Input Modeling

• The input data is the driving force for the simulation - the behavior of the simulation and all the results/conclusions that can be reached depend on appropriate inputs.
Developing a Model of Input Data

- Is there time to collect enough data?
- Are there other sources available to obtain relevant information?

- Start with a histogram of data to enable visualization.
- Is anything known about process?

- Valid data is especially important for this step
- $\chi^2$ and Kolmogorov-Smirnov tests

1. Collect data from real system
2. Identify a probability distribution to model observed data
3. Choose parameters for a specific instance of distribution family
4. Evaluate goodness-of-fit

Repeat?
Data Collection

- Data collection can be the hardest tasks in solving a real problem
- Data collection is one of the most important and hardest tasks in simulation
  - Data is often either scarce or overly abundant

- “GIGO” (Garbage-In = Garbage-Out) often applies
  - The simulation often abstracts real data, hiding its inadequacies

- Suggestions to improve data collection:
  - PLAN! Do some trial runs to see if there are any special circumstances that will have to be captured
  - Analyze/summarize data during collection - this might highlight a problem with data being collected
  - Look for homogeneity with plan to combine similar data sets
  - Watch for data censoring - is an observation of a process complete? Or are the long procedures truncated artificially?
  - Use a scatter plot to see relationships between variables - other senses help, as well
  - Look for correlation in data
  - Distinguish input data (independent variables) from performance data (dependent variables)
Identifying the Distribution

- Methods to identify distributions:
  - Histograms
  - Select the Family of Distributions
  - Quantile-quantile plots
Histograms

- Bins for individual observations
- Range of observed values
- Height proportional to number of observations
Histograms

Height proportional to number of observations

Range of observed values

Bins for individual observations

Number of bins ~ sqrt(observations)
Histograms - Picking the Number of Bins

\[ \mu := 2 \quad N := 1000 \]
\[ \sigma := 2.5 \quad R := \text{norm}(N, \mu, \sigma) \]

\[ M_1 := 5 \]
\[ h_1 := \text{histogram}(M_1, R) \]
\[ k_1 := 0..M_1 - 1 \]

\[ M_2 := \text{floor}(\sqrt{N}) \]
\[ h_2 := \text{histogram}(M_2, R) \]
\[ k_2 := 0..M_2 - 1 \]

\[ M_3 := 100 \]
\[ h_3 := \text{histogram}(M_3, R) \]
\[ k_3 := 0..M_3 - 1 \]

- All Mathcad examples are in the file “Chapter9.mcd”
Meaning of the Histogram for Input Modeling

- Scaled by the total number of points, the histogram approximates the p.d.f. of the input distribution

- Use the histogram to visualize the p.d.f. of the observed data to enable selection of a known distribution function
Selecting the Family of the Distribution

• There are a large number of probability distributions
  – generated from observations of the real world
  – proposed as theoretical models

• What is known about the physical characteristics of the input process?
  – Is it naturally discrete or continuous valued?
  – Are the observable values inherently bounded or is there no natural bound?
  – Can you infer a distribution from what you know about the process that generates input values?
    • E.g., Normal (Gaussian) process is derived from the sum of a large number of independent random variables
    • E.g., Erlang process is sum of several exponential processes
    • E.g., Lognormal process is derived from product of several component processes
    • E.g., Poisson process models the number of independent events that occur in a bounded period of time or area in space.
Quantile-Quantile Plots

Given a set of input data, $R$
Calculate $\mu$ and $\sigma^2$
Sort $R$ ($S$ in this Mathcad file)
Generate a set of $\gamma_j$, evenly distributed between 0 and 1
For a given assumed distribution (using calculated $\mu$ and $\sigma^2$), calculate the inverse of the c.d.f. for each $\gamma$
Plot the sorted data vs. the calculated values

If the assumed distribution matches, the plot should be a straight line with slope=1, intercept=0
Q-Q Plots with Incorrect Assumptions

- Generating Exponentially distributed random numbers
- Matching a Normal distribution

- The resulting plot is not linear, as expected.

