Slides 10: Input Modeling

Purpose and Overview

• Input models provide the driving force for a simulation model.

• The quality of the output is no better than the quality of the inputs.

• We’ll discuss:
  − Collect data from the real system.
  − Identify a probability distribution to represent the input process.
  − Choose parameters for the distribution.
  − Evaluate the chosen distribution and parameters for the goodness of fit.
Data Collection

- One of the biggest tasks in solving a real problem, GIGO, garbage-in-garbage-out.

- Suggestions that may enhance and facilitate data collection:
  - Plan ahead: begin by a practice of pre-observing session, watch for unusual circumstances.
  - Analyze the data as it is being collected: check adequacy.
  - Combine homogeneous data sets, e.g., successive time periods, during the same time period on successive days.
  - Be aware of data censoring: the quantity is not observed in its entirety, danger of leaving out long process times.
  - Check for relationship between variables, e.g., build scatter plots.
  - Check for autocorrelation.
  - Collect input data, not performance data.
Identifying the Distribution

- Histograms.
- Selecting families of distributions.
- Parameter estimation.
- Goodness-of-fit tests.
- Fitting a non-stationary process.
Histograms

• A frequency distribution or histogram is useful in determining the shape of a distribution.

• The number of class intervals depends on:
  – The number of observations.
  – The dispersion of the data.
  – A suggestion: the square root of the sample size.

• For continuous data:
  – Corresponds to the probability density function of a theoretical distribution.

• For discrete data:
  – Corresponds to the probability mass function.

• If few data points are available: combine adjacent cells to eliminate the ragged appearance of the histogram.
Vehicle arrival example: number of vehicles arriving at an intersection between 7 am and 7:05 am was monitored for 100 random workdays.
Selecting the Family of Distributions

- A family of distributions is selected based on:
  - The context of the input variable.
  - Shape of the histogram.
- Frequently encountered distributions:
  - Easier to analyze: exponential, normal and Poisson.
  - Harder to analyze: beta, gamma and Weibull.
Selecting the Family of Distributions

- Use the physical basis of the distribution as a guide, for example:
  - Binomial: number of success in $n$ trials.
  - Poisson: number of independent events that occur in a fixed amount of time or space.
  - Normal: distribution of a process that is the sum of a number of component processes.
  - Exponential: time between independent events, or a process time that is memoryless.
  - Weibull: time to failure for components.
  - Discrete or continuous uniform: Each outcome equally likely.
  - Triangular: a process for which only the minimum, most likely, and maximum values are known.
  - Empirical: resamples from the actual data collected.
Selecting the Family of Distributions

• Remember the physical characteristics of the process:
  – Is the process naturally discrete or continuous valued?
  – Is it bounded?

• No ‘true’ distribution for any stochastic input process.

• Goal: obtain a good approximation.
Quantile-Quantile Plots

- Q-Q plot is a useful tool for evaluating distribution fit.

- If $X$ is a random variable with cdf $F$, then the $q$-quantile of $X$ is the $\gamma$ such that $F(\gamma) = P(X \leq \gamma) = q$, for $0 < q < 1$.
  - When $F$ has an inverse $\gamma = F^{-1}(q)$.

- Let $\{x_i, i = 1, 2, \ldots, n\}$ be a sample of data from $X$ and $\{y_j, j = 1, 2, \ldots, n\}$ be the observations in ascending order. Then $y_j$ is approximately $F^{-1}\left(\frac{j-0.5}{n}\right)$ where $j$ is the ranking or order number.
Quantile-Quantile Plots

- The plot of $y_j$ versus $F^{-1}\left(\frac{j-0.5}{n}\right)$ is:
  - Approximately a straight line if $F$ is a member of an appropriate family of distributions.
  - The line has slope 1 if $F$ is a member of an appropriate family of distributions with appropriate parameter values.
Quantile-Quantile Plots

- Example: Check whether the door installation times follow a normal distribution:

  - The observations are ordered from smallest to largest, and the $y_j$ are plotted verses $F^{-1}\left(\frac{j-0.5}{n}\right)$ where $F$ has a normal distribution with the sample mean (99.99 sec) and sample variance (0.2832 sec$^2$).

