unit 6: INPUT MODELING

6. INPUT MODELING

- Input data provide the driving force for a simulation model. In the simulation of a queuing system, typical input data are the distributions of time between arrivals and service times.
- For the simulation of a reliability system, the distribution of time-to=failure of a component is an example of input data.

There are four steps in the development of a useful model of input data:

- Collect data from the real system of interest. This often requires a substantial time and resource commitment. Unfortunately, in some situations it is not possible to collect data
- Identify a probability distribution to represent the input process. When data are available, this step typically begins by developing a frequency distribution, or histogram, of the data.
- Choose parameters that determine a specific instance of the distribution family. When data are available, these parameters may be estimated from the data.
- Evaluate the chosen distribution and the associated parameters for good-of- fit. Goodness-of-fit may be evaluated informally via graphical methods, or formally via statistical tests. The chisquare and the Kolmo-gorov-Smirnov tests are standard goodness-of-fit tests. If not satisfied that the chosen distribution is a good approximation of the data, then the analyst returns to the second step, chooses a different family of distributions, and repeats the procedure. If several iterations of this procedure fail to yield a fit between an assumed distributional form and the collected data

6.1 Data Collection

- Data collection is one of the biggest tasks in solving real problem. It is one of the most important and difficult problems in simulation. And even if when data are available, they have rarely been recorded in a form that is directly useful for simulation input modeling.

The following suggestions may enhance and facilitate data collection, although they are not all – inclusive.

1. A useful expenditure of time is in planning. This could begin by a practice or pre observing session. Try to collect data while pre-observing.
2. Try to analyze the data as they are being collected. Determine if any data being collected are useless to the simulation. There is no need to collect superfluous data.
3. Try to combine homogeneous data sets. Check data for homogeneity in successive time periods and during the same time period on successive days.
4. Be aware of the possibility of data censoring, in which a quantity of interest is not observed in its entirety. This problem most often occurs when the analyst is interested in the time required to complete some process (for example, produce a part, treat a patient, or have a component fail), but the process begins prior to, or finishes after the completion of, the observation period.
5. To determine whether there is a relationship between two variables, build a scatter diagram.
6. Consider the possibility that a sequence of observations which appear to be independent may possess autocorrelation. Autocorrelation may exist in successive time periods or for successive customers.
7. Keep in mind the difference between input data and output or performance data, and be sure to collect input data. Input data typically represent the uncertain quantities that are largely beyond the control of the system and will not be altered by changes made to improve the system.

6.2 Identifying the Distribution with Data.

- In this section we discuss methods for selecting families of input distributions when data are available.

6.2.1 Histogram

- A frequency distribution or histogram is useful in identifying the shape of a distribution. A histogram is constructed as follows:

1. Divide the range of the data into intervals (intervals are usually of equal width;

however, unequal widths however, unequal width may be used if the heights of the frequencies are adjusted).

2. Label the horizontal axis to conform to the intervals selected.
3. Determine the frequency of occurrences within each interval.
4. Label the vertical axis so that the total occurrences can be plotted for each interval.
5. Plot the frequencies on the vertical axis.

- If the intervals are too wide, the histogram will be coarse, or blocky, and its shape and other details will not show well. If the intervals are too narrow, the histogram will be ragged and will not smooth the data.
- The histogram for continuous data corresponds to the probability density function of a theoretical distribution.

Example 6.2 : The number of vehicles arriving at the northwest corner of an intersection in a 5 min period between 7 A.M. and 7:05 A.M. was monitored for five workdays over a 20-week period. Table shows the resulting data. The first entry in the table indicates that there were 12:5 min periods during which zero vehicles arrived, 10 periods during which one vehicles arrived, and so on,

Table 6:1 Number of Arrivals in a 5 Minute period

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

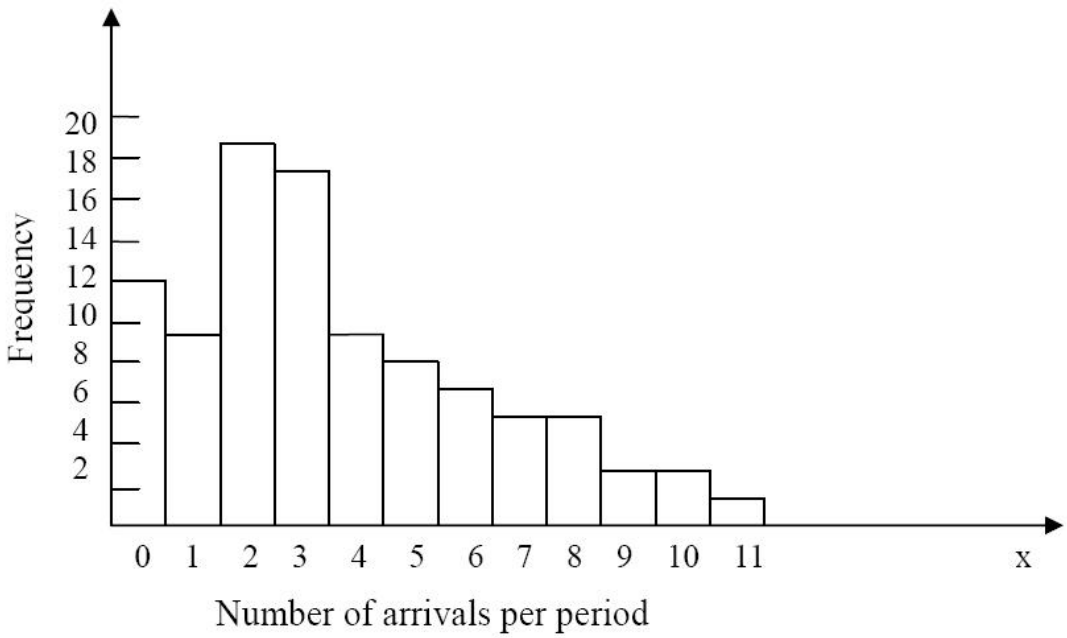

Fig 6.2 Histogram of number of arrivals per period.

