

## System Simulation Chapter 9: Input Modeling

Fatih Cavdur  
fatihcavdur@uludag.edu.tr

May 19, 2013

### Input Data

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

- Collect data from the real system of interest.
- Identify a probability distribution to represent the input process.
- Choose parameters that determine a specific instance of the distribution family.
- Evaluate the chosen distribution and associated parameters.

## Data Collection

- A useful expenditure of time is in planning. This could begin by practice or pre-observing session.
- Try to analyze the data as they are being collected. Figure out whether the data being collected are adequate to provide the distributions needed as input to the simulation.
- Try to combine homogeneous data sets.

## Data Collection-cont.

- Be aware of the possibility data censoring, in which a quantity of interest is not observed in its entirety.
- To discover whether there is a relationship between two variables, build a scatter diagram.
- Consider the possibility that a sequence of observations that appear to be independent actually has autocorrelation.
- Keep in mind the difference between input data and output or performance data, and be sure to collect input data.

## Identifying the Distribution with Data

- Divide the range of the data into intervals.
- Label the horizontal axis to conform to the intervals selected.
- Find the frequency of occurrences within each interval.
- Label the vertical axis so that the total occurrences can be plotted for each interval.
- Plot the frequencies on the vertical axis.

## Identifying the Distribution with Data-cont.

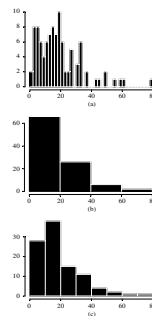


Figure: Examples of Ragged, Coarse and Appropriate Histograms

## Selecting the Family of Distributions

- **Binomial:** Models the number of successes in  $n$  trials, when the trials are independent with common success probability,  $p$ .
- **Negative Binomial (also includes the Geometric distribution):** Models the number of trials required to achieve  $k$  successes.
- **Poisson:** Models the number of independent events that occur in a fixed amount of time or space.
- **Normal:** Models the number of a process that can be thought of as the sum of a number of component processes.

## Selecting the Family of Distributions-cont.

- **Log-Normal:** Models the distribution of a process that can be thought of as the product of (meaning to multiply) a number of component processes.
- **Exponential:** Models the time between independent events or a process time that is memoryless.
- **Gamma:** An extremely flexible distribution used to model non-negative RVs.
- **Beta:** An extremely flexible distribution used to model bounded RVs.

## Selecting the Family of Distributions-cont.

- **Erlang:** Models processes that can be viewed as the sum of several exponentially distributed processes.
- **Weibull:** Models the time to failure for components.
- **Uniform (continuous or discrete):** Models complete uncertainty; all outcomes are equally likely.
- **Triangular:** Models a process for which only the minimum, most likely and maximum values of the distribution are known.
- **Empirical:** Re-samples from the actual data collected; often used when no theoretical distribution seems appropriate.

## Quantile-Quantile Plots

- The construction of histograms and the recognition of a distributional shape are necessary ingredients of for selecting a family of distributions to represent a sample of data.
- However, a histogram is not as useful for evaluating the *fit* of the chosen distribution.
- Even if the intervals are chosen well, it is difficult to compare a histogram to a continuous PDF.
- A quantile-quantile ( $q - q$ ) plot is a useful tool for evaluating distribution fit that does not suffer from these problems.

## Quantile-Quantile Plots

- If  $X$  is a RV with CDF  $F$ , then, the  $q$ -quantile of  $X$  is that value  $\gamma$  such that  $F(\gamma) = P(X \leq \gamma) = q$ , for  $0 < q < 1$ . When  $F$  has an inverse, we write  $\gamma = F^{-1}(q)$ .
- Now let  $\{X_i, i = 1, 2, \dots, n\}$  be a sample of data from  $X$ . Order the observations from the smallest to the largest, and denote these as  $\{y_j, j = 1, 2, \dots, n\}$ , where  $j$  denotes the ranking. The  $q - q$  plot is based on the fact that  $y_j$  is an estimate of the  $(j - 1/2)/n$  quantile of  $X$ .

$$y_j \sim F^{-1}\left(\frac{j - \frac{1}{2}}{n}\right)$$

## Quantile-Quantile Plots

- Now suppose that we have chosen a distribution with CDF  $F$  as a possible representation of the distribution of  $X$ .
- If  $F$  is appropriate, then, a plot of  $y_j$  vs  $F^{-1}[(j - 1/2)/n]$  will be *approximately a straight line*.
- If not, the points will deviate from a straight line, usually in a systematic manner.
- The decision about whether to reject some hypothesized model is subjective.

