**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:**

1. 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.
2. 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
3. Choose parameters that determine a specific instance of the distribution family. When data are available, these parameters may be estimated from the data.
4. 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 chi-square and the Kolmogorov-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

**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.
* "GIGO," or "garbage-in, garbage-out," is a basic concept in computer science and it applies equally in the area of discrete system simulation.
* Many lessons can be learned from an actual experience in data collection.

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 (extra/surplus) 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.

**PURPOSE OF DATA COLLECTION**

By now you have known that data could be classified in the following three ways:

a) Quantitative and Qualitative Data.

b) Sample and Census Data.

c) Primary and Secondary data.

**A)Quantitative and Qualitative data:** Quantitative data are those set of information which are quantifiable and can be expressed in some standard units like rupees or birr, kilograms, liters, etc. For example, pocket money of students of a class and income of their parents can be expressed in so many birrs; production or import of wheat can be expressed in so many kilograms or quintals; consumption of petrol and diesel in a country as so many liters in one year and so on.

Qualitative data, on the other hand, are not quantifiable, that is, cannot be expressed in standard units of measurement like kilograms, liters, etc. This is because they are 'features', 'qualities' or 'characteristics' like eye colors, skin complexion, honesty, good or bad, etc. These are also referred to as attributes. In this case, however, it is possible to count the number of individuals (or items) possessing a particular attribute.

**B) Sample and Census Data:** It can be collected either by census method or sample method. Information collected through sample inquiry is called sample data and the one collected through census inquiry is called census data. Population census data are collected every ten years in a country.

**C) Primary and Secondary Data:** The Primary data are collected by the investigator through field survey. Such data are in raw form and must be refined before use. On the other hand, secondary data are extracted from the existing published or unpublished sources, which are from the data already collected by others.

Collection of data is the first basic step towards the statistical analysis of any problem. The collected data are suitably transformed and analyzed to draw conclusions about the population.

**The following methods to collect primary data**

a) Direct Personal Investigation

b) Indirect Oral Investigation

c) Use of Local Reports

d) Questionnaire Method

**Identifying the Distribution with Data:**

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

**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 width may be used if the heights of the frequencies are adjusted).

2. Label the horizontal axis to confirm to the intervals selected.

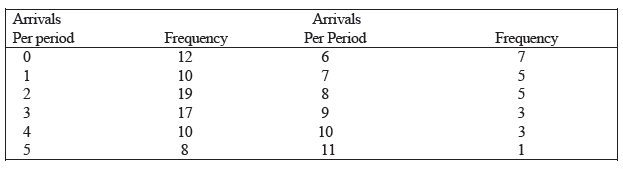
3. Determine the frequency of occurrences within each interval.

4. Label the vertical axis so that the total occurrences can be plotted for each interval.

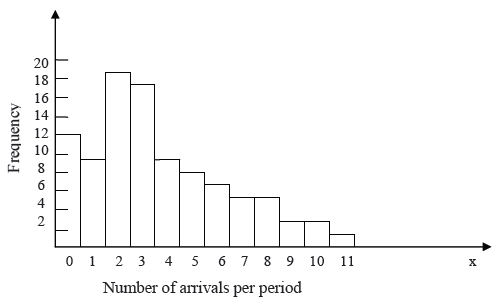
5. Plot the frequencies on the vertical axis.

* If the intervals are too wide, the histogram will be coarse or blocky (three-dimensional, boxy in shape), 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 4.1:** 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,

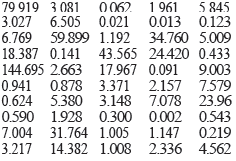


**Table 4:1 Number of Arrivals in a 5 Minute period**



**Fig 4.1** Histogram of number of arrivals per period

**Example 4.2:** Life tests were performed on a random sample of electronic chips at 1.5 times the nominal voltage, and their lifetime (or time to failure) in days was recorded:



Lifetime, usually considered a continuous variable, is recorded here to three decimal place accuracy. The histogram is prepared by placing the data in class intervals. The range of the data is rather large, from 0.002 day to 144.695 days. However, most of the value (30 0f 50) are in the zero to 5 days range. Using intervals of width three results in table 4.2.The data table 4.1 is then used to prepare the histogram shown in figure 4.2

|  |
| --- |
| Chips life(day) frequency |
| 0≤xj<3 23  3≤xj<6 10  6≤xj<9 5  9≤xj<12 1  12≤xj<15 1  15≤xj<18 2  18≤xj<21 0  21≤xj<24 1  24≤xj<27 1  27≤xj<30 0  30≤xj<33 1  33≤xj<36 1  .  .  42≤xj<45 1  .  .  57≤xj<60 1  .  .  78≤xj<81 1  .  .  144≤xj<147 1 |