6.2.2 Selecting the Family of Distributions

- Additionally, the shapes of these distributions were displayed. The purpose of preparing histogram is to infer a known pdf or pmf. A family of distributions is selected on the basis of what might arise in the context being investigated along with the shape of the histogram.

- Thus, if interarrival-time data have been collected, and the histogram has a shape similar to the pdf in Figure 5.9.the assumption of an exponential distribution would be warranted.

- Similarly, if measurements of weights of pallets of freight are being made, and the histogram appears symmetric about the mean with a shape like that shown in Fig 5.12, the assumption of a normal distribution would be warranted.

- The exponential, normal, and Poisson distributions are frequently encountered and are not difficult to analyze from a computational standpoint. Although more difficult to analyze, the gamma and Weibull distributions provide array of shapes, and should not be overlooked when modeling an underlying probabilistic process. Perhaps an exponential distribution was assumed, but it was found not to fit the data. The next step would be to examine where the lack of fit occurred.

- If the lack of fit was in one of the tails of the distribution, perhaps a gamma or Weibull distribution would more adequately fit the data.

- Literally hundreds of probability distributions have been created, many with some specific physical process in mind. One aid to selecting distributions is to use the physical basis of the distributions as a guide. Here are some examples:

<u>6.2.3 Quantile-Quantile Plots</u>

- Further, our perception of the fit depends on widths of the histogram intervals. But even if the intervals are well chosen, grouping of data into cells makes it difficult to compare a histogram to a continues probability density function

- If X is a random variable with cdf F, then the q-quintile of X is that y such that $F(y) = P(X < y) = q$, for $0 < q < 1$. When F has an invererse, we write $y = F\text{-}1(q)$.

- Now let $\{Xi, i = 1, 2,...,n\}$ be a sample of data from X. Order the observations from the smallest to the largest, and denote these as $\{yj, j =1,2 ,,,n\}$, where $y1 < y2 < \ldots < y_n$- Let j denote the ranking or order number. Therefore, j = 1 for the smallest and j = n for the largest. The q-q plot is based on the fact that y1 is an estimate of the (j — 1/2)/n quantile of X other words,

$$Yj \text{ is approximately } F^{-1}\left[\frac{J - ½}{n}\right]$$

- Now suppose that we have chosen a distribution with cdf F as a possible representation of the distribution of X. If F is a member of an appropriate family of distributions, then a plot of y_j versus $F^{-1}((j — 1/2)/n)$ will be approximately a straight line.

6.3 Parameter Estimation

- After a family of distributions has been selected, the next step is to estimate the parameters of the distribution. Estimators for many useful distributions are described in this section. In addition, many software packages—some of them integrated into simulation languages—are now available to compute these estimates.

6.3.1 Preliminary Statistics: Sample Mean and Sample Variance

- In a number of instances the sample mean, or the sample mean and sample variance, are used to estimate of the parameters of hypothesized distribution;

- If the observations in a sample of size n are X1, X2,..., Xn, the sample mean (X) is defined by

$$\overline{X} = \frac{\sum_{i=1}^{n} Xi}{n} \qquad 9.1$$

and the sample variance, s^2 is defined by

$$S^2 = \frac{\sum_{i=1}^{n} Xi^2 - n\overline{X}^2}{n - 1} \qquad 9.2$$

If the data are discrete and grouped in frequency distribution, Equation (9.1) and (.2) can be modified to provide for much greater computational efficiency, The sample mean can be computed by

$$\overline{X}^2 = \frac{\sum_{j=1}^{n} fjXj}{n} \qquad 9.3$$

And the sample variance by

$$X^2 = \frac{\sum_{j=1}^{k} fjXj^2 - n\overline{X}^2}{n - 1} \qquad 9\ 4$$

where k is the number of distinct values of X and fj is the observed frequency of the value Xj, of X.

6.3.2 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.
- These estimators are the maximum-likelihood estimators based on the raw data. (If the data are in class intervals, these estimators must be modified.)
- The triangular distribution is usually employed when no data are available, with the parameters obtained from educated guesses for the minimum, most likely, and maximum possible value's; the uniform distribution may also be used in this way if only minimum and maximum values are available.

Distribution	Parameter	Estimator
Poisson	α	$\hat{\alpha} = \overline{X}$
Exponential	λ	$\hat{\lambda} = \frac{1}{\overline{X}}$
Gamma	β, θ	$\hat{\theta} = \frac{1}{\overline{X}}$
Normal	μ, σ^2	$\hat{\mu} = \overline{X}, \hat{\sigma}^2 = S^2$
Lognormal	μ, σ^2	$\hat{\mu} = \overline{X}, \hat{\sigma}^2 = S^2$

6.4 Goodness-of-Fit Tests

- These two tests are applied in this section to hypotheses about distributional forms of input data. Goodness-of-fit tests provide help full guidance for evaluating the suitability of a potential input model.
- However, since there is no single correct distribution in a real application, you should not be a slave to the verdict of such tests.
- It is especially important to understand the effect of sample size. If very little data are available, then a goodness-of-fit test is unlikely to reject any candidate distribution; but if a lot of data are available, then a goodness-of-fit test will likely reject all candidate distribution.