## Quantile-Quantile Plots

Example 9.4: A sample of 20 installation times in seconds is shown in the text (DESS) where the sample mean is 99.99 seconds and the sample standard deviation is 0.2832. These can serve as the parameter estimates for the mean and variance of the normal distribution. The observations are now ordered from smallest to largest as follows:

## Quantile-Quantile Plots

99.79	99.56	100.17	100.33
100.26	100.41	99.98	99.83
100.23	100.27	100.02	100.47
99.55	99.62	99.65	99.82
99.96	99.90	100.06	99.85

Table: Sample Installation Times

## Quantile-Quantile Plots

$j$	Value	$j$	Value	$j$	Value	$j$	Value
1	99.55	6	99.82	11	99.98	16	100.26
2	99.56	7	99.83	12	100.02	17	100.27
3	99.62	8	99.85	13	100.06	18	100.33
4	99.65	9	99.90	14	100.17	19	100.41
5	99.79	10	99.96	15	100.23	20	100.47

Table: Ordered Sample Installation Times

## Quantile-Quantile Plots

- The ordered observations are plotted versus  $F^{-1}[(j - 1/2)/20]$ , for  $j = 1, 2, \dots, 20$ , where  $F$  is the CDF of the normal distribution with mean 99.99 and standard deviation 0.2832, to obtain a  $q - q$  plot.
- The plotted values are shown in Figure 2, along with a histogram of the data that the density function of the normal distribution superimposed.
- Note that it is difficult to tell whether the data are well represented by a normal distribution from the histogram, but the general perception of a straight line is quite clear in the  $q - q$  plot and supports the hypothesis of a normality



## Quantile-Quantile Plots

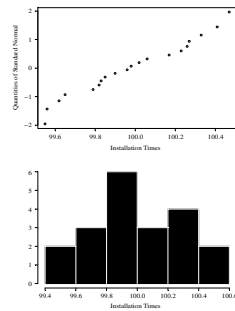


Figure: Histogram and  $q - q$  Plot for the Example

## Sample Mean and Sample Variance

- After a family of distributions has been selected, the next step is to estimate the parameters of the distribution.
- Preliminary Statistics: Sample Mean and Sample Variance
- If the observations in a sample of size  $n$  are  $X_1, X_2, \dots, X_n$ , we have

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

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

## Sample Mean and Sample Variance

If the data are discrete and have been grouped in a frequency distribution, the above equations can be modified to provide for much greater computational efficiency as

$$\bar{X} = \frac{\sum_{j=1}^k f_j X_j}{n}$$

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

where  $k$  is the number of distinct values of  $X$  and  $f_j$  is the observed frequency of the value  $X_j$  of  $X$ .

## Sample Mean and Sample Variance

When the raw data are not available, we can use the following

$$\bar{X} = \frac{\sum_{j=1}^c f_j m_j}{n}$$

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

where  $f_j$  and  $m_j$  are the frequency and the midpoint of the  $j$ th interval.

## Sample Mean and Sample Variance

Example 9.5: Grouped Data: The data in Table 9.1 in the text (DESS) can be analyzed to obtain  $n = 100$ ,  $f_1 = 12$ ,  $X_1 = 0$ ,  $f_2 = 10$ ,  $X_2 = 1$ , etc. We thus have

$$\bar{X} = \frac{\sum_{j=1}^k f_j X_j}{n} = \frac{364}{100} = 3.64$$

$$S^2 = \frac{\sum_{j=1}^k f_j X_j^2 - n\bar{X}^2}{n - 1} = \frac{2080 - 100(3.64)^2}{99} = 7.63$$

## Sample Mean and Sample Variance

Example 9.6: Continuous Data in Class Intervals Using the data in Table 9.2 in the text, we can approximate the mean and the variance as  $f_1 = 23$ ,  $m_1 = 1.5$ ,  $f_2 = 10$ ,  $m_2 = 4.5$ , etc. and

$$\bar{X} = \frac{\sum_{j=1}^c f_j m_j}{n} = \frac{614}{50} = 12.28$$

$$S^2 = \frac{\sum_{j=1}^c f_j m_j^2 - n\bar{X}^2}{n - 1} = \frac{37,226.5 - (50)(12.28)^2}{49} = 605.849$$

