Discrete-Event Simulation:

A First Course

Section 4.1: Sample Statistics

Section 4.1: Sample Statistics

- Simulation involves *a lot* of data
- Must compress the data into meaningful statistics
- Collected data is a *sample* from a much larger *population*
- Two types of statistical analysis:
  1. "Within-the-run"
  2. "Between-the-runs" (replication)
- Essence of statistics: analyze a sample and draw inferences

Sample Mean and Standard Deviation

- Consider a sample $x_1, x_2, \ldots, x_n$ (continuous or discrete)
- *Sample Mean:*

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- *Sample Variance:*

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- *Sample Standard Deviation:* $s = \sqrt{s^2}$
- *Coefficient of Variation:* $s/\bar{x}$

Understanding the Statistics

- Mean: a measure of *central tendency*
- Variance, Deviation: measures of *dispersion* about the mean
- Why variance — easier math (no square root)
- Why standard deviation — same units as data, mean
- Note that the coefficient of variation (C.V.) is unit-less
- But a common shift in data changes the C.V.

  E.g.: measure students' heights on the floor, in chairs

Biased and Unbiased Statistics

- An alternative definition of sample variance:

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \text{rather than} \qquad \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Why the $1/(n-1)$ version?
  - unbiased when data is independent (more in Ch. 8)
  - relates to analysis of variance (degrees of freedom)
- Why the $1/n$ version?
  - if n is large, the difference is irrelevant
  - unbiased property often doesn't apply in simulation
  - the math is easier
- For now, we will use the $1/n$ version

Relating the Mean and Standard Deviation

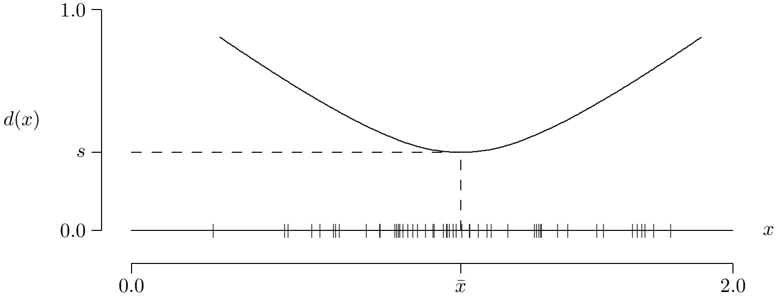
- Consider the root-mean-square (rms) function

$$d(x) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - x)^2}$$

- $d(x)$ measures dispersion about any value x
- The mean $\bar{x}$ gives the smallest possible value for $d(x)$ (Theorem 4.1.1)
- The standard deviation s is that smallest value

Example 4.1.1: Relating $\bar{x}$, s

- 50 samples from program `buffon`



- Here, $\bar{x} \cong 1.095$ and $s \cong 0.354$
- The smallest value of $d(x)$ is $d(\bar{x}) = s$ as shown

Chebyshev's Inequality

- Relates to the number of points that lie within k standard deviations of the mean
- Points farthest from the mean make the most contribution to s
- Define the set $\S_k = \{x_i \mid \bar{x} - ks < x_i < \bar{x} + ks\}$
- Let $p_k = |\S_k|/n$ be the proportion of x_i within $\pm ks$ of $\bar{x}$
- *Chebyshev's Inequality*:

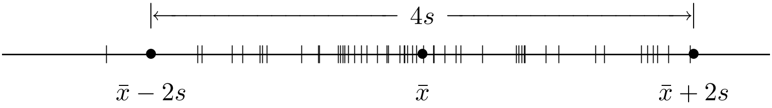
$$p_k \geq 1 - \frac{1}{k^2} \qquad (k > 1)$$

Understanding Chebyshev's Inequality

- For *any* sample, at least 75% of the points lie within $\pm 2s$ of $\bar{x}$
- For $k = 2$, Chebyshev's is very conservative

Typically 95% lie within $\pm 2s$ of $\bar{x}$

- $\bar{x} \pm 2s$ defines the "effective width" of a sample



- Most, but not all, points will lie in this interval
- *Outliers* should be viewed with suspicion