<table>
<thead>
<tr>
<th>j</th>
<th>Value</th>
<th>j</th>
<th>Value</th>
<th>j</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>99.55</td>
<td>6</td>
<td>99.98</td>
<td>11</td>
<td>100.26</td>
</tr>
<tr>
<td>2</td>
<td>99.56</td>
<td>7</td>
<td>100.02</td>
<td>12</td>
<td>100.27</td>
</tr>
<tr>
<td>3</td>
<td>99.62</td>
<td>8</td>
<td>100.06</td>
<td>13</td>
<td>100.33</td>
</tr>
<tr>
<td>4</td>
<td>99.65</td>
<td>9</td>
<td>100.17</td>
<td>14</td>
<td>100.41</td>
</tr>
<tr>
<td>5</td>
<td>99.79</td>
<td>10</td>
<td>100.23</td>
<td>15</td>
<td>100.47</td>
</tr>
</tbody>
</table>
Quantile-Quantile Plots

- Example (continued): Check whether the door installation times follow a normal distribution (straight line supports the hypothesis of a normal distribution, and this is overlayed on the histogram).
Quantile-Quantile Plots

- Consider the following while evaluating the linearity of a $q - q$ plot:
  - The observed values never fall exactly on a straight line.
  - The ordered values are ranked and hence not independent, unlikely for the points to be scattered about the line.
  - Variance of the extremes is higher than the middle. Linearity of the points in the middle of the plot is more important.

- Q-Q plot can also be used to check homogeneity:
  - Check whether a single distribution can represent both sample sets.
  - Plotting the order values of the two data samples against each other.
Parameter Estimation

- Next step after selecting a family of distributions.

- If observations in a sample of size $n$ are $X_1, X_2, \ldots, X_n$ (discrete or continuous), the sample mean and sample variance are:

\[
\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}, \quad S^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{n - 1}
\]

- If the data are discrete and have been grouped in a frequency distribution, then:

\[
\bar{X} = \frac{\sum_{j=1}^{n} f_j X_j}{n}, \quad S^2 = \frac{\sum_{j=1}^{n} f_j X_j^2 - n\bar{X}^2}{n - 1}
\]

- Here $f_j$ is the observed frequency of value $X_j$. 
Parameter Estimation

- When raw data are unavailable (data are grouped into class intervals), the approximate sample mean and variance are:

\[
\bar{X} = \frac{\sum_{j=1}^{c} f_j m_j}{n}, \quad S^2 = \frac{\sum_{j=1}^{c} f_j m_j^2 - n\bar{X}^2}{n - 1}
\]

- Here \(f_j\) is the observed frequency in the \(j\)-th class interval with mid-point \(m_j\) and \(c\) is the number of class intervals.

- A parameter is an unknown constant, but an estimator is a statistic.
Parameter Estimation

• Vehicle Arrival Example (continued): Assume \( n = 100, \ f_1 = 12, \ X_1 = 0, \ f_2 = 10, \ X_2 = 1 \) and so on with \( \sum_{j=1}^{k} f_j X_j = 364 \) and \( \sum_{j=1}^{k} f_j X_j^2 = 2080 \)

• The sample mean and variance are \( \bar{X} = \frac{364}{100} = 3.64 \) and \( S^2 = \frac{(2080 - 100(3.64)^2)}{99} = 7.63. \)

• The histogram suggests \( X \) has a Poisson distribution (but note the sample mean is not equal to the sample variance, but each estimator is a random variable and is hence not perfect).
Goodness-of-Fit Tests

• Conducting hypothesis testing on input data distribution using:
  – Kolmogorov-Smirnov test.
  – Chi-square test.

• No single correct distribution in a real application exists:
  – If very little data are available, it is unlikely to reject any candidate distribution.
  – If a lot of data are available, it is likely to reject all candidate distributions.
Chi-Square Test

- Intuition: comparing the histogram of the data to the shape of the candidate density or mass function.
- Valid for large sample sizes when parameters are estimated by maximum likelihood.
- By arranging the $n$ observations into a set of $k$ class intervals or cells, the test statistic is:

$$\chi^2_0 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

- Here $O_i$ is the observed frequency while $E_i$ is the expected frequency which is $E_i = np_i$ where $p_i$ is the theoretical probability of the $i$-th interval (suggested minimum is 5).
- The test-statistic approximately follows the chi-square distribution with $k - s - 1$ degrees of freedom where $s$ is the number of parameters of the hypothesized distribution estimated by the sample statistics.
Chi-Square Test

• The hypothesis of a chi-square test is:
  – \( H_0 \): The random variable \( X \) conforms to the distributional assumption with the parameter(s) given by the estimate(s).
  – \( H_1 \): The random variable \( X \) does not conform.

