histogram Mean and Standard Deviation

- Continuous-data histogram mean and standard deviation:

$$\bar{x} = \int_a^b x\hat{f}(x)\,dx \qquad\qquad s = \sqrt{\int_a^b (x-\bar{x})^2\hat{f}(x)\,dx}$$

- $\bar{x}$ and s can be evaluated *exactly* by summation

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

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

- Some choose to ignore the $\delta^2/12$ term

Quantization Error

- Continuous-data *histogram* $\bar{x}$, s will differ slightly from *sample* $\bar{x}$, s
- *Quantization error* associated with binning of continuous data
- If difference is not slight, a, b, and k (or δ) should be adjusted
- **Example 4.3.3**: 1000-point `buffon` sample

  Let $a = 0.0$, $b = 2.0$, and $k = 20$

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

  Essentially no impact of $\delta^2/12$ term

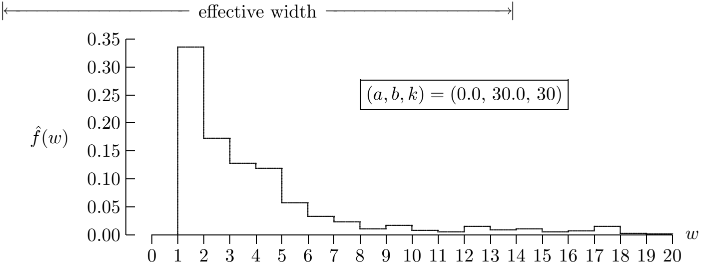
Computational Model: Program cdh

Algorithm 4.3.1

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


Example 4.3.4: Using cdh

- Use cdh to process first $n = 1000$ wait times
- $(a, b, k) = (0.0, 30.0, 30)$



- Histogram $\bar{x} = 4.57$ and $s = 4.65$

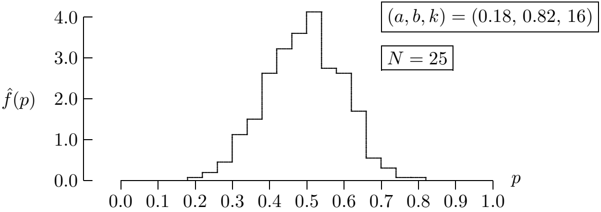
Point Estimation

- Inherent uncertainty in any MC simulation derived estimate
- Four-fold increase in replications yields a two-fold decrease in uncertainty (e.g., craps)
- As $n \to \infty$, a DDH will look like a CDH
- As such, natural to treat the discrete data as continuous to experiment with uncertainty
- You can use `cdh` on discrete data

  You cannot use `ddh` on continuous data

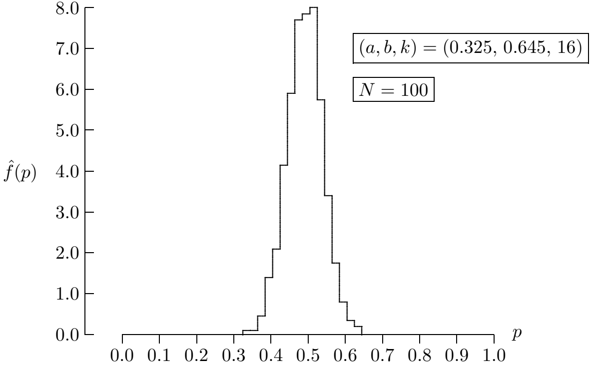
Example 4.3.5: The Square-Root Rule

- $n = 1000$ estimates of craps for $N = 25$ plays



- Note these are *density* estimates, not *relative frequency* estimates
- As $N \to \infty$, histogram will become taller and narrower
- Centered on mean, consistent with $\int_0^1 \hat{f}(p)\, dp = 1$

Example 4.3.5: The Square-Root Rule



- Four-fold increase in N yields two-fold decrease in uncertainty

Random Events, Exponential Inter-Events

- Generate n random events via calls to $\texttt{Uniform}(0, t)$ with $t > 0$
- Sort the event times in increasing order

$$0 < u_1 < u_2 < \cdots < u_n < t$$

- With $u_0 = 0$, define the inter-event times as

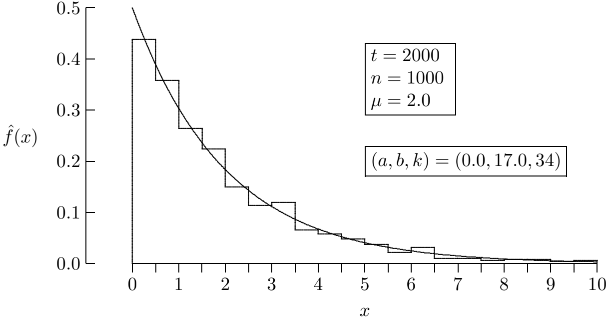
$$x_i = u_i - u_{i-1} \qquad i = 1, 2, \ldots, n$$

- Let $\mu = t/n$ and note that the sample mean is approximately μ

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

A Histogram Of Inter-Event Times

- A histogram of the inter-event times x_i has exponential shape



- Smallest inter-event times are the most likely
- As $n \to \infty$ and $\delta \to 0$, $\hat{f}(x) \to f(x) = (1/\mu)\exp(-x/\mu)$

Empirical Cumulative Distribution Functions

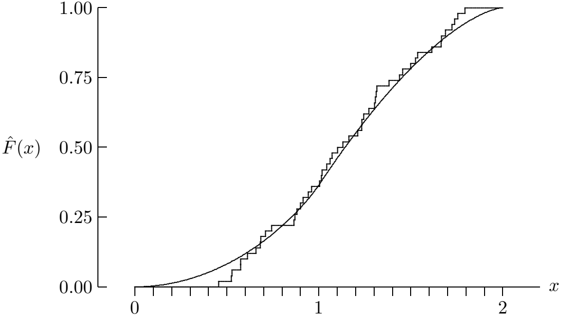
- Drawback of CDH: need to choose k
- Two different choices for k can give quite different histograms
- Estimated cumulative distribution function for random variable X:

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

- *Empirical cumulative distribution function*: plot of $\hat{F}(x)$ versus x
- With an empirical CDF, no parameterization required
- However, must store all the data and then sort

Example 4.3.7: An Empirical CDF

- $n = 50$ observations of the needle from `buffon`



- Upward step of 1/50 for each of the values generated

CDH Versus Empirical CDF

Continuous Data Histogram:

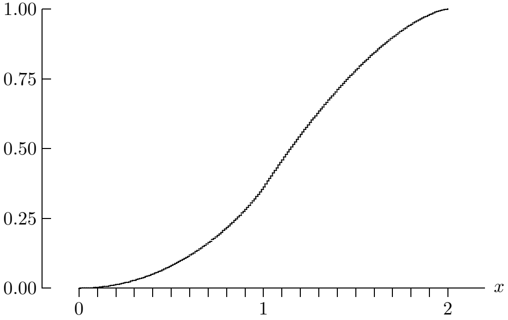
- Superior for detecting *shape* of distribution
- Arbitrary parameter selection is not ideal

Empirical Cumulative Distribution Function:

- Nonparametric, therefore less prone to sampling variability
- Shape is less distinct than that of a CDH
- Requires storing and sorting entire data set
- Often used for statistical "goodness-of-fit" tests

Example 4.3.8: Combining CDH and Empirical CDF

- Increase to $n = 1\,000\,000\,000$ samples from `buffon`
- Use 200 equal-width bins (a la CDH) to create an empirical CDF



- Very smooth curve — close to theoretical CDF