Linear Data Transformations

- Often need to convert to different units after data has been collected
- Let x_i' be the "new data": $x_i' = ax_i + b$
- Sample mean:

$$\bar{x}' = \frac{1}{n}\sum_{i=1}^{n} x_i' = \frac{1}{n}\sum_{i=1}^{n}(ax_i + b) = \frac{a}{n}\left(\sum_{i=1}^{n} x_i\right) + b = a\bar{x} + b$$

- Sample variance:

$$(s')^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i' - \bar{x}')^2 = \cdots = a^2 s^2$$

- Sample standard deviation: $s' = |a|s$

Examples of Linear Data Transformations

- **Example 4.1.2**: Suppose $x_1, x_2, \ldots, x_n$ measured in seconds
  - To convert to minutes, let $x_i' = x_i/60$

$$\bar{x}' = \frac{45}{60} = 0.75 \text{ (minutes)} \qquad s' = \frac{15}{60} = 0.25 \text{ (minutes)}$$

- **Example 4.1.3**: *Standardize* data — subtract $\bar{x}$, divide by s
  - For sample $x_1, x_2, \ldots, x_n$, standardized sample is

$$x_i' = \frac{x_i - \bar{x}}{s} \qquad i = 1, 2, \ldots, n$$

  - Then $\bar{x}' = 0$ and $s' = 1$
  - Used to avoid problems with very large (or small) valued samples

Nonlinear Data Transformations

- Usually involves a Boolean (two-state) outcome
- The *value* of x_i is not as important as the *effect*
- Let $\mathcal{A}$ be a fixed set; then

$$x'_i = \begin{cases} 1 & x_i \in \mathcal{A} \\ 0 & \text{otherwise} \end{cases}$$

- Let p be the proportion of x_i that fall in $\mathcal{A}$

$$p = \frac{\text{the number of } x_i \text{ in } \mathcal{A}}{n}$$

- Then $\bar{x}' = p$ and $s' = \sqrt{p(1-p)}$
- Similar to Bernoulli (see Ch. 6)

Examples of Nonlinear Data Transformations

- **Example 4.1.4**: Single Server Service Node
  - Let $x_i = d_i$ be the delay for job i from SSQ
  - Let $\mathcal{A} = \mathbb{R}^+$; then $x_i' = 1$ iff. $d_i > 0$
  - From Exercise 1.2.3, proportion of jobs delayed is $p = 0.723$
  - Then $\bar{x}' = 0.723$ and $s = \sqrt{(0.723)(0.277)} = 0.448$
- **Example 4.1.2**: Monte Carlo Simulation
  - Estimate a probability by generating a sequence of 0's and 1's
  - The probability estimate p is the ratio of 1's to trials
  - Then $\bar{x} = p$ and $s = \sqrt{p(1-p)}$

Computational Considerations

- Consider the sample standard deviation equation

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Requires two passes through the data
  1. Compute the mean $\bar{x}$
  2. Compute the squared differences about $\bar{x}$
- Must store or re-create the entire sample — bad when n is large

The Conventional One-Pass Algorithm

- A mathematically equivalent, one-pass equation for s^2:

$$\begin{aligned}
s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2\bar{x}x_i + \bar{x}^2) \\
&= \left(\frac{1}{n}\sum_{i=1}^{n}x_i^2\right) - \left(\frac{2}{n}\bar{x}\sum_{i=1}^{n}x_i\right) + \left(\frac{1}{n}\sum_{i=1}^{n}\bar{x}^2\right) \\
&= \left(\frac{1}{n}\sum_{i=1}^{n}x_i^2\right) - 2\bar{x}^2 + \bar{x}^2 \\
&= \left(\frac{1}{n}\sum_{i=1}^{n}x_i^2\right) - \bar{x}^2
\end{aligned}$$

- Round-off error is problematic

Welford's One-Pass Algorithm

- Running sample mean:

$$\bar{x}_i = \frac{1}{i}(x_1 + x_2 + \cdots + x_i)$$

- Running sample sum of squared deviations:

$$v_i = (x_1 - \bar{x}_i)^2 + (x_2 - \bar{x}_i)^2 + \cdots + (x_i - \bar{x}_i)^2$$

