1 Simulation: Introduction

Simulation is a lot like masturbation. If you do it too much, you begin to think it is the real thing.

1.1 Why and what - raison d'$\widehat{\text{e}}$tre.

Why simulate? What is simulation? This chapter endeavors to present the basic underpinnings of simulation so that subsequent chapters can address specific applications that build from these initial concepts. *Simulation imitates reality.* The reasons to simulate are many but they may be rolled into a desire to *better understand a system that is too complex, expensive, or dangerous to manipulate in the real world.*

Hydrologists studying contaminant transport in an aquifer would like to assess the likelihood of different flow paths and possible travel times. Mining and petroleum engineers are asked by management to quantify the risks and potential payoffs in the geologic resources buried in the earth. Industrial engineers want to reduce product cycle times, minimize scrap, improve staffing allocation, and increase profitability. The system to be studied may be a groundwater system, geologic depositional system, or a manufacturing/office system among many other possibilities.

A frequent objective is to understand and possibly modify the system to achieve desired goals. Different alternatives can be evaluated. For the hydrology study of contaminant transport, it is not unlikely that the full investigation would ascertain the impact of possible extraction wells including finding "optimal" placements for such. Mining and petroleum engineers want to assess the financial return for a given prospect. If the prospect is deemed to be worthwhile, how should it be best approached to address risk and financial concerns? The industrial engineer would like to investigate different scenarios that cannot be easily tested in the real world.

Thus the "why simulate" issue is based on the basic human need to understand and/or manipulate systems. But what is simulation? Unfortunately a simple global answer is not possible as different disciplines view simulation differently. However, *all simulations form a model of the system of interest.* It may not be so obvious in all simulations what the model is; however, in everything that is worth simulating in our opinion the model and its assumptions should be well understood. Successful application of simulation requires the ability to change the model and see if it better matches reality. Once a reasonable agreement between the simulation model and reality is obtained, then the simulation can be used to aid in decision making. Keep in mind that the simulation is attempting to mimic the real system and will never be a full replacement for it regardless of how much verification/validation of the model is undertaken.

What do we have to know before we simulate? Bottom line: a lot. Simulation without understanding the behavioral parameters and correct relationships generally leads to trouble. Getting dragged into lengthy litigation as the result of poor simulation modeling is not our recommendation for a good time. Doing one's homework is essential. A reasonable starting point (among many other options) are chapters 1-6 of *Practical Geostatistics 2000*. Understanding basic statistical concepts and the associated potential pitfalls are essential to anyone that attempts simulation.

1.1.1 Classifications of Simulation

There are a variety of ways to pigeon-hole simulation. Many of these are very specific and will be addressed later in this book. Some of the more general ways that simulation can be classified are briefly addressed in the paragraphs that follow.

Static vs. Dynamic: Is time a factor in your system? Can steady-state conditions be assumed? Time is a natural element in dynamic systems. If the system being modeled is a confined aquifer in which potentiometric head values are not rapidly changing as a function of time, then a static approach may be warranted. Time was an important component impacting the depositional patterns of the diamonds; however, geologic time works over such an incredibly long time span in human terms that the current location of the diamonds can be viewed as a static problem. If the system is water levels in the Mississippi river throughout the entire year, then time is an important component. It depends on the focus of the modeling effort. If one is focusing on a snapshot in time, then a static approach is acceptable.

How about a business operation? It depends. If the goal is to model the gambling operation of a roulette wheel, then the probabilities are static since the odds do not change. The term "Monte Carlo simulation" often refers to static simulations. Waiting time probabilities to get a drink at the casino's bar however are dynamic since the waiting time is a function of the state of the system, e.g., the number of people in line in front of you, number of bartenders on duty, etc.

Continuous vs. Discrete: Does the state of the system change continuously over time or in discrete jumps? The amount of oil recovered is often a continuous function of time. The state of many manufacturing systems changes based on the arrival/departure of parts, staffing changes, machine breakdowns, etc. These events such as arrivals/departures occur at discrete points in time and can be modeled with discrete-event simulation. Mixed systems have elements of both continuous and discrete.