6.4.1 Chi-Square Test

- One procedure for testing the hypothesis that a random sample of size n of the random variable X follows a specific distributional form is the chi-square goodness-offit test.
- This test formalizes the intuitive idea of comparing the histogram of the data to the shape of the candidate density or mass function, The test is valid for large sample sizes, for both discrete and continuous distribution assumptions, When parameters are estimated by maximum likelihood.

$$X_0^2 = \sum_{I=1}^{k} \frac{(Oi - Ei)^2}{Ei} \qquad 9.16$$

- where 0, is the observed frequency in the ith class interval and Ei, is the expected frequency in that class interval. The expected frequency for each class interval is computed as Ei=npi, where pf is the theoretical, hypothesized probability associated with the ith class interval.
- It can be shown thatX02 approximately follows the chi-square distribution with k-s-1 degrees of freedom, where s represents the number of parameters of the hypothesized distribution estimated by sample statistics. The hypotheses are :

H0: the random variable, X, conforms to the distributional assumption with the parameter(s) given by the parameter estimate(s)

H1 : the random variable X does not conform

- If the distribution being tested is discrete, each value of the random variable should be a class interval, unless it is necessary to combine adjacent class intervals to meet the minimum expected cell-frequency requirement. For the discrete case, if combining adjacent cells is not required,

$Pi = P(X_I) = P(X\ X_i)$

Otherwise, pi, is determined by summing the probabilities of appropriate adjacent cells.

- If the distribution being tested is continuous, the class intervals are given by $[a_{i-1},ai)$, , where ai-1 and ai, are the endpoints of the ith class interval. For the continuous case with assumed pdf f(x), or assumed cdf F(x), pi, can be computed By

$Pi= {}_{ai-1}^{ai} f(x)\,dx = F(a_i) - F(a_{i-1})$

6.4.2 Chi-Square Test with Equal Probabilities

- If a continuous distributional assumption is being tested, class intervals that are equal in probability rather than equal in width of interval should be used.
- Unfortunately, there is as yet no method for deter mining the; probability associated with each interval that maximize the; power of a test o f a given size.

$Ei = n\ p\ i \quad 5$

- Substituting for p i yields $n/k \quad 5$
- and solving for k yields $k \quad n/5$

6.4.3 Kolmogorov - Smirnov Goodness-of-Fit Test

- The chi-square goodness-of-fit test can accommodate the estimation of parameters from the data with a resultant decrease in the degrees of freedom (one for J each parameter estimated). The chi-square test requires that the data be placed in class intervals, and in the case of continues distributional assumption, this grouping is arbitrary.

- Also, the distribution of the chi-square test statistic is known only approximately, and the power of the test is sometimes rather low. As a result of these considerations, goodness-of-fit tests, other than the chi-square, are desired.

- The Kolmogorov-Smirnov test is particularly useful when sample sizes are small and when no parameters have been estimated from the data.

- (Kolmogoro-Smirnov Test for Exponential Distribution)

Ho : the interarrival times are exponentially distributed

H1: the interarrival times are not exponentially distributed

- The data were collected over the interval 0 to T = 100 min. It can be shown that if the underlying distribution of interarrival times { T1, T2, … } is exponential, the arrival times are uniformly distributed on the interval (0,T).

- The arrival times T1, T1+T2, T1+T2+T3,…..,T1+…..+T50 are obtained by adding interarrival times.
- On a (0,1) interval, the points will be [T1/T, (T1+T2)/T,…..,(T1+….+T50)/T].

6.5 Selecting Input Models without Data

Unfortunately. it is often necessary in practice to develop a simulation model for demonstration purposes or a preliminary study—before any i data are available.) In this case the modeler must be resourceful in choosing input models and must carefully check the sensitivity of results to the models.

Engineering data : Often a product or process has performance ratings pro vided by the manufacturer.

Expert option : Talk to people who are experienced with the procesws or similar processes. Often they can provide optimistic, pessimistic and most likely times.

Physical or conventional limitations : Most real processes have physical limit on performance. Because of company policies, there may be upper limits on how long a process may take. Do not ignore obvious limits or bound: that narrow the range of the input process.

The nature of the process It can be used to justify a particular choice even when no data are available.

6.6 Multivariate and Time-Series Input Models

The random variables presented were considered to be independent of any other variables within the context of the problem. However, variables may be related, and if the variables appear in a simulation model as inputs, the relationship should be determined and taken into consideration.

Step 1. Generate Z_1 and Z_2, independent standard normal random variables.

Step 2. Set $X_1 = \mu_1 + \sigma_1 Z_1$

Step 3. Set $X_2 = \mu_2 + \sigma_2 \left(\rho Z_1 + \sqrt{1-\rho^2} Z_2\right)$

6.7 Time series input model:

If X1,X2..Xn is a sequence of identically distributed,but dependent and convarianc stationary random variables,then there are a number of times series model that can be used to represent the process. The two models that have the characteristics that the autocorrelatrion take the form.

$$\rho_h = \text{corr}(X_t, X_{t+h}) = \rho^h$$

for h=1,2,..n that the log-h autocorrelation decreases geometrically as the lag increases.

