Chapter 9
Input Modeling

Input Modeling

- **In real-world simulation applications, determining appropriate distributions for input data is a major task.**
- **Input data provide the driving force for a simulation model.**
  - A queuing system: distributions of inter-arrivals and services
  - An inventory system: distributions of demand & lead time.
  - A reliability system: distribution of time-to-failure.

Steps in the development of a useful model of input data

1. Collect data from the real system of interest.
   - Sometimes, it is not possible to collect data (e.g. limited resources). In this case, domain knowledge plays vital role to make educated guess.
2. Identify probability distribution to represent the input process to choose a family of distributions.
   - Development of a frequency distribution or histogram.
3. Choose parameters to determine a specific instance of the distribution family.
4. Evaluate the chosen distribution and parameters for goodness-of-fit.
   - May be done informally (graphically) or formally (K-S / Chi-^{2}).

Data Collection

- **One of the most important and difficult problems in simulation.**
- **Data may not be recorded properly that is directly useful for simulation.**
- **Example 9.1: Laundromat problem**
  - Interarrival-time distribution may not be homogeneous.
  - Service times: various service combinations
    - # of washers, # of dryers, user stays or leaves, ..
  - Breakdown and repair times.

Recommendations for Data Collection

- **Planning**
- **Analyze the data as they are being collected.**
- **Combine homogeneous data sets.**
- **The quantity of data may not be observed in its entirety.**
- **Check if there is a relationship between two variables.**
- **Observation may process autocorrelation.**
- **Separate input from output or performance data.**

Identifying the Distribution with Data

- Histograms
  - An important step is to determine the number of intervals.
  - A heuristic: the # of class intervals approximately equal to the square root of the sample size.
- Steps:
  - Divide the range of data into intervals.
  - Label the x-axis based on the intervals.
  - Determine the frequency of occurrences within each interval.
  - Label the y-axis based on the frequency.
  - Plot the frequencies on the y-axis.
- Examples 9.2 & 9.3

Selecting the Family of Distributions

- **The purpose of preparing a histogram is to** infer **a known pdf or pmf.**
- **A family of distributions is selected based on the context being investigated and the shape of the histogram.**
- **See chapter 5 slides on distributions.**

Quantile-Quantile (Q-Q) Plots

- What are quantiles?
  - Median: the middle value that divides the set into two equal parts.
  - Quartiles (Q_1, Q_2, Q_3): Those values that divide the set into four equal parts.
  - Deciles (D_1, D_2, …, D_9): The values that divide the data into 10 equal parts.
  - Percentiles (P_1, P_2, …, P_{99}): The values that divide the data into 100 equal parts.
  - Collectively, quartiles, deciles, percentiles, and other values obtained by equal subdivisions of the data are called quantiles.

Q-Q Plots (cont'd)

- Histograms are useful for selecting a family of distributions.
- Problems with histograms:
  - Not as useful for evaluating the **fit** of the chosen distribution. The fit depends on the widths of the intervals.
  - Small # of data points, a histogram is ragged.
- A q-q plot is useful for evaluating distribution fit that does not suffer these problems.

Q-Q Plots (cont'd)

- The q-q plot is a graphical technique for determining if two data sets come from populations with a common distribution.
  - If A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line.
- A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.
- If F is a member of an approximate family of distributions, the plot is approximately a straight line.

Q-Q Plots (cont'd)

- If X is a RV with cdf F, then the q-quantile of X is that value r s.t. $F(r) = P(X = r) = q$, for $0 < q < 1$. When F has an inverse, $r = F^{-1}(q)$.
- How:
  - {let x_i, i = 1, 2, …, n} be a sample of data from X.
  - Order the observations from the smallest to the largest and denote these as {y_j, j = 1, 2, …, n}, $y_1 = y_2 = \ldots = y_n$.
  - Let j denote the order number. j =1 for the smallest and j = n for the largest.
  - $F(X_i) = i / n$. But for the purpose of probability plotting, it turns to be inconvenient to have $F(X_i) = 1$ for a finite value of x. So, we use the following empirical cdf
    - $F(X_i)\sim = F(X_i) - 0.5/n = (i - 0.5) / n$
  - Yi is approximately $F^{-1}((j - 0.5) / 5)$