- If you have Mathcad, try experimenting with other distributions, parameters

Using the wrong distribution:

\[
N := 1000 \\
R := \text{rexp}(N, \mu) \\
S := \text{sort}(R) \\
\text{mean}(R) = 0.504 \\
\text{var}(R) = 0.24 \\
j := 1..N \\
\gamma_j := \frac{j - 1}{N} \\
\text{INV}_j := \text{qnorm}(\gamma_j, \text{mean}(R), \text{var}(R))
\]
Parameter Estimation

- Mean and variance:
  - n samples:

\[
\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}
\]

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

- Discrete data, grouped in k frequency distribution classes:

\[
\bar{X} = \frac{\sum_{j=1}^{k} f_j X_j}{n}
\]

\[
S^2 = \frac{\sum_{j=1}^{k} f_j X_j^2 - n\bar{X}^2}{n-1}
\]
Parameter Estimation

• Mean and variance:
  – Continuous data in $c$ frequency classes when raw data is not available. $m_j$ are the midpoints of the frequency classes:

$$
\bar{X} = \frac{\sum_{j=1}^{c} f_j m_j}{n}
$$

$$
S^2 = \frac{\sum_{j=1}^{c} f_j m_j^2 - n\bar{X}^2}{n-1}
$$
Parameter Estimation

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Parameter(s)</th>
<th>Suggested Estimator(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson</td>
<td>$\alpha$</td>
<td>$\hat{\alpha} = \bar{X}$</td>
</tr>
<tr>
<td>Exponential</td>
<td>$\lambda$</td>
<td>$\hat{\lambda} = \frac{1}{\bar{X}}$</td>
</tr>
</tbody>
</table>
| Normal       | $\mu, \sigma^2$ | $\hat{\mu} = \bar{X}$  
|              |              | $\hat{\sigma^2} = S^2$ |

- The estimated parameters using the suggested estimators are maximum-likelihood estimators, based on raw data.
- The true parameters, assuming that the distribution was known, are not expected to be the same as the experimentally measured parameters
  - small sample size
  - noise, randomness in measurements
Parameter Estimation

\[ \mu := 3.5 \]
\[ \sigma := 2 \]
\[ N_{\text{norm}} := 50 \]
\[ R_{\text{norm}} := \text{rnorm}(N_{\text{norm}}, \mu, \sigma) \]
\[ \text{mean}(R_{\text{norm}}) = 3.883 \]
\[ \sqrt{\text{var}(R_{\text{norm}})} = 1.653 \]

\[ \lambda := .1 \]
\[ N_{\text{exp}} := 50 \]
\[ R_{\text{exp}} := \text{rexp}(N_{\text{exp}}, \frac{1}{\lambda}) \]
\[ \text{mean}(R_{\text{exp}}) = 0.093 \]
\[ \sqrt{\text{var}(R_{\text{exp}})} = 0.138 \]

These should be equal

• Distinguish the parameter of the distribution: \( \alpha \)

• From the estimator or statistic: \( \hat{\alpha} \)
Goodness-of-Fit

• Hypothesis testing was introduced 2 weeks ago to test random number distributions
  – Kolmogorov-Smirnov
  – \( \chi^2 \)

• Goodness-of-fit tests should *guide* the choice of a distribution, not *establish* it: there is often no perfect answer with real-world data.

• Sample size can have a significant effect on results:
  – With a small number of data points, few or no candidate distributions will be rejected
  – With a large number of data points, almost all candidate distributions will be rejected.
Goodness-of-Fit Tests

• $\chi^2$ test:
  – Compares the histogram of candidate density function
  – Valid for large sample sizes
  – Assumes parameters are estimated by maximum likelihood function

• $\chi^2$ test with equal probabilities:
  – If distribution is assumed to be continuous, class intervals should be equal probability, rather than equal width. Example 9.14 (next slide)

• Kolmogorov-Smirnov:
  – With $\chi^2$ test, grouping of data may influence accept/reject decision
  – K-S test is based on examining a q-q plot
  – Especially useful when no parameters have been estimated from data
Example 9.14 - 
χ² Test for Exponential Distribution

- The data, X is generated from an exponential distribution, and the χ² test is applied.

- Here, χ₀² = 14.32, which exceeds the tabulated value for a significance of 0.05, but not at a significant of 0.01 - we might reject this distribution if the level of significance were tight enough.
\( p \)-Value and “Best Fits”

- In the prior example, with the particular data examined, the hypothesis that the data came from an exponential distribution would have been rejected at a significance level of 0.05, but not at a significance level of 0.01.