• If the distribution tested is discrete and combining adjacent cells is not required (so that \( E_i \) is bigger than the minimum requirement):
  – Each value of the random variable should be a class interval, unless combining is necessary, and \( p_i = p(x_i) = P(X = x_i) \).
Chi-Square test

• If the distribution tested is continuous:

\[ p_i = \int_{a_{i-1}}^{a_i} f(x) \, dx = F(a_i) - F(a_{i-1}) \]

– Here \( a_{i-1} \) and \( a_i \) are the endpoints of the \( i \)-th class interval and \( f(x) \) is the assumed pdf, while \( F(x) \) is the assumed cdf.

• Recommended number of class intervals (\( k \)) [Caution: Different grouping of data, \( i.e., \), \( k \), can affect the hypothesis testing result].

<table>
<thead>
<tr>
<th>Sample Size, ( n )</th>
<th>Number of Class Intervals, ( k )</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>Do not use the chi-square test</td>
</tr>
<tr>
<td>50</td>
<td>5 to 10</td>
</tr>
<tr>
<td>100</td>
<td>10 to 20</td>
</tr>
<tr>
<td>&gt; 100</td>
<td>( n^{1/2} ) to ( n/5 )</td>
</tr>
</tbody>
</table>
Chi-Square test

- Vehicle Arrival Example (continued): $H_0$ is the random variable is Poisson distributed, with $H_1$ being the random variable is not Poisson Distributed:

\[
E_i = np(x_i) = \frac{ne^{-\lambda} \lambda^x}{x_i!}
\]

- Values 0 and 1 are combined because of minimum $E_i$ value, as are values 7 through > 11.
- Degree of freedom is $k - s - 1 = 7 - 1 - 1 = 5$, hence hypothesis is rejected at the 0.05 level of significance: $\chi^2_0 = 27.68 > \chi^2_{0.05,5} = 11.1$. 

Chi-Square test

<table>
<thead>
<tr>
<th>x_i</th>
<th>Observed Frequency, O_i</th>
<th>Expected Frequency, E_i</th>
<th>((O_i - E_i)^2 / E_i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>12</td>
<td>2.6</td>
<td>7.87</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>9.6</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>19</td>
<td>17.4</td>
<td>0.15</td>
</tr>
<tr>
<td>3</td>
<td>17</td>
<td>21.1</td>
<td>0.8</td>
</tr>
<tr>
<td>4</td>
<td>19</td>
<td>19.2</td>
<td>4.41</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>14.0</td>
<td>2.57</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>8.5</td>
<td>0.26</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>4.4</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>2.0</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>0.8</td>
<td>11.62</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>&gt; 11</td>
<td>1</td>
<td>0.1</td>
<td></td>
</tr>
</tbody>
</table>

| Σ    | 100                     | 100.0                   | 27.68                    |
Kolmogorov-Smirnov Test

- Intuition: formalize the idea behind examining a q-q plot.

- Recall from slides 3:
  - The test compares the continuous cdf $F(x)$ of the hypothesized distribution with the empirical cdf $S_N(x)$ of the $N$ sample observations.
  - Based on the maximum difference statistics $D = \max |F(x) - S_N(x)|$.

- A more powerful test, particularly useful when:
  - Sample sizes are small.
  - No parameters have been estimated from the data.

- When parameter estimates have been made.
p-values and ‘Best Fits’

• p-values for the test statistics:
  – The significance level at which one would reject $H_0$ for the given test statistic value.
  – A measure of fit, the larger the better.
  – Large p-value means good fit.
  – Small p-value means poor fit.

• Vehicle Arrival Example (continued):
  – $H_0$: data is Poisson.
  – Test statistic $\chi^2_0 = 27.68$ with 5 degrees of freedom.
  – p-value is 0.00004 meaning we would reject $H_0$ with 0.00004 significance level, hence Poisson is a poor fit.
p-values and ‘Best Fits’

• Many software use p-values as the ranking measure to automatically determine the ‘best fit’. Things to be cautious about:
  – Software may not know about the physical basis of the data, distribution families it suggests may be inappropriate.
  – Close conformance to the data does not always lead to the most appropriate input model.
  – p-value does not say much about where the lack of fit occurs.

• Recommended: always inspect the automatic selection using graphical methods.
Fitting a Non-stationary Poisson Process

• Fitting a NSPP to arrival data is difficult, possible approaches are:
  – Fit a very flexible model with lots of parameter or,
  – (Our focus) Approximate constant arrival rate over some basic interval of time, but vary it from time interval to time interval.

• Suppose we need to model arrivals over time $[0, T]$, our approach is the most appropriate when we can:
  – Observe the time period repeatedly and,
  – Count arrivals/record arrival times.
The estimated arrival rate during the \( i \)-th time period is:

\[
\hat{\lambda}(t) = \frac{1}{n\Delta t} \sum_{j=1}^{n} C_{ij}
\]

Here \( n \) is the number of observation periods, \( \Delta t \) is the time interval length, \( C_{ij} \) is the number of arrivals during the \( i \)-th time interval on the \( j \)-th observation period.