Table 4.2 electronic chip data

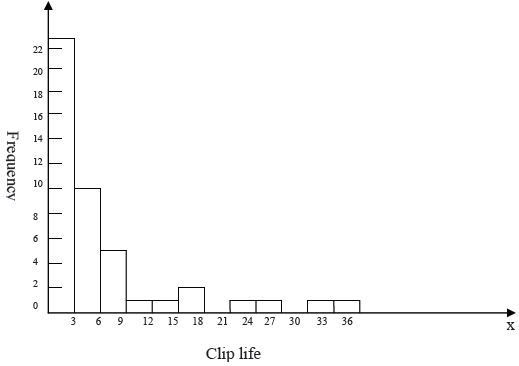
****

Fig 4.2 Histogram of chip life

**Selecting the Family of Distributions**

* Additionally, the shapes of these distributions were displayed. The purpose of preparing histogram is to infer (gather/conclude/understand) a known pdf *(probability density function)* or pmf *(probability mass function)*. 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 inter-arrival time data have been collected and the histogram has a shape similar to the pdf (*probability density function*).
* Similarly, if measurements of weights of pallets (platform) of freight (goods/shipment) are being made, and the histogram appears symmetric about the mean with a shape, 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:***

1. **Binomial:** Models the number of successes in ‘n’ trials, when the trials are independent with common success probability p, for example, the number of defective computer chips found in a lot of ‘n’ chips.
2. **Negative Binomial (includes the geometric distribution):** Models the number of trials required to achieve k successes; for example, the number of computer chips that we must inspect to find 4 defective chips.
3. **Poisson:** Models the number of independent events that occur in a fixed amount of time or space: for example, the number of customers that arrive to a store during 1 hour, or the number of defects found in 30 square meters of sheet metal.
4. **Normal:** Models the distribution of a process that can be thought of as the sum of a number of component processes; for example, the time to assemble a product which is the sum of the times required for each assembly operation. Notice that the normal distribution admits negative values, which may be-impossible for process times.
5. **Lognormal:** Models the distribution of a process that can be thought of as the product of (meaning to multiply together) a number of component processes; for example, the rate of return on an investment, when interest is compounded, is the product of the returns for a number of periods.
6. **Exponential:** Models the time between independent events, or a process time which is Memory less (knowing how much time has passed gives no information about how much additional time will pass before the process is complete); for example, the times between the arrivals of a large number of customers who act independently of each other. The exponential is a highly variable distribution and is sometimes overused because it often leads to mathematically tractable models. Recall that, if the time between events is exponentially distributed, then the number of events in a fixed period of time is Poisson.
7. **Gamma:** An extremely flexible distribution used to model nonnegative random variables. The gamma can be shifted away from 0 by adding a constant.
8. **Beta:** An extremely flexible distribution used to model bounded (fixed upper and lower limits) random variables. The beta can be shifted away from 0 by adding a constant and can have a larger range than [0, 1] by multiplying by a constant.
9. **Erlang:** Models processes that can be viewed as the sum of several exponentially distributed processes; for example, a computer network fails when a computer and two backup computers fail, and each has a time to failure that is exponentially distributed. The Erlang is a special case of the gamma.
10. **Weibull:** Models the time to failure for components; for example, the time to failure for a disk drive. The exponential is a special case of the Weibull. Discrete or Continuous Uniform Models complete uncertainty, since all outcomes are equally likely. This distribution is often overused when there are no data.
11. **Discrete or Continuous Uniform** Models complete uncertainty, since all outcomes are equally likely. This distribution is often overused when there are no data.
12. **Triangular** Models a process when only the minimum, most-likely, and maximum values of the distribution are known; for example, the minimum, most likely, and maximum time required testing a product.
13. **Empirical** Re-samples from the actual data collected; often used when no theoretical distribution seems appropriate:

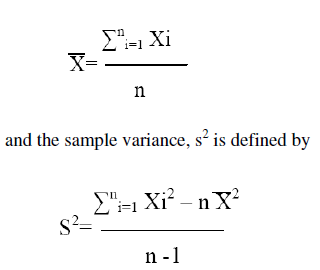
Do not ignore physical characteristics of the process when selecting distributions. Is the process naturally discrete or continuous valued? Is it bounded or is there no natural bound. This knowledge, which does not depend on data, can, help narrow the family of distributions from which to choose.

**4.3 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.

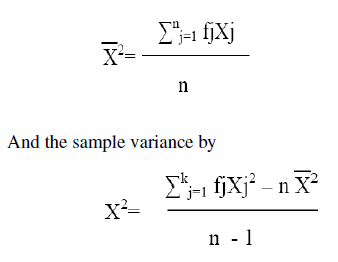
**4.3.1 Preliminary Statistics: Sample Mean and Sample Variance**

In a number of instances the sample mean or the sample mean and sample variance, are used to estimate of the parameters frequency distribution. Equations (4.5) and (4.6) are used when the data are discrete or continuous and-have been placed in class intervals. Equations (4.5) and (9.6) are approximations and should be of hypothesized distribution; see Example 9.4. In the following paragraphs, three sets of equations are given for computing the sample mean and sample variance, -Equations (4.1) and (4.2). Equations (4.3}.and (4.4} are used when the data are discrete and have been grouped in used only when the raw data are unavailable.