- How should you choose a significance level? 0.01, 0.05, & 0.10 are commonly used.

- The significance level is equivalent to the probability of falsely rejecting \( H_0 \)

- Many software packages compute a \( p \)-value:
  - The \( p \)-value is the significance level which just rejects \( H_0 \)
  - The \( p \)-value can be viewed as a measure of fit: larger \( p \)-value indicates a better fit
  - One possible approach to choosing a distribution:
    - Test every distribution available, choose the one that has largest \( p \)-value
Selecting Input Models without Data

• What about where there is no input data available to model, e.g., for a preliminary study when no time or funds are available to gather data?

• Creating data where none exists:
  – base it on published performance data
  – get expert opinion for the same or similar systems - this might help to bound or otherwise characterize inputs
  – use physical limitations to bound problem (e.g., maximum possible car arrivals at an intersection is related to minimum car spacing and maximum velocity)
  – The nature of the process: use the descriptions of various distributions to pick the one most closely related to the underlying input process

• Uniform, triangular, and beta distributions are used when nothing else is known
Multivariate and Time Series Input Models

• Previous discussion dealt with independent processes
• What if multiple inputs to the simulation are related to each other?
  – multivariate input models with a fixed, finite number of random variables
  – time-series input models of a sequence of related random variables

• There are two measures of dependence between random variables:
  – Covariance
  – Correlation
Multivariate and Time Series Input Models

• $X_1$ and $X_2$ are two random variables:
  - $\mu_i = E(X_i)$ and $\sigma^2_i = \text{Var}(X_i)$

• Covariance and correlation define how well the relationship between $X_1$ and $X_2$ is described by:
  
  $$ (X_1 - \mu_1) = \beta (X_2 - \mu_2) + \epsilon $$

• $\epsilon$ is a zero mean R.V., independent of $X_2$

• Covariance:
  
  $$ \text{cov}(X_1, X_2) = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E(X_1X_2) - \mu_1\mu_2 $$

• Correlation:
  
  $$ \rho = \text{corr}(X_1, X_2) = \frac{\text{cov}(X_1, X_2)}{\sigma_1\sigma_2} $$

  $$ -1 \leq \text{corr}(X_1, X_2) \leq 1 $$

  $$ -\infty \leq \text{cov}(X_1, X_2) \leq \infty $$


Multivariate and Time Series Input Models

• If $X_1, X_2, \ldots$ are identically distributed (but possibly dependent), this is referred to as a time-series.

• $\text{cov}(X_t, X_{t+h})$ and $\text{corr}(X_t, X_{t+h})$ are the lag-$h$ covariance and correlation

• If $\text{cov}(X_t, X_{t+h})$ depends on $h$ and not $t$, the time series is covariance-stationary. For such a situation, it must also be the case that $\text{corr}(X_t, X_{t+h})$ depends on $h$ and not $t$. In this case, we can leave $t$ out of the expression and:

$$\rho_h = \text{corr}(X_t, X_{t+h})$$
Comments on Correlation

$-1 \leq \rho = \text{corr}(X_1, X_2) \leq 1$

- Correlation is a relative measure on a continuous scale
  - $|\rho| < .1$ is generally considered uncorrelated
  - $|\rho| > .33$ is often considered correlated
More Comments on Correlation

\[ \rho_X(\tau) = \text{corr}(X_t, X_{t+\tau}) \]

- \( \rho_X(\tau) \) is referred to as the Autocorrelation of \( X \)

- \( \rho_X(\tau) \) is a measure of periodic or non-independent behavior of \( X \)
Guidance for Course Projects

• Project report should (at least) address:
  – **Introduction/background**: What was the problem you studied?
  – **Assumptions**: How did you develop your model?
  – **Observations of physical system**: Representative input data you collected
  – **Simulation program**: Listing of all code needed to build your simulation
  – **Simulation results**: Representative simulation execution outputs
  – **Validation/Verification**: What leads you to believe that your model & simulation are meaningful?
  – **Conclusions**: What did your simulation show?
  – **Recommendations**: How would you modify the physical system or how you use it to improve performance?
  – **Future work**: What follow on activities would be appropriate if you were to continue this simulation work?
  – **References**: Cite previous results or information you based your work on
Homework

• Ch. 9 exercises 16, 19