For example, divide a 10 hour business day (8am to 6pm) into equal intervals \( k = 20 \) whose length \( \Delta t = 1/2 \) and observe over \( n = 3 \) days.

On next slide, the second time period has estimated arrival rate of \((1/3)(1/2)(23 + 26 + 32) = 54\) arrivals per hour.
## Fitting a Non-stationary Poisson Process

<table>
<thead>
<tr>
<th>Time Period</th>
<th>Number of Arrivals</th>
<th>Estimated Arrival Rate (arrivals/hr)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Day 1</td>
<td>Day 2</td>
</tr>
<tr>
<td>8:00 - 8:00</td>
<td>12</td>
<td>14</td>
</tr>
<tr>
<td>8:30 - 9:00</td>
<td>23</td>
<td>26</td>
</tr>
<tr>
<td>9:00 - 9:30</td>
<td>27</td>
<td>18</td>
</tr>
<tr>
<td>9:30 - 10:00</td>
<td>20</td>
<td>13</td>
</tr>
</tbody>
</table>
Selecting Model without Data

- If data is not available, some possible sources to obtain information about the process are:
  - Engineering data: often product or process has performance ratings provided by the manufacturer or company rules specify time or production standards.
  - Expert opinion: people who are experienced with the process or similar processes, often they can provide optimistic, pessimistic, and most-likely times, and they may know the variability as well.
  - Physical or conventional limitations: physical limits on performance, limits or bounds that narrow the range of the input process.
  - The nature of the process.

- The uniform, triangular, and beta distributions are often used as input models.
Selecting Model without Data

- Example: Production planning simulation.
  - Input of sales volume of various products is required, salesperson of product XYZ says that:
    * No fewer than 1000 units and no more than 5000 units will be sold.
    * Given her experience, she believes there is a 90\% chance of selling more than 2000 units and a 25\% chance of selling more than 3000 units, and only a 1\% chance of selling more than 4000 units.
  - Translating these information into a cumulative probability of being less than or equal to those goals for simulation input.
## Selecting Model without Data

<table>
<thead>
<tr>
<th>$i$</th>
<th>Interval (Sales)</th>
<th>Cumulative Frequency, $c_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$1000 \leq x \leq 2000$</td>
<td>0.10</td>
</tr>
<tr>
<td>2</td>
<td>$2000 &lt; x \leq 3000$</td>
<td>0.75</td>
</tr>
<tr>
<td>3</td>
<td>$3000 &lt; x \leq 4000$</td>
<td>0.99</td>
</tr>
<tr>
<td>4</td>
<td>$4000 &lt; x \leq 5000$</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Multivariate and Time-Series Input Models

• Multivariate:
  – For example, lead time and annual demand for an inventory model, increase in demand results in lead time increase, hence variables are dependent.

• Time-series:
  – For example, time between arrivals or orders to buy and sell stocks, buy and sell orders tend to arrive in bursts, hence times between arrivals are dependent.
Covariance and Correlation

- Consider the model that describes the relationship between $X_1$ and $X_2$:
  \[(X_1 - \mu) = \beta(X_2 - \mu_2) + \epsilon\]

- Here epsilon is a random variable with mean 0 and is independent of $X_2$.
  - $\beta = 0$ means $X_1$ and $X_2$ are statistically independent.
  - $\beta > 0$ means $X_1$ and $X_2$ tend to be above or below their means together.
  - $\beta < 0$ means $X_1$ and $X_2$ tend to be on opposite sides of their means.

- Covariance between $X_1$ and $X_2$:
  \[\text{Cov}(X_1, X_2) = \mathbb{E}[(X_1 - \mu_1)(X_2 - \mu_2)] = \mathbb{E}[X_1 X_2] - \mu_1 \mu_2\]

- Here $\text{Cov}(X_1, X_2) = 0$ if $\beta = 0$, $\text{Cov}(X_1, X_2) < 0$ if $\beta < 0$, and $\text{Cov}(X_1, X_2) > 0$ if $\beta > 0.$
Covariance and Correlation

• Correlation between $X_1$ and $X_2$ (values between -1 and 1):

$$
\rho = \text{Corr}(X_1, X_2) = \frac{\text{Cov}(X_1, X_2)}{\sigma_1 \sigma_2}
$$

• Here $\text{Corr}(X_1, X_2) = 0$ if $\beta = 0$, $\text{Corr}(X_1, X_2) < 0$ if $\beta < 0$, and $\text{Corr}(X_1, X_2) > 0$ if $\beta > 0$.