## Suggested Estimators

Distribution	Parameter(s)	Suggested Estimator(s)
Poisson	$\alpha$	$\hat{\alpha} = \bar{X}$
Exponential	$\lambda$	$\hat{\lambda} = 1/\bar{X}$
Gamma	$\beta, \theta$	$\hat{\beta}$
		$\hat{\theta} = 1/\bar{X}$

Table: Suggested Estimators for Distributions

## Suggested Estimators

Distribution	Parameter(s)	Suggested Estimator(s)
Normal	$\mu, \sigma^2$	$\hat{\mu} = \bar{X}$
		$\hat{\sigma}^2 = S^2$
Log-Normal	$\mu, \sigma^2$	$\hat{\mu} = \bar{X}$ (after taking the log)
		$\hat{\sigma}^2 = S^2$ (after taking the log)

Table: Suggested Estimators for Distributions

## Suggested Estimators

Distribution	Parameter(s)	Suggested Estimator(s)
Weibull ( $v = 0$ )	$\alpha, \beta$	$\hat{\beta}_0 = \bar{X}/S$ $\hat{\beta}_j = \hat{\beta}_{j-1} - \frac{f(\hat{\beta}_{j-1})}{f'(\hat{\beta}_{j-1})}$
Beta	$\beta_1, \beta_2$	$\psi(\hat{\beta}_1) + \psi(\hat{\beta}_1 - \hat{\beta}_2) = \ln(G_1)$ $\psi(\hat{\beta}_2) + \psi(\hat{\beta}_1 - \hat{\beta}_2) = \ln(G_2)$ where $G_1 = (\prod_{i=1}^n X_i)^{1/n}$ $G_2 = [\prod_{i=1}^n (1 - X_i)]^{1/n}$

Table: Suggested Estimators for Distributions

## Suggested Estimators

Example 9.7: Assume that the arrival data in Table 9.1 require analysis. By comparison with Poisson PFM and CDF, an examination of the histogram of the data suggests a Poisson distributional assumption with unknown parameter  $\alpha$ . The estimator for  $\alpha$  is  $\bar{X}$ , which was found in Example 9.5. Thus,  $\hat{\alpha} = 3.64$ . Recall that the true mean and variance are equal for the Poisson distribution.

## Suggested Estimators

Example 9.10: The estimator  $\hat{\beta}$  for the gamma distribution requires the computation of the quantity  $1/M$ , where

$$M = \ln \bar{X} - \frac{1}{n} \sum_{i=1}^n \ln X_i$$

Also,  $\hat{\theta}$  is given by

$$\hat{\theta} = \frac{1}{\bar{X}}$$

## Suggested Estimators

Using the data given in the Example, to estimate  $\hat{\beta}$  and  $\hat{\theta}$ , it is necessary to compute  $M$  using the above equation.

$$\bar{X} = \frac{564.32}{20} = 28.22 \Rightarrow \ln \bar{X} = 3.34$$

$$\sum_{i=1}^{20} \ln X_i = 63.99 \Rightarrow M = 3.34 - \frac{63.99}{20} = 0.14 \Rightarrow \frac{1}{M} = 7.14$$

By interpolation,  $\hat{\beta} = 3.728$ . Finally,

$$\hat{\theta} = \frac{1}{28.22} = 0.035$$

## Chi-Square Test

Goodness-of-fit tests provide helpful guidance for evaluating the suitability of a potential input model; however, there is no single correct distribution in a real application, so you should not be a slave to the verdict of such a test. 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 distributions; but if a lot of data are available, then, a goodness-of-fit test will likely to reject all candidate distributions.

## Chi-Square Test

This test formalizes the intuitive idea of comparing the histogram to the shape of the PDF or PMF. This test is valid for large sample sizes and for both discrete and continuous distributions when parameters are estimated by maximum likelihood. The test procedure begins by arranging the  $n$  observations into a set of  $k$  class intervals or cells. The test statistic is given by

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

where  $O_i$  and  $E_i$  are the observed and expected frequencies in the  $i$ th class interval.

## Chi-Square Test

The expected frequency is computed as  $E_i = np_i$ , where  $p_i$  is the theoretical probability associated with the  $i$ th class interval. We can show that  $\chi_0^2$  approximately follows the chi-square distribution with  $k - s - 1$  degrees of freedom, where  $s$  is the number of parameters of the hypothesized distribution estimated by the sample statistics. The hypotheses are the following:

$H_0$ : The RV  $X$  conforms to the distributional assumption

$H_1$ : The RV  $X$  does not conform.

The null hypothesis is rejected if  $\chi_0^2 > \chi_{\alpha, k-s-1}^2$ .

