# 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

5

## 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

6

## 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

7

## 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

8

## 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



## 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 $\gamma$ such that

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

  - 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

## 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



## 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 |

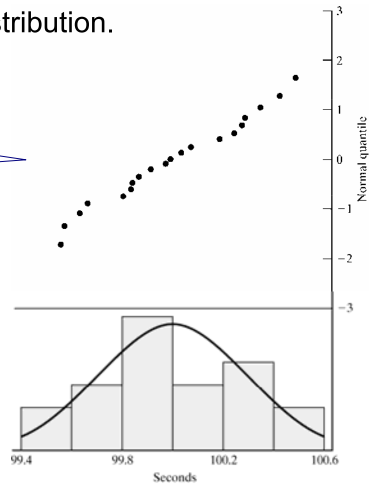
  - $y_j$ are plotted versus $F^{-1}((j-0.5)/n)$ where $F$ has a normal distribution with the sample mean *(99.99 sec)* and sample variance ($0.2832^2$ $sec^2$)

## Quantile-Quantile Plots [Identifying the distribution]

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

Straight line, supporting the hypothesis of a normal distribution

Superimposed density function of the normal distribution




13

## 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

14

## 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 $j$th class interval
$m_j$ is the midpoint of the $j$th 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:

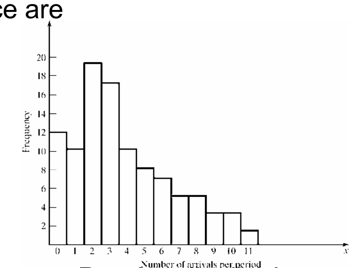
$$n = 100, f_1 = 12, X_1 = 0, f_2 = 10, X_2 = 1, \ldots,$$

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

  - 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 $i$th 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.

20

# 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.

.

## 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