Chapter 9
Input Modeling

Banks, Carson, Nelson & Nicol

Discrete-Event System Simulation

Purpose & Overview

- Input models provide the driving force for a simulation model.
- The quality of the output is no better than the quality of inputs.
- In this chapter, we will discuss the 4 steps of input model development:
  - 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 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 or 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 diagram
  - Check for autocorrelation
  - Collect input data, not performance data

3

Identifying the Distribution

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

4

Histograms

[Identifying the distribution]

- 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
  - Suggested: 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

Histograms

[Identifying the distribution]

- Vehicle Arrival Example: # of vehicles arriving at an intersection between 7 am and 7:05 am was monitored for *100* random workdays.

Arrivals per Period	Frequency
0	12
1	10
2	19
3	17
4	10
5	8
6	7
7	5
8	5
9	3
10	3
11	1

- There are ample data, so the histogram may have a cell for each possible value in the data range

Selecting the Family of Distributions

[Identifying the distribution]

- 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

[Identifying the distribution]

- Use the physical basis of the distribution as a guide, for example:
  - Binomial: # of successes in *n* trials
  - Poisson: # of independent events that occur in a fixed amount of time or space
  - Normal: dist'n 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: models complete uncertainty
  - 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

[Identifying the distribution]

- 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

9

Quantile-Quantile Plots

[Identifying the distribution]

- *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 γ such that

$$F(\gamma) = P(X \le \gamma) = q, \qquad \text{for } 0 < q < 1$$

  - 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:

$$y_j \text{ is approximately } F^{-1}\left(\frac{j - 0.5}{n}\right)$$

where j is the ranking or order number

10

Quantile-Quantile Plots

[Identifying the distribution]

- The plot of y_j versus $F^{-1}((j-0.5)/n)$ 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

11

Quantile-Quantile Plots

[Identifying the distribution]

- Example: Check whether the door installation times follows a normal distribution.
  - The observations are now ordered from smallest to largest:

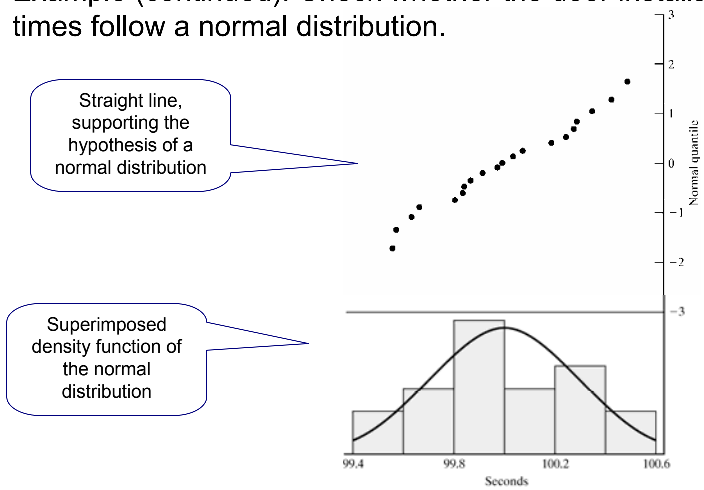
j	Value	j	Value	j	Value
1	99.55	6	99.98	11	100.26
2	99.56	7	100.02	12	100.27
3	99.62	8	100.06	13	100.33
4	99.65	9	100.17	14	100.41
5	99.79	10	100.23	15	100.47

  - y_j are plotted versus $F^{-1}((j-0.5)/n)$ where F has a normal distribution with the sample mean *(99.99 sec)* and sample variance (0.2832^2 sec^2)

12

Quantile-Quantile Plots [Identifying the distribution]

- Example (continued): Check whether the door installation times follow a normal distribution.



Quantile-Quantile Plots [Identifying the distribution]

- 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

[Identifying the distribution]

- 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 variance are:

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

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

$$\bar{X} = \frac{\sum_{j=1}^{n} f_j X_j}{n} \qquad S^2 = \frac{\sum_{j=1}^{n} f_j X_j^2 - n\bar{X}^2}{n-1}$$

where f_j is the observed frequency of value X_j

15

Parameter Estimation

[Identifying the distribution]

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

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

where f_j is the observed frequency of in the jth class interval
m_j is the midpoint of the jth interval, and c is the number of class intervals

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

16

Parameter Estimation [Identifying the distribution]

- Vehicle Arrival Example (continued): Table in the histogram example on slide 6 (Table 9.1 in book) can be analyzed to obtain:

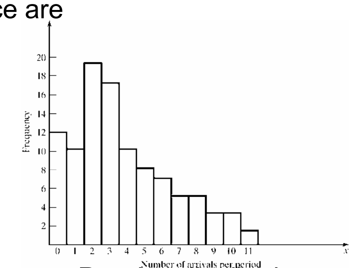
$$n = 100, f_1 = 12, X_1 = 0, f_2 = 10, X_2 = 1, \ldots,$$

$$\text{and } \sum_{j=1}^{k} f_j X_j = 364, \text{ and } \sum_{j=1}^{k} f_j X_j^2 = 2080$$

  - The sample mean and variance are

$$\bar{X} = \frac{364}{100} = 3.64$$

$$S^2 = \frac{2080 - 100 * (3.64)^2}{99} = 7.63$$



  - The histogram suggests X to have a Possion distribution
    - However, note that sample mean is not equal to sample variance.
    - Reason: each estimator is a random variable, is not perfect.