## Chi-Square Test

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	$\sqrt{n}$ to $n/5$

**Table:** Recommendations for # of Class Intervals for Continuous Data



## Chi-Square Test

Example 9.14: Consider the data in Example 9.2 and 9.7. The data appeared to follow a Poisson distribution; hence the parameter  $\hat{\alpha} = 3.64$ , was found. Thus the following hypotheses are formed:

$H_0$ : The RV  $X$  is Poisson distributed

$H_1$ : The RV  $X$  is not Poisson distributed

The PMF for the Poisson distribution is given as

$$p(x) = \begin{cases} \frac{e^{-\alpha} \alpha^x}{x!}, & x = 0, 1, 2, \dots \\ 0, & \text{otherwise} \end{cases}$$

## Chi-Square Test

For  $\alpha = 3.64$ , the probabilities associated with various values of  $x$  are obtained as

$$p(0) = 0.026$$

$$p(1) = 0.096$$

$$p(2) = 0.174$$

$$\vdots$$

$$p(10) = 0.003$$

$$p(\geq 11) = 0.001$$

## Chi-Square Test

From this information, Table 9.6 in the text (DESS) is constructed. The value of  $E_1$  is given by  $np_0 = 100(0.026) = 2.6$ . In a similar manner, the remaining  $E_i$  values are computed. Since  $E_1 = 2.6 < 5$ ,  $E_1$  and  $E_2$  are combined. In that case,  $O_1$  and  $O_2$  are also combined, and  $k$  is reduced by one. The last five classes are also combined, for the same reason, and  $k$  is further reduced by four. The calculated  $\chi_0^2 = 27.68$ . The degrees of freedom for the tabulated value of  $\chi^2$  is  $k - s - 1 = 7 - 1 - 1 = 5$ . Here,  $s = 1$ , since one parameter,  $\hat{\alpha}$  was estimated from the data. At the 0.05 level of significance, the critical value  $\chi_{0.05,5}^2 = 11.1$ . Thus,  $H_0$  would be rejected at the level of significance 0.05.

## 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 figuring out the probability associated with each interval that maximizes the power for a test of a given size. (The power of a test is defined as the probability of rejecting a false hypothesis.) However, if using equal probabilities, then,  $p_i = 1/k$ . We recommend,

$$E_i = np_i \geq 5 \Rightarrow \frac{n}{k} \geq 5 \Rightarrow k \leq \frac{n}{5}$$

The above expression was used in coming up with the recommendations for maximum number of class intervals.

## Chi-Square Test

Example 9.15 (Chi-Square Test for Exponential Distribution): The failure data in Example 9.11 appeared to follow an exponential distribution, and the parameter  $\hat{\lambda} = 1/\bar{X} = 0.084$  was computed. Thus, the following hypotheses are formed:

$H_0$ : the RV is exponentially distributed

$H_1$ : the RV is not exponentially distributed

In order to perform the chi-square test with intervals of equal probability, the endpoints of the class intervals must be found.

## Chi-Square Test

Above equations indicates that the number of intervals should be less than or equal to  $n/5$ . Here,  $n = 50$ , and so  $k \leq 10$ . It is recommended that 7 to 10 class intervals be used. Let  $k = 8$ , then each interval will have probability  $p = 0.125$ . The endpoints for each interval are computed from the CDF for the exponential distribution as

$$F(a_i) = 1 - e^{-\lambda a_i}$$

where  $a_i$  represents the endpoint of the  $i$ th interval,  $i = 1, 2, \dots, k$ . Since,  $F(a_i)$  is the cumulative area from zero to  $a_i$ ,  $F(a_i) = ip$ , so the above equation can be written as

$$ip = 1 - e^{-\lambda a_i} \Rightarrow e^{-\lambda a_i} = 1 - ip$$

## Chi-Square Test

Taking the log of both sides and solving for  $a_i$  gives a general result for the endpoints of  $k$  equiprobable intervals for the exponential distribution.

$$a_i = -\frac{1}{\lambda} \ln(1 - ip), \quad i = 0, 1, \dots, k$$

Regardless of the value of  $\lambda$ , the above equation will always result in  $a_0 = 0$  and  $a_k = \infty$ . With  $\hat{\lambda} = 0.084$  and  $k = 8$ ,  $a_1$  is computed from the equation as (see others in the text)

$$a_1 = -\frac{1}{0.084} \ln(1 - 0.125) = 1.590$$

Since  $\chi_0^2 = 39.6 > \chi_{0.05,6}^2 = 12.6$ , the null hypothesis is rejected.