Continuous and discrete can be applied in different ways other than time based. In spatially distributed geologic systems the item of interest may change continuously over space such as groundwater levels or mineral concentrations.

Continuous modeling can also incorporate discrete discontinuities such as geologic faults. Diamonds, on the other hand, generally must be modeled using discrete techniques.

Deterministic vs. Stochastic: A system having no random inputs is deterministic. Systems relying totally on random inputs are stochastic. Many systems have both deterministic and random inputs but are still classified as stochastic. In a business environment the time to perform a given task is often modeled by a statistical distribution that is used to generate random values while staffing changes are often deterministically modeled as occurring at set times. Geologic systems are often modeled as a combination of a stochastic spatial component with a deterministic trend.

Spatial vs. Nonspatial: By the nature of this book most readers will be interested in simulating spatially distributed phenomena. For such the spatial analysis as covered in basic geostatistics texts must be an integral part of the overall simulation analysis. For most of the rest of the world simulation is nonspatial. Many of the nonspatial applications fall into what is termed discrete-event simulation which is overviewed in Chapter XXX with a mining example in which daily operations of a mine are simulated including the breakdown of equipment.

1.2 Where do the random numbers come from?

Most simulation modeling is based on logical mathematical models and not on physical models (e.g., a scale model of a river). This type of simulation analysis is solved using appropriate computer software. The variability and uncertainty inherent in the systems studied is modeled using random inputs. How are these random numbers generated? How does one decide what distributions to use? Can you have faith in your random number generator? This section will not fully address these concerns but will discuss some aspects concerning the birth of the random numbers.

Selecting an appropriate statistical distribution for a given application is both an art and a science. Knowledge of what is being modeled, knowing when results look reasonable, and blending this with the goodness-of-fit aspects of statistical science requires one to use statistics as an aid but not a diviner of truth. Blindly selecting distributions based on passing hypothesis tests with certain p-values is an unjustified act of faith. Visual checks are important aspects of assessing what is satisfactory.

The random numbers referred to in this book are more corrected identified as pseudo-random numbers since each is generated according to some specified algorithm. The hope is that the numbers satisfy two general criteria: 1) they appear to be independent; 2) they follow the desired statistical distribution. This is discussed in more detail later in this chapter.

1.2.1 Linear Congruential Method of generating Uniform(0, 1) random numbers

Underlying the generating of random numbers for any statistical distribution is the ability to generate Uniform (0, 1) pseudo-random numbers. The Uniform (0,1) continuous distribution covers the interval from 0 to 1 with equal probability. If a good method of creating Uniform (0,1) numbers is available, then other distributions can be developed from these Uniform (0,1) pseudo-random numbers.

One of the most common methods for generating Uniform (0,1) numbers is the Linear Congruential Method. From a starting seed X_0, a series of X_i are generated and converted to R_i that are then pseudo-random Uniform (0, 1) numbers as indicated below.

$X_{i+1} = (aX_i + c) \bmod m$, $i = 0, 1, 2, ...$

$R_i = X_i \ / m$, $i = 1, 2, ...$

The development of good Uniform (0,1) numbers is contingent on the selection of the constant multiplier a, the increment c, and the modulus m and is beyond the scope of this book. The intent here is to show a simple application of the this method. Setting the initial seed X_0= 27, a = 17, c = 43, and m = 87,653 (no particular reason for these selections), we generate the 20 pseudo-random hopefully Uniform (0,1) numbers rounded to 3 decimal places seen in Table 2.1. The mod function returns the integer remainder after division. Some sample calculations below should help explain this procedure. An Excel spreadsheet contained on the CD version of this book or downloadable via the web allows one to experiment with different settings of X_0, a, c, and m.