AR(1) Model:

consider the time series model

$$X_t = \mu + \phi(X_{t-1} - \mu) + \varepsilon_t$$

for t=2,3,..n where ε2, ε3 are the independent and identically distributed with men 0 and variance σ^2_ε and -1< ϕ<1. If the initial value x1 is chosen appropriately,then x1,x2..are all normal distributed with mean u and variance $\sigma_\varepsilon^2/(1-\phi^2)$,

Step 1. Generate X_1 from the normal distribution with mean μ and variance $\sigma_\varepsilon^2/(1-\phi^2)$. Set $t = 2$.

Step 2. Generate ε_t from the normal distribution with mean 0 and variance σ_ε^2.

Step 3. Set $X_t = \mu + \phi(X_{t-1} - \mu) + \varepsilon_t$.

Step 4. Set $t = t + 1$ and go to Step 2.

EAR(1) Model:

Consider the time series model

$$X_t = \begin{cases} \phi X_{t-1}, & \text{with probability } \phi \\ \phi X_{t-1} + \varepsilon_t, & \text{with probability } 1-\phi \end{cases}$$

for t=2,3,..n where ε2, ε3 are the independent and identically distributed with mean $1/\lambda$ and 0< ϕ<1. If the initial value x1 is chosen appropriately, then x1,x2.. are all exponentially distributed with mean $1/\lambda$ and variance $\sigma_\varepsilon^2/(1-\phi^2)$,

Step 1. Generate X_1 from the exponential distribution with mean $1/\lambda$. Set $t = 2$.

Step 2. Generate U from the uniform distribution on [0, 1]. If $U \le \phi$, then set

$$X_t = \phi X_{t-1}$$

Otherwise, generate ε_t from the exponential distribution with mean $1/\lambda$ and set

$$X_t = \phi X_{t-1} + \varepsilon_t$$

Step 3. Set $t = t + 1$ and go to Step 2.

Goodness-of-fit Tests

⊙ Chi-Square Test with poisson Assumption

Step 1: Compute poisson distribution using

$$P(x) = \begin{cases} \dfrac{e^{-\alpha}\alpha^{x}}{x!} & x = 0, 1, 2, \ldots \end{cases}$$

Step 2: Compute expected frequency

$E_i = n . P(i)$ where n is sum of Sample data

Reduce interval i.e

$E_i > 5$

Step 3: Compute chi-square test i.e

$$X_0^2 = \sum_{i=0}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Step 4: Obtain chi-square test value from table A.6

$$X^2_{\alpha, K-s-1}$$

Step 5: Check hypothesis or Null hypothesis

$X_0^2 \lessgtr X^2_{\alpha, K-s-1}$ Accepted hypothesis / Rejected Null hypothesis

*) Chi-Square test for Exponential distribution (Equal probability)

Step 1: Determine the probability

$$\boxed{P = 1/K}$$ where K is interval

Step 2: Determine the mean

$$\lambda = \frac{1}{\bar{X}}$$

$$\bar{X} = \sum_{i=0}^{n} \frac{X_i}{n}$$

Step 3: Compute class interval

$$\boxed{a_i = -\frac{1}{\lambda} \ln(1 - iP)}$$ $i = 0, 1, 2, \ldots K$

Step 4: Compute expected frequency

$$\boxed{E_i = \frac{N}{K}}$$

N - Sum of Sample data

K - interval

Step 5: Compute chi-Square test

$$\boxed{X_0^2 = \sum_{i=0}^{n} \frac{(O_i - E_i)^2}{E_i}}$$

Step 6: Obtain chi-square test value from table A.6

$$\boxed{X^2_{\alpha, K-s-1}}$$

Step 7: Check hypothesis or Null hypothesis

$$\boxed{X_0^2 < X^2_{\alpha, K-s-1}}$$ accepted

*) Kolmogorov - Smirnov test for exponential distributions

Step 1: Calculate inter arrival points

$$\boxed{R_i = \{ T_1/T,\ (T_1+T_2)/T,\ (T_1+T_2+T_3)/T,\ (T_1+T_2+T_n)/T \}}$$

T is total No's of sample data

T_i - is the sample data

Step 2: Compute

$$D^+ = \max_{1 \le i \le n} \{ i/n - R(i) \}$$

$$D^- = \max_{1 \le i \le n} \{ R(i) - \frac{i-1}{n} \}$$

Step 3: Compute

$$\boxed{D = \max(D^+, D^-)}$$

Step 4: Obtain KS-test value from table A.8

$$\boxed{D_{\alpha, n}}$$

Step 5: Check hypothesis or Null hypothesis

$\boxed{D \le D_{\alpha, n}}$ accepted

(1) UNIT-6 : INPUT MODELLING

I. Chi Square Test using Poisson Assumption

1. Using goodness of fit test, test whether random Nos are uniformly distributed based on poisson assumption with level of significance $\alpha = 0.05$. $\hat{\alpha} = 3.64$. Sample data are :

Interval :	0	1	2	3	4	5	6	7	8	9	10	11
observed Frequency :	12	10	19	17	10	8	7	5	5	3	3	1

⇒ **Given** : $\alpha = 0.05$

$\hat{\alpha} = 3.64$

$n = 12 + 10 + 19 + 17 + 10 + 8 + 7 + 5 + 5 + 3 + 3 + 1 = 100$

Step 1 : Compute Poisson Distribution

$$P(x) = \frac{e^{-\alpha} \alpha^{x}}{x!} \quad \text{where } x = 0, 1, \ldots, 11$$

$$P(0) = \frac{e^{-3.64} * (3.64)^0}{0!} = 0.026$$

$P(1) = 0.096$

$P(2) = 0.174$

$P(3) = 0.211$

$P(4) = 0.192$

$P(5) = 0.140$

$P(6) = 0.085$

$P(7) = 0.044$

$P(8) = 0.020$

$P(9) = 0.008$

$P(10) = 0.003$

$P(11) = 0.001$

Step 2 : Apply Chi Square with poisson assumption.

