Empirical Distributions
Distributions

• In the discrete case, a probability distribution is just a set of values, each with some probability of occurrence
  – Probabilities don’t change as values occur
  • Example, tossing a coin
• The distribution must cover all possibilities, so there is no probability of a value occurring that is not in the distribution
• An empirical distribution is one for which the values and their associated probabilities are determined by observation or experiment
• Since values may occur in ranges, empirical distributions that are not discrete are also used
Expected Value
Finite Discrete Case

- Finite case
  - Values $v_1, v_2, \ldots, v_n$
  - Probability of occurrence for each $p_1, p_2, \ldots p_n$
    - $\sum p_i = 1$ (intuitively, all possibilities are covered by requiring the sum to be 1)
  - Expected value $E = \sum p_i v_i$
    - Corresponds to the “weighted” sum of the $v_i$
    - where the $p_i$ are the weights
  - The expected value is the average value when each value is equally probable
    - Average $= (\sum v_i)/n = \sum (1/n)v_i$
Expected Value
Infinite Discrete Case

- Infinite, discrete case
  - Values are “countable” (one to one correspondence with the counting numbers 1, 2, 3, …)
  - The infinite sequence of probabilities $p_1$, $p_2$, … still must have the property that $\sum p_i = 1$
- Determined as the limit of the sequence of partial sums; i.e.,
  $$\sum p_i = \lim_{n \to \infty} \sum_{1}^{n} p_i$$
Expected Value – Non-discrete Case

• Generalize from the discrete case by observing a graphical representation

- Area under the curve is 1

• This suggests that for values spanning a range, it is natural to think of the probability of the range as an area

• Uniformly spaced rectangles is how we approximate the value of integrals, the area under a curve
Probability Density Function

- The function describing the curve is called a “probability density function” (pdf)
  - Can assume the pdf takes values over real line from $-\infty$ to $+\infty$ (by letting it be 0, where not otherwise defined)
    - To be a pdf, p must have the property that $\int p(x) = 1$ over all real numbers x (area under the curve is 1)
    - Note that $\int_{a}^{b} p(x) = $ probability of the range of values between a and b
    - A point’s probability is 0, since the area under the curve for it is 0
  - $E = \int p(x) \cdot x \, dx$
    - Corresponds to the weighted sum
## Defining an Empirical Distribution

- **Discrete case**
  - List each value with its corresponding probability
  - **Example**
    | Value | Probability |
    |-------|-------------|
    | $v_1$ - | -2 | 0.3 |
    | $v_2$ - | 4  | 0.2 |
    | $v_3$ - | 5  | 0.15 |
    | $v_4$ - | 7  | 0.35 |
  - Sum of the probabilities must be 1
  - Expected value is
    $$-2 \times 0.3 + 4 \times 0.2 + 5 \times 0.15 + 7 \times 0.35 = 3.4$$
Defining an Empirical Distribution

- **Ranges case**

  - List the start point for each range with the probability of the range

  - Example

<table>
<thead>
<tr>
<th>Value</th>
<th>probability</th>
<th>range represented</th>
<th>cum</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2</td>
<td>0.3</td>
<td>-2 – 4</td>
<td>0.3</td>
</tr>
<tr>
<td>4</td>
<td>0.2</td>
<td>4 – 5</td>
<td>0.5</td>
</tr>
<tr>
<td>5</td>
<td>0.15</td>
<td>5 – 7</td>
<td>0.65</td>
</tr>
<tr>
<td>7</td>
<td>0.35</td>
<td>7 – 10</td>
<td>1.00</td>
</tr>
<tr>
<td>10</td>
<td>0.35</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

  - Note that 4 ranges requires 5 values

  - Expected value is

\[
1 \times 0.3 + 4.5 \times 0.2 + 6 \times 0.15 + 8.5 \times 0.35 = 5.075
\]

  - Use the midpoint (average value) of each range
Sampling from Empirical Distributions

• Algorithms used will be covered later

• ExtendSim procedure
  – Select Empirical table
  – Select the sampling method
    • Stepped for sampling by linear interpolation (values drawn uniformly across each of the ranges)
    • Discrete for the sampling on step (discrete values)
  • Enter the data
    – *Time* and probability for a Create block
    – Enter times in the *Value* column
Warning

• If you switch between discrete and stepped, you will need to adjust your data
  – Stepped sampling requires an extra entry to terminate the final interval and has the final interval probability repeated