## Kolmogorov-Smirnov Goodness-of-Fit Test

The Kolmogorov-Smirnov (K-S) test formalizes the idea behind examining a  $q - q$  plot. It is particularly useful when sample sizes are small and when no parameters are estimated from the data. When parameter estimates have been made, the critical values in Table A.8 are biased; in particular, they are too conservative.

$H_0$ : the inter-arrival times are exponentially distributed

$H_1$ : the inter-arrival times are not exponentially distributed

## Kolmogorov-Smirnov Goodness-of-Fit Test

We can show that if the underlying distribution of the inter-arrival times is exponential, then, the arrival times are uniformly distributed on the interval  $(0, T)$ . Hence, we first obtain the arrival times as  $T_1, T_1 + T_2, \dots, T_1 + \dots + T_{50}$ . We normalize the data to a  $(0, 1)$  interval so that we can apply the K-S test as before. Performing the same tasks as we did before, we obtain  $D^+ = 0.1054$  and  $D^- = 0.0080$ , and hence,  $D = 0.1054$ . The critical value of  $D$  is obtained from the K-S table for a level of significance  $\alpha = 0.05$  and  $n = 50$  is  $D_{0.05} = 1.36/\sqrt{n} = 0.1923$ . Since  $D < D_{0.05}$ , we cannot reject the null hypothesis.

## $p$ -Values and “Best Fits”

To apply a goodness-of-fit test, a significance level must be chosen. Recall that the significance level is the probability of falsely rejecting the null hypothesis. Due to the advances in computing, we can now specify a different level of significance, such as 0.07 etc. However, rather than require a pre-specified significance level, many software packages compute a  $p$ -value for the test statistic. This is the value of the significance level at which one would *just reject* the null hypothesis for the given value of the test statistic. Therefore, a large  $p$ -value tends to indicate a good fit (we would have to accept a large chance of error in order to reject), 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.

## Fitting a Non-Stationary Poisson Process

Fitting a non-stationary Poisson process (NSPP) to arrival data is a difficult problem, in general, because we seldom have knowledge about the appropriate form of the arrival rate function  $\lambda(t)$ . One approach is to choose a very flexible model with lots of parameters and fit it with a method such as maximum likelihood. Another one, that we consider here, is to approximate the arrival rate as being constant over some basic interval of time, such as an hour or a day or a month etc., but varying from time interval to time interval. The problem then becomes choosing the basic time interval and estimating the arrival rate within each interval.

## Selecting Input Models without Data

Unfortunately, it is often necessary in practice to develop a simulation model before any process data are available. There are a number of ways to obtain information about a process even if data are not available:

- Engineering Data
- Expert Option
- Physical or Conventional Limitations
- The Nature of the Processes

When data are not available, the uniform, triangular and beta distributions are often used as input models.

## Multivariate and Time-Series Input Models

Variables may be related, and if the variables are the inputs of a simulation model, the relationship should be investigated and taken into consideration. Let  $X_1$  and  $X_2$  be two RVs, and let  $\mu_i = E(X_i)$  and  $\sigma_i^2 = V(X_i)$  be the mean and variance of  $X_i$ , respectively. The covariance and correlation are measures of the linear dependence between  $X_1$  and  $X_2$ . The covariance and correlation between  $X_1$  and  $X_2$  is defined as

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

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

## Multivariate and Time-Series Input Models

Now suppose that we have a sequence of RVs  $X_1, X_2, \dots$  that are identically distributed, but could be dependent. We refer to such a sequence as a *time series* and to  $\text{cov}(X_t, X_{t+h})$  and  $\text{corr}(X_t, X_{t+h})$  as the *lag- $h$  auto-covariance* and *lag- $h$  auto-correlation*, respectively. If the value of the auto-covariance depends only on  $h$  and not on  $t$ , then, we say that the time series is *covariance stationary*.

## Multivariate and Time-Series Input Models

If  $X_1$  and  $X_2$  each are normally distributed, then dependence between them can be modeled by the bivariate normal distribution with parameters  $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$  and  $\rho = \text{corr}(X_1, X_2)$ . To estimate,  $\rho$ , assume that we have  $(X_{11}, X_{21}), (X_{12}, X_{22}), \dots, (X_{1n}, X_{2n})$  which are  $n$  IID pairs.