X_i	O_i	$E_i = n \cdot P_i$	$O_i - E_i$	$(O_i - E_i)^2$	$X_0^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$
0	12	2.6			
1	10 (22)	9.6 (12.2)	9.8	96.04	7.87
2	19	17.4	1.6	2.56	0.15
3	17	21.1	-4.1	16.81	0.79
4	10	19.2	-9.2	84.64	4.41
5	8	14.0	-6	36	2.57
6	7	8.5	-1.5	2.25	0.27
7	5	4.4	9.4	88.36	11.63
8	5 (17)	2.0			
9	3	0.8 (7.6)			
10	3	0.3			
11	1	0.1			

We have $k = 7$, $s = 1$ $\quad$ $X_0^2 = 27.69$

Step 3 : Compute level of Significance from Table A6

$$X^2_{0,\alpha, k-s-1} = X^2_{0\ 0.05,\ 7-1-1}$$

$$= X^2_{0\ 0.05,\ 5} = 11.1$$

Step 4 : Check whether Random No.s are uniformly distributed.

Compare X_0^2 & $X^2_{0\ 0.05,5}$

$\therefore$ $27.69 > 11.1$ $\Rightarrow$ Random No.s are not uniformly distributed.

2. Using goodness of fit test, check whether Random No.s are uniformly distributed over interval [0, 1] using poisson assumption with level of significance = 0.05. Simulation table for critical values is given:

(2)

Interval (X_i) :	0	1	2	3	4	5	6	7
Frequency (f_i) :	5	10	5	8	12	10	8	12

⇒ Given: $\alpha = 0.05$

$n = 5+10+5+8+12+10+8+12 = 70$

$\hat{\alpha} = ?$

$$\hat{\alpha} = \bar{X} = \frac{\sum_{i=1}^{n} f_i X_i}{n} = \frac{0+10+10+24+48+50+48+84}{70}$$

$$\hat{\alpha} = \frac{274}{70} = 3.91$$

Step 1: Compute Poisson Distribution

$$P(x) = \frac{e^{-\alpha} \alpha^{x}}{x!}, \quad x = 0, 1, \ldots, 7. \quad \& \quad \alpha = 3.91$$

$P(0) = 0.020$

$P(1) = 0.078$

$P(2) = 0.153$

$P(3) = 0.199$

$P(4) = 0.195$

$P(5) = 0.153$

$P(6) = 0.099$

$P(7) = 0.056$

Step 2 : Apply Chi Square test with poisson assumption

X_i	O_i	$E_i = n \cdot P_i$	$O_i - E_i$	$(O_i - E_i)^2$	$X_0^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$
0	5 } 15	1.4 } 6.86	8.14	66.26	9.66
1	10	5.46			
2	5	10.71	−5.71	32.60	3.04
3	8	13.93	−5.93	35.16	2.52
4	12	13.65	−1.65	2.72	0.19
5	10	10.71	−0.71	0.50	0.05
6	8 } 20	6.93 } 10.85	9.15	83.72	7.72
7	12	3.92			

$X_0^2 = 23.18$

Here $k = 6$, $s = 1$

Step 3 : Compute level of Significance from Table A6

$X^2_{\alpha, k-s-1} = X^2_{0.05, 6-1-1} = 9.49$

Step 4 : Check whether Random No.s are uniformly distributed.

Compare X_0^2 & $X^2_{0.05, 4}$

$\because 23.18 > 9.49 \Rightarrow$ Random No.s are not uniformly distributed

3. Apply goodness of fittest, check whether Random Nos are uniformly distributed over Interval [0, 1] with given size of data 100. Assume $\alpha = 0.01$. Simulation table to check critical values using poisson assumption is given below :

Interval :	1	2	3	4	5	6	7	8	9	10
Frequency :	8	6	10	11	12	8	10	12	12	11

⇒ Given : $\alpha = 0.01$ $\hat{\alpha} = ?$

$n = 100$

$$\hat{\alpha} = \bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{n} = \frac{586}{100} = 5.86$$

Step 1 : Compute Poisson Distribution

$$P(x) = \frac{e^{-\alpha} \cdot \alpha^{x}}{x!} \quad \text{where } x = 1, 2 \ldots 10 \quad \& \quad \alpha = 5.86$$