- $\bar{x}_i$ and v_i can be computed recursively ($\bar{x}_0 = 0, v_0 = 0$) (Theorem 4.1.2):

$$\begin{aligned} \bar{x}_i &= \bar{x}_{i-1} + \frac{1}{i}(x_i - \bar{x}_{i-1}) \\ v_i &= v_{i-1} + \left(\frac{i-1}{i}\right)(x_i - \bar{x}_{i-1})^2 \end{aligned}$$

- Then $\bar{x}_n$ is the sample mean, v_n/n is the variance

Algorithm 4.1.1: Welford's One-Pass

- No *a priori* knowledge of the sample size n required
- Less prone to accumulated round-off error

Algorithm 1.1.1

```
n = 0;
x̄ = 0.0;
v = 0.0;
while (more data ) {
    x = GetData();
    n++;
    d d = x - x̄;
    v = v + d * d * (n - 1) / n;
    x̄ = x̄ + d / n;
}
s = sqrt(v / n);
return n, x̄, s;
```

- Program uvs implements Welford's algorithm

Example 4.1.6: Using Welford's Algorithm

- Let $x_1, x_2, \ldots, x_n$ be *Uniform(a,b)* random variates
- In the limit as $n \to \infty$

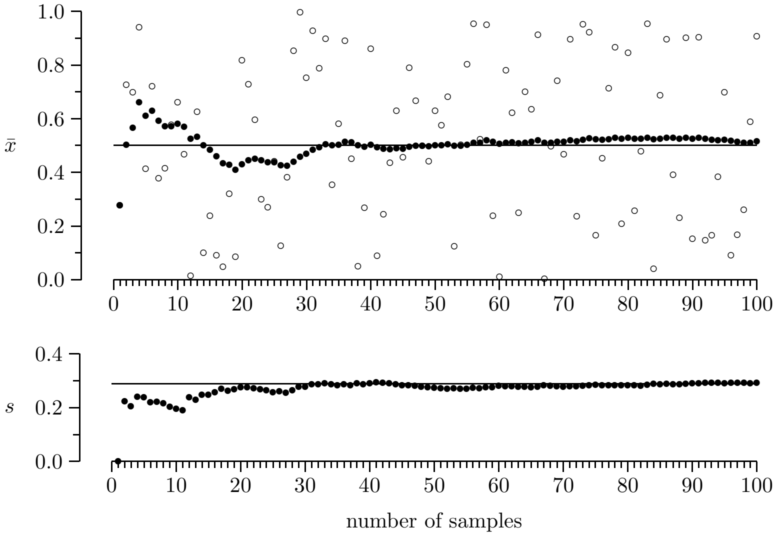
$$\bar{x} \to \frac{a+b}{2} \qquad\qquad s \to \frac{b-a}{\sqrt{12}}$$

- Using *Uniform*(0,1) random variates, $\bar{x}$ and s should converge to

$$\frac{0+1}{2} = 0.5 \qquad\qquad \frac{1-0}{\sqrt{12}} \cong 0.2887$$

- Convergence of $\bar{x}$ and s to theoretical values is not necessarily monotone

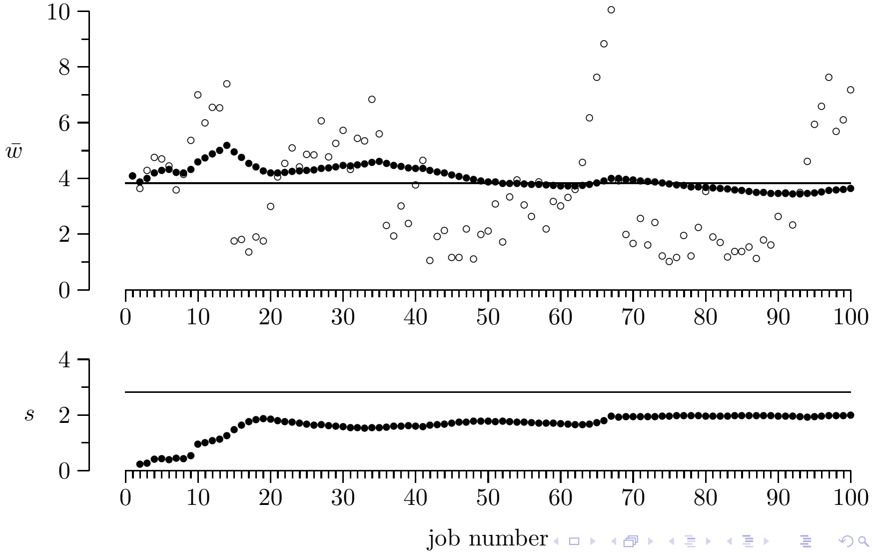
Example 4.1.6: Using Welford's Algorithm