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

4.1

4.2

If the data are discrete and grouped in frequency distribution, Equation (4.1) and (.2) can be modified to provide for much greater computational efficiency, the sample mean can be computed by

4.3

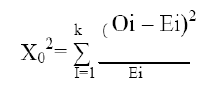
4.4

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

**4.3.2 Suggested Estimators**

1. Numerical estimates of the distribution parameters are needed to reduce the family of distributions to a specific distribution and to test the resulting hypothesis.
2. These estimators are the maximum-likelihood estimators based on the raw data. (If the data are in class intervals, these estimators must be modified.)
3. 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 values; the uniform distribution may also be used in this way if only minimum and maximum values are available.
4. Examples of the use of the estimators are given in the following paragraphs. The reader should keep in mind that a parameter is an unknown constant, but the estimator is a statistic or random variable because it depends on the sample values. To distinguish the two clearly, if, say, a parameter is denoted by a, the estimator will be denoted by.
   1. **Goodness-of-Fit Tests**
5. 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.
6. However, since there is no single correct distribution in a real application, you should not be a slave to the verdict of such tests.
7. 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
   * 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-of fit- 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.



4

4.5

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 (XI) = P(X =Xi)**

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 [ai-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



* + 1. **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.

Suppose that 50 inter arrival times (in minutes) are collected over the following 100 minute interval (arranged in order of occurrence)

0.44 0.53 2.04 2.74 2.00 0.30 2.54 0.52 2.02 1.89 1.53 0.21

2.80 0.04 1.35 8.32 2.34 1.95 0.10 1.42 0.46 0.07 1.09 0.76

5.55 3.93 1.07 2.26 2.88 0.67 1.12 0.26 4.57 5.37 0.12 3.19

1.63 1.46 1.08 2.06 0.85 0.83 2.44 2.11 3.15 2.90 6.58 0.64

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].

0.0044 0.0097 0.0301 0.0575 0.0775 0.0805 0.1059 0.1111 0.1313 0.1502

0.1655 0.1676 0.1956 0.1960 0.2095 0.2927 0.3161 0.3356 0.3366 0.3508

0.3553 0.3561 0.3670 0.3746 0.4300 0.4694 0.4796 0.5027 0.5315 0.5382

0.5494 0.5520 0.5977 0.6514 0.6526 0.6845 0.7008 0.7154 0.7262 0.7468

0.7553 0.7636 0.7880 0.7982 0.8206 0.8417 0.8732 0.9022 0.9680 0.9744

* + 1. **p-Values and "Best Fits"**

To apply a goodness-of-fit test a significance level must be chosen. The traditional significance levels are 0.1, 0.05, and 0.01. Prior to the availability of high-speed computing, having a small set of standard values made it possible to produce tables of useful critical values. Now most statistical software computes critical values as needed, rather than storing them in tables. Thus, if the analyst prefers a level of significance of, say, 0.07, then he or she can choose it.

However, rather than require a prespecified significance level, many software packages compute a p-value for the test statistic. The p-value is the significance level at which one would just reject H0 for the given value of the test statistic. While a small p-value suggests a poor fit (to accept we would have to insist on almost no risk).

The p-value can be viewed as a measure of fit, with larger values being better. This suggests that we could fit every distribution at our disposal, compute a test statistic for each fit, and then choose the distribution that yields the largest p-value.

While we know of no input modeling software that implements this specific algorithm, many such packages do include a "best-fit"' option in which the software recommends an input model to the user based on evaluating all feasible models.

In the end some summary measure of fit, like the p-value, is used to rank the distributions. There is nothing wrong with this, but there are several things to keep in mind:

1. The software may know nothing about the physical basis of the data and that information can suggest distribution families that are appropriate.
2. Recall that both the Erlang and the exponential distributions are special cases of the gamma, while the exponential is also a special case of the more flexible Weibull. Automated best-fit procedures tend to choose the more flexible distributions (gamma and Weibull over Erlang and exponential) because the extra flexibility allows closer conformance to the data and a better summary measure of fit. But again, close conformance to the data may not always lead to the most appropriate input model.
3. A summary statistic, like the p-value, is just that, a summary measure. It says little or nothing about where the lack of fit occurs (in the body of the distribution, in the right tail or in the left tail). A human, using graphical tool can see where the lack of fit occurs and decide whether or not it is important for the application at hand.
   1. **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 data is 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 process 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.

* 1. **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.