• The closer $\rho$ is to -1 or 1, the stronger the linear relationship is between $X_1$ and $X_2$. 
Covariance and Correlation

- A time series is a sequence of random variables $X_1, X_2, \ldots$ that are identically distributed (same mean and variance) but dependent.
  - $\text{Cov}(X_t, X_{t+h})$ is the lag-$h$ autocovariance.
  - $\text{Corr}(X_t, X_{t+h})$ is the lag-$h$ autocorrelation.
  - If the autocovariance value depends only on $h$ and not on $t$, the time series is covariance stationary.
Multivariate Input Models

- If $X_1$ and $X_2$ are normally distributed, dependence between them can be modeled by the bivariate normal distribution with $\mu_1$, $\mu_2$, $\sigma_1^2$, $\sigma_2^2$ and correlation $\rho$.
  - To estimate $\mu_1$, $\mu_2$, $\sigma_1^2$, $\sigma_2^2$ see earlier notes on parameter estimation.
  - To estimate $\rho$, suppose we have $n$ independent and identically distributed pairs $(X_{11}, X_{21}), (X_{12}, X_{22}), \ldots, (X_{1n}, X_{2n})$:

$$
\hat{\text{Cov}}(X_1, X_2) = \frac{1}{n-1} \sum_{j=1}^{n} (X_{1j} - \hat{X}_1)(X_{2j} - \hat{X}_2) = \frac{1}{n-1} \left( \sum_{j=1}^{n} X_{1j}X_{2j} - n\hat{X}_1\hat{X}_2 \right)
$$

$$
\hat{\rho} = \frac{\hat{\text{Cov}}(X_1, X_2)}{\hat{\sigma}_1 \hat{\sigma}_2}
$$
Time-Series Input Models

- If $X_1, X_2, X_3, \ldots$ is a sequence of identically distributed, but dependent and covariance-stationary random variables, then we can represent the process as follows:
  - Autoregressive order-1 model AR(1).
  - Exponential autoregressive order-1 model, EAR(1).
  * Both have characteristics that:

\[
\rho_h = \text{Corr}(X_t, X_{t+h}) = \rho^h \text{ for } h = 1, 2, \ldots
\]

  * Lag-$h$ autocorrelation decreases geometrically as the lag increases, hence observations far apart in time are nearly independent.
AR(1) Time-Series Input Models

- Consider the time-series model:
  \[ X_t = \mu + \phi(X_{t-1} - \mu) + \epsilon_t \text{ for } t = 2, 3, \ldots \]

- Here \( \epsilon_2, \epsilon_3, \ldots \) are i.i.d. normally distributed with \( \mu_{\epsilon} = 0 \) and variance \( \sigma_\epsilon^2 \).

- If \( X_1 \) is chosen appropriately, then:
  - \( X_1, X_2, \ldots \) are normally distributed with mean \( \mu \) and variance \( \sigma^2/(1 - \phi^2) \).
  - Autocorrelation \( \rho_h = \phi^h \).

- To estimate \( \phi, \mu, \sigma_\epsilon^2 \):
  \[
  \hat{\mu} = \bar{X}, \hat{\sigma}_\epsilon^2 = \hat{\sigma}^2(1 - \hat{\phi}^2), \hat{\phi} = \frac{\hat{\text{Cov}}(X_t, X_{t+1})}{\hat{\sigma}^2}
  \]
Consider the time-series model:

\[ X_t = \phi X_{t-1} \quad \text{with probability } \phi, \phi X_{t_1} + \epsilon_t \quad \text{otherwise } t = 2, 3, \ldots \]

Here \( \epsilon_2, \epsilon_3, \ldots \) are i.i.d. exponentially distributed with \( \mu_\epsilon = 1/\lambda \) and \( 0 \leq \phi < 1 \).

If \( X_1 \) is chosen appropriately, then:

- \( X_1, X_2, \ldots \) are exponentially distributed with mean \( 1/\lambda \).
- Autocorrelation \( \rho_h = \phi^h \), and only positive correlation is allowed.

To estimate \( \phi, \lambda \):

\[ \hat{\lambda} = 1/\bar{X}, \hat{\phi} = \hat{\rho} = \frac{\hat{Cov}(X_t, X_{t+1})}{\hat{\sigma}^2} \]
Summary

• We have considered the 4 steps in developing input data models:
  – Collecting the raw data.
  – Identifying the underlying statistical distribution.
  – Estimating the parameters.
  – Testing for goodness of fit.