Serial Correlation

- *Independence*: each x_i value does not depend on any other point
- Time-sequenced DES output is typically not independent
- E.g.: wait times of consecutive jobs have positive *serial correlation*
- Independence is appropriate only for Monte Carlo simulation
- **Example 4.1.7**: Consider output from `ssq2`
    - *Exponential*(2) interarrivals, *Uniform*(1,2) service
    - Wait times $w_1, w_2, \ldots, w_{100}$ have high positive serial correlation
    - The correlation produces a *bias* in the standard deviation

Example 4.1.7: Serial Correlation



Time-Averaged Sample Statistics

- Let $x(t)$ be the sample path of a stochastic process for $0 < t < \tau$
- *Sample-path mean*:

$$\bar{x} = \frac{1}{\tau} \int_0^\tau x(t)\, dt$$

- *Sample-path variance*:

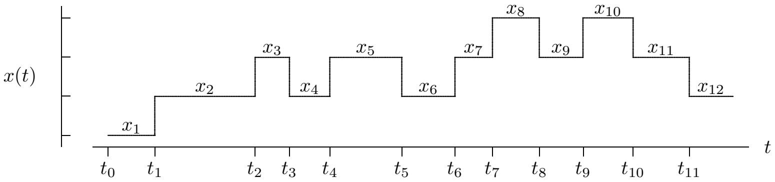
$$s^2 = \frac{1}{\tau} \int_0^\tau \left(x(t) - \bar{x}\right)^2 dt$$

- *Sample-path standard deviation*: $s = \sqrt{s^2}$
- One-pass equation for variance:

$$s^2 = \left(\frac{1}{\tau} \int_0^\tau x^2(t)\, dt \right) - \bar{x}^2$$

Computational Considerations

- For DES, a sample path is *piecewise constant*
- Changes in the sample path occur at *event times* $t_0, t_1, \ldots$



- For computing statistics, integrals reduce to summations

Computational Sample-Path Formulas

Theorem (4.1.3)

Consider a piecewise constant sample path

$$x(t) = \begin{cases} x_1 & t_0 < t \leq t_1 \\ x_2 & t_1 < t \leq t_2 \\ \vdots & \vdots \\ x_n & t_{n-1} < t \leq t_n \end{cases}$$

- *Sample-path mean:*

$$\bar{x} = \frac{1}{\tau} \int_0^\tau x(t)\, dt = \frac{1}{t_n} \sum_{i=1}^{n} x_i\, \delta_i$$

- *Sample-path variance:*

$$s^2 = \frac{1}{\tau} \int_0^\tau \left(x(t) - \bar{x} \right)^2 dt = \frac{1}{t_n} \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 \delta_i = \left(\frac{1}{t_n} \sum_{i=1}^{n} x_i^2\, \delta_i \right) - \bar{x}^2$$

Welford's Sample Path Algorithm

- Based on the definitions

$$\begin{aligned}\bar{x}_i &= \frac{1}{t_i}(x_1\delta_1 + x_2\delta_2 + \cdots + x_i\delta_i)\\ v_i &= (x_1 - \bar{x}_i)^2\delta_1 + (x_2 - \bar{x}_i)^2\delta_2 + \cdots + (x_i - \bar{x}_i)^2\delta_i\end{aligned}$$

- $\bar{x}_i$ is the sample-path mean of $x(t)$ for $t_0 \le t \le t_i$
- v_i/t_i is the sample-path variance
- $\bar{x}_i$ and v_i can be computed recursively ($\bar{x}_0 = 0, v_0 = 0$) (Theorem 4.1.4):

$$\begin{aligned}\bar{x}_i &= \bar{x}_{i-1} + \frac{\delta_i}{t_i}(x_i - \bar{x}_{i-1})\\ v_i &= v_{i-1} + \frac{\delta_i t_{i-1}}{t_i}(x_i - \bar{x}_{i-1})^2\end{aligned}$$