17

Goodness-of-Fit Tests [Identifying the distribution]

- Conduct 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 distributions
  - If a lot of data are available, it is likely to reject all candidate distributions

18

Chi-Square test [Goodness-of-Fit Tests]

- 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 statistics is:

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Observed Frequency

Expected Frequency
$E_i = n^*p_i$
where p_i is the theoretical prob. of the ith interval.
Suggested Minimum = 5

which **approximately** follows the chi-square distribution with *k-s-1* degrees of freedom, where s = # of parameters of the hypothesized distribution estimated by the sample statistics.

19

Chi-Square test [Goodness-of-Fit Tests]

- 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 cell is not required (so that E_i > minimum requirement):
  - Each value of the random variable should be a class interval, unless combining is necessary, and

$$p_i = p(x_i) = P(X = x_i)$$

20

Chi-Square test [Goodness-of-Fit Tests]

- If the distribution tested is continuous:

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

where a_i-1 and a_i are the endpoints of the i^{th} class interval and $f(x)$ is the assumed pdf, $F(x)$ is the assumed cdf.

- Recommended number of class intervals (k):

Sample Size, n	Number of Class Intervals, k
20	Do not use the chi-square test
50	5 to 10
100	10 to 20
> 100	$n^{1/2}$ to n/5

- Caution: Different grouping of data (i.e., k) can affect the hypothesis testing result.

21

Chi-Square test [Goodness-of-Fit Tests]

- Vehicle Arrival Example (continued):

H_0: the random variable is Poisson distributed.

H_1: the random variable is not Poisson distributed.

x_i	Observed Frequency, O_i	Expected Frequency, E_i	$(O_i - E_i)^2/E_i$
0	12	2.6	7.87
1	10	9.6	
2	19	17.4	0.15
3	17	21.1	0.8
4	19	19.2	4.41
5	6	14.0	2.57
6	7	8.5	0.26
7	5	4.4	11.62
8	5	2.0	
9	3	0.8	
10	3	0.3	
> 11	1	0.1	
	100	100.0	27.68

$$E_i = np(x) = n\frac{e^{-\alpha}\alpha^x}{x!}$$

Combined because of min E_i

- Degree of freedom is k-s-1 = 7-1-1 = 5, hence, the hypothesis is rejected at the 0.05 level of significance.

$$\chi_0^2 = 27.68 > \chi_{0.05,5}^2 = 11.1$$

22

Kolmogorov-Smirnov Test

[Goodness-of-Fit Tests]

- Intuition: formalize the idea behind examining a *q-q* plot
- Recall from Chapter 7.4.1:
  - 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 (Tabulated in A.8):

$$D = max| F(x) - S_N(x)|$$

- 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:
  - Critical values in Table A.8 are biased, too large.
  - More conservative, i.e., smaller Type I error than specified.

23

p-Values and "Best Fits"

[Goodness-of-Fit Tests]

- *p-value* for the test statistics
  - The significance level at which one would just reject H_0 for the given test statistic value.
  - A measure of fit, the larger the better
  - Large *p-value*: good fit
  - Small *p-value*: poor fit
- Vehicle Arrival Example (cont.):
  - H_0: data is Possion
  - Test statistics: $\chi_0^2 = 27.68$, with *5* degrees of freedom
  - *p-value = 0.00004*, meaning we would reject H_0 with *0.00004* significance level, hence Poisson is a poor fit.

24

p-Values and "Best Fits"

[Goodness-of-Fit Tests]

- Many software use *p-value* 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.

25

Fitting a Non-stationary Poisson Process

- Fitting a NSPP to arrival data is difficult, possible approaches:
  - Fit a very flexible model with lots of parameters or
  - Approximate constant arrival rate over some basic interval of time, but vary it from time interval to time interval. **Our focus**
- 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.

26

Fitting a Non-stationary Poisson Process

- The estimated arrival rate during the *i*th time period is:

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

where n = # of observation periods, Δt = *time interval length*

C_{ij} = # of arrivals during the i^{th} time interval on the j^{th} observation period

- Example: Divide a *10*-hour business day [*8am,6pm*] into equal intervals $k = 20$ whose length $\Delta t = ½$, and observe over n =3 days

Time Period	Number of Arrivals Day 1	Day 2	Day 3	Estimated Arrival Rate (arrivals/hr)
8:00 - 8:00	12	14	10	24
8:30 - 9:00	23	26	32	54
9:00 - 9:30	27	18	32	52
9:30 - 10:00	20	13	12	30

For instance,
1/3(0.5)*(23+26+32)
= 54 arrivals/hour

27

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 option: 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.