$P(1) = 0.017$

$P(2) = 0.049$

$P(3) = 0.096$

$P(4) = 0.140$

$P(5) = 0.164$

$P(6) = 0.160$

$P(7) = 0.134$

$P(8) = 0.098$

$P(9) = 0.064$

$P(10) = 0.038$

Step 2 : Apply Chi Square with poisson assumption

X_i	O_i	$E_i = n \cdot P_i$	$O_i - E_i$	$(O_i - E_i)^2$	$X_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$
1	8 } 14	1.7 } 6.6	7.4	54.76	8.29
2	6	4.9			
3	10	9.6	0.4	0.16	0.02
4	11	14.0	-3	9	0.64
5	12	16.4	-4.4	19.36	1.18
6	8	16.0	-8	64	4
7	10	13.4	-3.4	11.56	0.86
8	12	9.8	2.2	4.84	0.49
9	12 } 23	6.4 } 10.2	12.8	163.84	16.06
10	11	3.8			

$k = 8, \; s = 1$

$X_0^2 = 31.54$

Step 3 : Compute level of Significance from Table A6

$X_{0\,\alpha, \, k-s-1}^2 = X_{0\,0.01, \, 8-1-1}^2 = 20.1$

Step 4 : Check whether Random No.s are uniformly distributed

Compare X_0^2 & $X_{0\,0.05, 6}^2$

$31.54 > 20.1 \Rightarrow$ Random No.s are not uniformly distributed

(4)

II. Chi Square test with Equal Probability (Exponential Dist.)

1. Apply goodness of fit test to check whether random No.s are uniformly distributed over [0, 1] using equal probability. Use $\alpha = 0.05$, interval $k = 8$ to check whether given sample datas are accepted or rejected.

79.918	3.081	0.062	1.961	5.845	3.027	6.505
0.021	0.013	0.123	6.769	59.899	1.192	34.760
5.009	18.387	0.141	43.565	24.420	0.433	144.695
2.663	17.967	0.091	9.003	0.941	0.878	3.371
2.157	7.579	0.624	5.380	3.148	7.078	23.960
0.590	1.928	0.300	0.002	0.543	7.004	31.764
1.005	1.147	0.219	3.217	14.382	1.008	2.336
4.562						

⇒ Given : $k = 8$, $\alpha = 0.05$

Step 1 : Compute mean

$$\bar{\lambda} = \frac{1}{\bar{X}} \quad \text{where} \quad \bar{X} = \frac{\Sigma X_i}{n}$$

$$\bar{X} = \frac{594.674}{50} = 11.894$$

$$\bar{\lambda} = \frac{1}{11.894}$$

$$\boxed{\bar{\lambda} = 0.084}$$

Step 2 : Compute class intervals

$$P = \frac{1}{k} = \frac{1}{8} = 0.125$$

$$a_i = -\frac{1}{\lambda} \ln [1 - i * p] \quad \text{where } i = 0 \ldots 8$$

$\lambda = 0.084$

$P = 0.125$

$a_0 = 0$

$a_1 = 1.589$ $\qquad$ $a_5 = 11.677$

$a_2 = 3.425$ $\qquad$ $a_6 = 16.504$

$a_3 = 5.595$ $\qquad$ $a_7 = 24.755$

$a_4 = 8.252$ $\qquad$ $a_8 = \infty$

Step 3 : Compute Chi square with equal probability

Class Interval	O_i	$E_i = \frac{N}{k}$ (50/8)	$O_i - E_i$	$(O_i - E_i)^2$	$X_0^2 = \sum_{i=1} \frac{(O_i - E_i)^2}{E_i}$
0 - 1.589	19	6.25	12.75	162.563	26.01
1.589 - 3.425	10	6.25	3.75	14.063	2.25
3.425 - 5.595	3	6.25	-3.25	10.563	1.69
5.595 - 8.252	6	6.25	-0.25	0.0625	0.01
8.252 - 11.677	1	6.25	-5.25	27.563	4.41
11.677 - 16.504	1	6.25	-5.25	27.563	4.41
16.504 - 24.755	4	6.25	-2.25	5.0623	0.81
24.755 - ∞	6	6.25	-0.25	0.063	0.01
	50				$X_0^2 = 39.60$

Step 3 : Compute level of Significance from table A6.

$X_{0\,\alpha,\, k-s-1}^2 = X_{0\; 0.05,\; 8-1-1}^2 = 12.6$

Step 4 : Check whether random No.s are uniformly distributed.

$39.60 > 12.6 \Rightarrow$ Random No.s are rejected.

3. Consider goodness of fit test using Chi Square test with equal probability. Given $k = 6$, $\alpha = 0.05$. Sample data:

(5)

0.34	0.90	1.88	1.90	0.74	2.62	2.67	3.53	4.91
5.50	1.10	1.03	1.73	1.00	2.03	1.49	2.16	0.80
0.48	5.60	0.45	0.26	0.24	0.63	0.36	1.28	0.82
2.16	0.05	0.04	0.39	0.21	0.79	0.53	3.53	2.62
0.53	1.50	2.81						

Given: $k = 6$ $\alpha = 0.05$ $N = 39$

Step 1: Compute Mean

$\bar{\lambda} = \frac{1}{\bar{X}}$ where $\bar{X} = \frac{\Sigma X_i}{N} = \frac{61.61}{39} = 1.579$