## Multivariate and Time-Series Input Models

The sample covariance and the correlation are

$$\begin{aligned}\widehat{\text{cov}}(X_1, X_2) &= \frac{1}{n-1} \sum_{j=1}^n (X_{1j} - \bar{X}_1)(X_{2j} - \bar{X}_2) \\ &= \frac{1}{n-1} \left( \sum_{j=1}^n X_{1j}X_{2j} - n\bar{X}_1\bar{X}_2 \right) \\ \hat{\rho} &= \frac{\widehat{\text{cov}}(X_1, X_2)}{\hat{\sigma}_1\hat{\sigma}_2}\end{aligned}$$

where  $\bar{X}_1$  and  $\bar{X}_2$  are the sample means, and  $\hat{\sigma}_1$  and  $\hat{\sigma}_2$  are the sample standard deviations.



## Multivariate and Time-Series Input Models

The following simple algorithm can be used to generate bivariate normal RVs.

- ① Generate  $Z_1$  and  $Z_2$ , independent standard normal RVs.
- ② Set  $X_1 = \mu_1 + \sigma_1 Z_1$ .
- ③ Set  $X_2 = \mu_2 + \sigma_2 \left( \rho Z_1 + \sqrt{1 - \rho^2} Z_2 \right)$ .

## AR(1) Model

If  $X_1, X_2, \dots$  is a sequence of identically distributed, but dependent and covariance-stationary RVs, then, there are a number of time series models that can be used to represent the process. We will describe two models that have the characteristic that the autocorrelations take the form

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

for  $h = 1, 2, \dots$

## AR(1) Model

Consider the time-series model

$$X_t = \mu + \phi(X_{t-1} - \mu) + \epsilon_t \quad (2)$$

for  $t = 2, 3, \dots$ , where  $\epsilon_2, \epsilon_3, \dots$  are IID normal with mean 0 and variance  $\sigma_\epsilon^2$  and  $-1 \leq \phi \leq 1$ . If the initial value of  $X_1$  is chosen appropriately, then,  $X_1, X_2, \dots$  are all normally distributed with mean  $\mu$ , variance  $\sigma_\epsilon^2/(1 - \phi^2)$  and  $\rho_h = \phi^h$ , for  $h = 1, 2, \dots$ . This time-series model is called the autoregressive order-1 model, or AR(1) for short. Estimation of the parameter  $\phi$  can be obtained from the fact that  $\phi = \rho^1 = \text{corr}(X_t, X_{t+1})$  the lag-1 autocorrelation.

## EAR(1) Model

$$X_t = \begin{cases} \phi X_{t-1}, & \text{with probability } \phi \\ \phi X_{t-1} + \epsilon_t, & \text{with probability } 1 - \phi \end{cases} \quad (3)$$

for  $t = 2, 3, \dots$ , where  $\epsilon_2, \epsilon_3, \dots$  are IID exponential with mean  $1/\lambda$  and  $0 \leq \phi \leq 1$ . If the initial value  $X_1$  is chosen appropriately, then,  $X_1, X_2, \dots$  are all exponentially distributed with mean  $1/\lambda$  and  $\rho_h = \phi^h$ , for  $h = 1, 2, \dots$ . This time-series model is called the exponential autoregressive order-1 model, or EAR(1) for short. Only autocorrelations greater than 0 can be represented by this model. Estimating of the parameters proceeds as for the AR(1) by setting  $\hat{\phi} = \widehat{\rho^1}$ , the estimated lag-1 autocorrelation, and setting  $\hat{\lambda} = 1/\bar{X}$ .

## EAR(1) Algorithm

- ❶ Generate  $X_1$  from the exponential distribution with mean  $1/\lambda$ . Set  $t = 2$ .
- ❷ Generate  $U$  from the uniform distribution on  $U[0, 1]$ . If  $U \leq \phi$ , then, set  $X_t = \phi X_{t-1}$ . Otherwise, generate  $\epsilon_t$  from the exponential distribution with mean  $1/\lambda$  and set  $X_t = \phi X_{t-1} + \epsilon_t$ .
- ❸ Set  $t = t + 1$  and go to Step 2.

## The Normal-to-Anything Transformation

The bivariate normal distribution and the AR(1) and EAR(1) time-series models are useful input models that are easy to fit and simulate. However, the marginal distribution is either normal or exponential, which is certainly not the best choice for many applications. Fortunately, we can start with a bivariate normal or AR(1) model and transform it to have any marginal distributions we want (including exponential). We refer this as the *normal to anything transformation*, or NORTA for short. See the text (DESS) for more details.

## Summary

- Reading HW: Chapter 9.
- Chapter 9 Exercises.