28

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 *1,000* units and no more than *5,000* units will be sold.
    - Given her experience, she believes there is a *90%* chance of selling more than *2,000* units, a *25%* chance of selling more than *2,500* units, and only a *1%* chance of selling more than *4,500* units.
  - Translating these information into a cumulative probability of being less than or equal to those goals for simulation input:

i	Interval (Sales)	Cumulative Frequency, c_i
1	$1000 \le x \le 2000$	0.10
2	$2000 < x \le 3000$	0.75
3	$3000 < x \le 4000$	0.99
4	$4000 < x \le 5000$	1.00

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 of orders to buy and sell stocks, buy and sell orders tend to arrive in bursts, hence, times between arrivals are dependent.

Covariance and Correlation

[Multivariate/Time Series]

- Consider the model that describes relationship between X_1 and X_2:

$$(X_1 - \mu_1) = \beta(X_2 - \mu_2) + \varepsilon$$

ε is a random variable with mean *0* and is independent of X_2

  - $\beta = 0$, X_1 and X_2 are statistically independent
  - $\beta > 0$, X_1 and X_2 tend to be above or below their means together
  - $\beta < 0$, X_1 and X_2 tend to be on opposite sides of their means

- Covariance between X_1 and X_2 :

$$\mathrm{cov}(X_1, X_2) = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E(X_1 X_2) - \mu_1 \mu_2$$

  - where $\quad cov(X_1, X_2) \begin{cases} = 0, \\ < 0, \\ > 0, \end{cases} \quad$ then $\beta \begin{cases} = 0 \\ < 0 \\ > 0 \end{cases}$

31

Covariance and Correlation

[Multivariate/Time Series]

- Correlation between X_1 and X_2 (values between *-1* and *1*):

$$\rho = \mathrm{corr}(X_1, X_2) = \frac{\mathrm{cov}(X_1, X_2)}{\sigma_1 \sigma_2}$$

  - where $\quad corr(X_1, X_2) \begin{cases} = 0, \\ < 0, \\ > 0, \end{cases} \quad$ then $\beta \begin{cases} = 0 \\ < 0 \\ > 0 \end{cases}$
  - The closer ρ is to *-1* or *1*, the stronger the linear relationship is between X_1 and X_2.

32

Covariance and Correlation

[Multivariate/Time Series]

- A time series is a sequence of random variables $X_1, X_2, X_3, \ldots$, are identically distributed (same mean and variance) but dependent.
  - cov(X_t, X_{t+h}) is the *lag-h autocovariance*
  - 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

[Multivariate/Time Series]

- 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 ρ
  - To Estimate $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$, see "Parameter Estimation" (slide 15-17, Section 9.3.2 in book)
  - To Estimate ρ, suppose we have n independent and identically distributed pairs (X_{11}, X_{21}), (X_{12}, X_{22}), … (X_{1n}, X_{2n}), then:

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

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

$$\hat{\rho} = \frac{\hat{\text{cov}}(X_1, X_2)}{\hat{\sigma}_1\hat{\sigma}_2}$$

Sample deviation

Time-Series Input Models

[Multivariate/Time Series]

- 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 the characteristics that:

$$\rho_h = corr(X_t, X_{t+h}) = \rho^h, \quad \text{for } h = 1,2,\ldots$$

    - *Lag-h* autocorrelation decreases geometrically as the lag increases, hence, observations far apart in time are nearly independent

35

AR(1) Time-Series Input Models

[Multivariate/Time Series]

- Consider the time-series model:

$$X_t = \mu + \phi(X_{t-1} - \mu) + \varepsilon_t, \quad \text{for } t = 2,3,\ldots$$

where $\varepsilon_2, \varepsilon_3, \ldots$ are i.i.d. normally distributed with $\mu_\varepsilon = 0$ and variance σ_ε^2

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

$$\hat{\mu} = \bar{X}, \qquad \hat{\sigma}_\varepsilon^2 = \hat{\sigma}^2(1 - \hat{\phi}^2), \qquad \hat{\phi} = \frac{\hat{\mathrm{cov}}(X_t, X_{t+1})}{\hat{\sigma}^2}$$

where $\hat{\mathrm{cov}}(X_t, X_{t+1})$ is the *lag*-1 autocovariance

36

EAR(1) Time-Series Input Models

[Multivariate/Time Series]

- Consider the time-series model:

$$X_t = \begin{cases} \phi X_{t-1}, & \text{with probability } \phi \\ \phi X_{t-1} + \varepsilon_t, & \text{with probability } 1\text{-}\varphi \end{cases} \qquad \text{for } t = 2,3,...$$

where $\varepsilon_2, \varepsilon_3, \ldots$ are i.i.d. exponentially distributed with $\mu_\varepsilon = 1/\lambda$, and $0 \le \phi < 1$

- 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 ϕ, λ:

$$\hat{\lambda} = 1/\bar{X} \ , \qquad \hat{\phi} = \hat{\rho} = \frac{\hat{\mathrm{cov}}(X_t, X_{t+1})}{\hat{\sigma}^2}$$

where $\hat{\mathrm{cov}}(X_t, X_{t+1})$ is the *lag*-1 autocovariance

Summary

- In this chapter, we described 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