$\bar{\lambda} = \frac{1}{1.579} = \underline{\underline{0.63}}$

Step 2: Compute class intervals

$p = \frac{1}{k} = \frac{1}{6} = \underline{\underline{0.17}}$

$a_i = \frac{-1}{\lambda} \ln [1 - i * p]$ where $i = 0, 1 \dots 6$

$\lambda = 0.63$

$p = 0.17$

$a_0 = 0$

$a_1 = 0.29$

$a_2 = 0.66$

$a_3 = 1.13$

$a_4 = 1.81$

$a_5 = 3.01$

$a_6 = \infty$

Class Interval	O_i	$E_i = \frac{N}{K}$	$O_i - E_i$	$(O_i - E_i)^2$	$X_0^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$
0 - 0.29	5	6.5	-1.5	2.25	0.35
0.29 - 0.66	8	6.5	1.5	2.25	0.35
0.66 - 1.13	8	6.5	1.5	2.25	0.35
1.13 - 1.81	4	6.5	-2.5	6.25	0.96
1.81 - 3.01	9	6.5	2.5	6.25	0.96
3.01 - ∞	5	6.5	-1.5	2.25	0.35

$X_0^2 = 3.32$

Step 3 : Compute level of significance from table A6

$X_0^2{}_{\alpha,\ k-s-1} = X_0^2{}_{0.05,\ 6-1-1} = 9.49$

Step 4 : Check whether Random Nos are uniformly distributed.

$\because 3.32 < 9.49 \Rightarrow$ Random Nos are accepted

III K-S Test

1. Apply goodness of fit test to check whether Random Nos are uniformly distributed over $(0, T)$ for an interval 100. Take $\alpha = 0.05$

Simulation table for Critical values :

0.44 0.53 2.04 2.74 2.00 0.30 2.54 0.52 2.02 1.89

Given : $\alpha = 0.05$, $n = 10$, $T = 100$

$$R_{(i)} = \left\{ \frac{T_1}{T}, \frac{T_1 + T_2}{T}, \ldots\ldots, \frac{T_1 + T_2 + \ldots T_n}{T} \right\}$$

⑥ Step 1 :

$R_{(i)} = \{ 0.0044, 0.0097, 0.0301, 0.0575, 0.0775, 0.0805, 0.1059, 0.1111, 0.1313, 0.1502 \}$

Step 2 :

i	$R_{(i)}$	i/n	$i-1/n$	$D^+ = \max\{\frac{i}{n} - R_{(i)}\}$	$D^- = \max\{R_{(i)} - \frac{i-1}{n}\}$
1	0.0044	0.1	0	0.0956	0.0044
2	0.0097	0.2	0.1	0.1903	~
3	0.0301	0.3	0.2	0.2699	~
4	0.0575	0.4	0.3	0.3425	~
5	0.0775	0.5	0.4	0.4225	~
6	0.0805	0.6	0.5	0.5195	~
7	0.1059	0.7	0.6	0.5941	~
8	0.1111	0.8	0.7	0.6889	~
9	0.1313	0.9	0.8	0.7687	~
10	0.1502	1.0	0.9	0.8498	~

Step 3 :

$D = \max\{D^+, D^-\} = \max\{0.8498, 0.0044\}$

$D = \underline{\underline{0.8498}}$

Step 4 :

$D_{\alpha, n}$ from A8 table.

$D_{0.05, 10} = \underline{\underline{0.410}}$

Step 5 :

$\therefore 0.8498 > 0.410 \Rightarrow$ Random Nos are rejected

2. Consider Sample data. Perform KS Test.

0.10 1.42 0.46 0.07 1.09 0.76 5.53 3.93 1.07

2.26 2.88 0.67 1.12 0.26

Interval (0-T) = 100 min $\alpha = 0.05$ $n = 14$

= Step 1 :

$R_{(i)} = \{$ 0.0010, 0.0152, 0.0198, 0.0205, 0.0314, 0.039, 0.0943, 0.1336, 0.1443, 0.1669, 0.1957, 0.2024, 0.2136, 0.2162 $\}$

Step 2 :

i	$R_{(i)}$	i/n	i-1/n	$D^+ = \max\{\frac{i}{n} - R_{(i)}\}$	$D^- = \max\{R_{(i)} - \frac{i-1}{n}\}$
1	0.0010	0.0714	0	0.0704	0.0010
2	0.0152	0.1429	0.0714	0.1277	~
3	0.0198	0.2143	0.1429	0.1945	~
4	0.0205	0.2857	0.2143	0.2652	~
5	0.0314	0.3571	0.2857	0.3257	~
6	0.039	0.4286	0.3571	0.3896	~
7	0.0943	0.5	0.4286	0.4057	~
8	0.1336	0.5714	0.5	0.4378	~
9	0.1443	0.6429	0.5714	0.4986	~
10	0.1669	0.7143	0.6429	0.5474	~
11	0.1957	0.7857	0.7143	0.59	~
12	0.2024	0.8571	0.7857	0.6547	~
13	0.2136	0.9286	0.8571	0.715	~
14	0.2162	1	0.9286	0.7838	~

Step 3 : $D = \max\{0.7838, 0.0010\} = \underline{0.7838}$

Step 4 : $D_{0.05, 14} = 0.349$ (A8 table)

Step 5 : $\because$ 0.7838 > 0.349 $\Rightarrow$ <u>Random No.s</u> are rejected

OUTPUT ANALYSIS FOR A SINGLE MODEL

Estimate system performance via simulation

- If q is the system performance, the precision of the estimator can be measured by:
  1. The standard error of .
  2. The width of a confidence interval (CI) for q.
- Purpose of statistical analysis:
  1. To estimate the standard error or CI .
  2. To figure out the number of observations required to achieve desired error/CI.
- Potential issues to overcome:
  1. Autocorrelation, e.g. inventory cost for subsequent weeks lack statistical independence.
  2. Initial conditions, e.g. inventory on hand and # of backorders at time 0 would most likely influence the performance of week 1.