Parameter Estimation

- After a family of distributions has been selected, the next step is to estimate the parameters of the distribution.
- Sample mean and sample variance are used to estimate the parameters (such as population mean and variance) of a hypothesized distribution.
- Sample mean X bar – (9.1)
- Unbiased sample variance, S^2 – (9.2)
- Based on frequency – (9.3) & (9.4)
- If data are missing, use midpoints – (9.5) & (9.6)

Suggested Estimators

- Numerical estimates of the distribution parameters are needed to reduce the family of distributions to a specific distribution and to test the resulting hypothesis.
- Table 9.3 lists suggested estimators for distributions often used in simulation.
- Examples 9.7-9.12

Goodness-of-Fit Tests

- Goodness-of fit tests provide helpful guidance for evaluating the suitability of a potential input model.
- Important to understand the effect of sample size.
  - Very little data: a goodness-of-fit test is unlikely to reject any candidate distribution.
  - A lot of data: a goodness-of-fit test will likely reject all candidate distributions.
- Kolmogorov-Smirnov and chi-square tests are applied to hypotheses about distributional forms of input data.

Chi-Square Test

- The test formalizes the idea of comparing the histogram of the data to the shape of the candidate density or mass function.
- Valid of large sample sizes, for both discrete and continuous distributions.
- Procedure:
  - Arrange the n observations into k class intervals
  - Test statistic: $?_0 = S^k (O_i - E_i)2 / E_i$
  - Perform hypothesis test:
    - Ho: the random variable, X, conforms the distributional assumption with the parameter(s) given by the parameter estimates(s). The degrees of freedom is k-s-1, where k is the # of intervals and s is the # of estimated parameters.
  - Example 9.13

Chi-Square Test with Equal Probabilities

- For continuous distributional assumption, class intervals that are equal in probability should be used.
- An important issue is to determine the # of intervals, k. See Table 9.5.
- Solve $F(a_i)$ in terms of p, where a_i represents the endpoint of the i_{th} interval and p is the equal probability.
- Calculate a_i's
- Perform chi-square test.

Kolmogorov-Smirnov Goodness-of-Fit Test

- K-S goodness-of-fit test formalizes the idea behind examining a q-q plot.
- K-S test also overcome some problems with the chi-square test.
- Problems with the chi-square g-o-f test:
  - Requires the data be placed in class intervals. Changing the # of classes & the interval width affects the value.
  - A hypothesis may be accepted when data are grouped one way but rejected another way.
  - The estimation of parameters from the data results in decrease in the degrees of freedom.
  - Valid for large sample sizes.
- Example 9.15

Selecting Input Models without Data

- In practice, simulation model is often built before any data are available.
- Ways to obtain information in this case:
  - Engineering data: values obtained from company rules, industry standards … are used as central value as a starting point.
  - Expert opinions: based on similar products or processes.
  - Physical or conventional limitations: obvious limits or bounds for the input process.
  - The nature of process: the feature of the distributions can be used to justify a particular choice.
- Example 9.16

Multivariate Input Models

- Variables may be related. The relationship should be taken into consideration.
- Examples:
  - An increase in demand results in an increase in lead time.
  - Network traffic often arrives in bursts.
  - An increase in delay results in an increase in packet losses.

Covariance and Correlation

- Covariance & Correlation of 2 RVs: how well the relationship between X_1 and X_2 or linear dependence between X_1 and X_2.
  - cov $(X_1, X_2) = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E(X_1 X_2) - \mu_1 \mu_2$
  - *corr* $(X_1, X_2) =$ cov $(X_1, X_2) / s_1 s_2$
  - If cov $(X_1, X_2) = 0$, X_1 and X_2 are uncorrelated.
  - The closer *corr* (X_1, X_2) is to –1 or 1, the stronger the linear relationship is between X_1 and X_2.
  - Example 9.19.

Summary

- Input data collection and analysis require major commitments.
- Unreliable inputs may lead to incorrect outputs and faulty recommendations.
- Input modeling steps:
  - Collect data
  - Hypothesize a statistical model: histogram
  - Estimate parameters: sample mean & sample var.
  - Test distributional hypothesis
    - q-q plot, chi square, K-S
    - When a distributional assumption is rejected, another distribution is tried. When all else fails, the empirical distribution may be used.