7.1 Type of Simulations

- Terminating verses non-terminating simulations
- Terminating simulation:
  1. Runs for some duration of time T_E, where E is a specified event that stops the simulation.
  2. Starts at time *0* under well-specified initial conditions.
  3. Ends at the stopping time T_E.
  4. Bank example: Opens at 8:30 am (time *0*) with no customers present and *8* of the *11* teller working (initial conditions), and closes at 4:30 pm (Time $T_E = 480$ minutes).
  5. The simulation analyst chooses to consider it a terminating system because the object of interest is one day's operation.

7.2 Stochastic Nature of Output Data

- Model output consist of one or more random variables (r. v.) because the model is an input-output transformation and the input variables are r.v.'s.

- M/G/1 queuing example:

  1. Poisson arrival rate = *0.1* per minute;
     service time ~ $N(m = 9.5, s = 1.75)$.

  2. System performance: long-run mean queue length, $L_Q(t)$.

  3. Suppose we run a single simulation for a total of 5,000 minutes

- Divide the time interval [*0, 5000*) into *5* equal subintervals of *1000* minutes.

Average number of customers in queue from time *(j-1)1000* to *j(1000)* is Y_j .

- M/G/1 queueing example (cont.):

- Batched average queue length for 3 independent replications:

Batching Interval (minutes)	Batch, j	Replication 1, Y_{1j}	Replication 2, Y_{2j}	Replication 3, Y_{3j}
[0, 1000)	1	3.61	2.91	7.67
[1000, 2000)	2	3.21	9.00	19.53
[2000, 3000)	3	2.18	16.15	20.36
[3000, 4000)	4	6.92	24.53	8.11
[4000, 5000)	5	2.82	25.19	12.62
[0, 5000)		3.75	15.56	13.66

- Inherent variability in stochastic simulation both within a single replication and across different replications.
- The average across 3 replications, can be regarded as independent observations, but averages within a replication, $Y_{11}, \ldots, Y_{15}$, are not.

7.3 Measures of performance

- Consider the estimation of a performance parameter, *q* (or *f*), of a simulated system.

1. Discrete time data: $[Y_1, Y_2, \ldots, Y_n]$, with ordinary mean: *q*
2. Continuous-time data: $\{Y(t), 0 \le t \le T_E\}$ with time-weighted mean: *f*

7.3.1 Point Estimator

- Point estimation for discrete time data$[Y_1, Y_2, \ldots, Y_n]$ is defined by.

  The point estimator:

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

- Where $\hat{\theta}$ is a sample mean based on sample of size n The pointer estimator $\hat{\theta}$ is said to be unbiased for θ if its expected value is θ, that is if: Is biased

$$E(\hat{\theta}) = \theta$$

- Point estimation for continuous-time data. The point estimator:

$$\hat{\phi} = \frac{1}{T_E}\int_0^{T_E} Y(t)\,dt$$

- An unbiased or low-bias estimator is desired.

- Usually, system performance measures can be put into the common framework of *q* or *f:* the proportion of days on which sales are lost through an out-of-stock situation, let:

$$Y(t) = \begin{cases} 1, & \text{if out of stock on day } i \\ 0, & \text{otherwise} \end{cases}$$

- Performance measure that does not fit: quantile or percentile:
- Estimating quantiles: the inverse of the problem of estimating a proportion or probability. $\Pr\{Y \le \theta\} = p$
- Consider a histogram of the observed values *Y*:
- Find such that *100p*% of the histogram is to the left of (smaller than) .

7.3.2 Confidence-Interval Estimation

To understand confidence intervals fully, it is important to distinguish between measures of error, and measures of risk, e.g., confidence interval versus prediction interval.

Suppose the model is the normal distribution with mean *q*, variance s^2 (both unknown).

- Let Y_i be the average cycle time for parts produced on the i^{th} replication of the simulation (its mathematical expectation is *q*).
- Average cycle time will vary from day to day, but over the long-run the average of the averages will be close to *q*.
- Sample variance across *R* replications: $$S^2 = \frac{1}{R-1}\sum_{i=1}^{R}(Y_{i.} - Y_{..})^2$$

7.3.3 Confidence-Interval Estimation

- Confidence Interval (CI):
  - A measure of error.
  - Where $Y_{i.}$ are normally distributed.

$$Y_{..} \pm t_{\alpha/2,R-1}\frac{S}{\sqrt{R}}$$

- We cannot know for certain how far is from q but CI attempts to bound that error.
- A CI, such as 95%, tells us how much we can trust the interval to actually bound the error between and q .
- The more replications we make, the less error there is in (converging to *0* as *R* goes to infinity).

7.3.4 Confidence-Interval Estimation

- Prediction Interval (PI):
  - A measure of risk.
  - A good guess for the average cycle time on a particular day is our estimator but it is unlikely to be exactly right.
  - PI is designed to be wide enough to contain the *actual* average cycle time on any particular day with high probability.
  - Normal-theory prediction interval:

$$\bar{Y}_{..} \pm t_{\alpha/2,R-1} S \sqrt{1+\frac{1}{R}}$$

  - The length of PI will not go to *0* as *R* increases because we can never simulate away risk.
  - PI's limit is: $\theta \pm z_{\alpha/2}\sigma$