$X_1 = 17 * X_0 + 43 \bmod 87,653 = 502 \bmod 87,653$. Doing integer division to get a remainder (what the mod function is doing) results in 502/(87,653) = 0 with a remainder of $X_1 = 502$. $R_1 = X_1/m = 502/(87,653) = 0.006$ when rounded to 3 decimal places. The rounding is shown below only for cosmetic purposes. In practice the unrounded result would be used.

$X_2 = 17 * X_1 + 43 \bmod 87,653 = 8577 \bmod 87,653$ which results in X_2 = 8577 since the modulus 87,653 is larger than 8577 and thus goes into it zero times doing integer division leaving the entire 8577 as the remainder. $R_2 = X_2/m = 8577/(87,653) = 0.098$ when rounded to 3 decimal places.

$X_3 = 17 * X_2 + 43 \bmod 87,653 = 145,852 \bmod 87,653$. Now when we perform the integer division we find that 87,653 goes into 145,852 once leaving a remainder of X_3 = 58,199. $R_3 = X_3/m = (58,199)/(87,653) = 0.664$ when rounded to 3 decimal places.

$X_4 = 17 * X_3 + 43 \bmod 87,653 = 989,426 \bmod 87,653$. We find that 87,653 goes into 989,426 eleven times leaving a remainder of X_4 =25,243. $R_4 = X_4/m = (25,243)/(87,653) = 0.288$ when rounded to 3 decimal places.

It is the R_i below that are the resulting pseudo-random Uniform (0, 1) numbers desired. The X_i are just an intermediate step. The twenty X_i and R_i are shown in Table 2.1 below.

X_o=	27
c=	43
a=	17
m=	87653

i	$\mathbf{X}_i$	$\mathbf{R}_i$
1	502	0.006
2	8577	0.098
3	58199	0.664
4	25243	0.288
5	78562	0.896
6	20802	0.237
7	3065	0.035
8	52148	0.595
9	10029	0.114
10	82883	0.946
11	6606	0.075
12	24692	0.282
13	69195	0.789
14	36869	0.421
15	13245	0.151
16	49902	0.569
17	59500	0.679
18	47360	0.540
19	16286	0.186
20	13946	0.159

Table 2.1 Linear Congruential generated pseudo-random numbers.

Later we discuss how to use these U(0, 1) pseudo-random numbers to generate other statistical distributions. Before we do that it is worthwhile to address some aspects of assessing the quality of the U(0, 1) random numbers generated.

1.3 Testing Uniform (0, 1) random numbers

Are the numbers generated in the prior section good pseudo-random numbers? To individuals that study random number generators in detail, good generally implies that the generator passes two major classes of tests. The first is a test of independence of the numbers being generated and the second is a distributional test. The former wants to make sure that the sequence of the numbers (or possibly digits within the numbers) being generated does not result in an easily predictable pattern to an outside observer and that the numbers are not obviously related to each other based on the order in which they are generated. For example, generating numbers .001, .002, .003, ... may make the output indeed

look like it is coming from a Uniform (0, 1) distribution but the numbers are easily seen not to have been generated in random order. There are many tests that fall into this class, but our favorites are the runs up and down test and the Poker test (simple 3 digit version). Excellent references for both class of tests are Banks, Carson, Nelson, and Nichol [*Discrete-Even System Simulation*, 3rd edition by Jerry Banks, John S. Carson II, Barry L. Nelson, and David M. Nicol, ISBN 0-13-088702-1, Prentice-Hall, New Jersey, 2001], Law and Kelton [*Simulation Modeling and Analysis*, 3rd Edition, by Averill M. Law and W. David Kelton, ISBN 0-07-059292-6, McGraw-Hill, New York, 2000], and Banks [Jerry Banks, Editor, Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice, ISBN 0-471-13403-1, co-published by Engineering and Management Press and John Wiley & Sons, Inc., New York, 1998].

The second class of tests is a distributional test (or more specifically for this distribution a uniformity test) to see if the output of the generator fits the desired distribution. In this class of tests, the order in which the numbers are generated is not important but rather an evaluation of a full set of the numbers generated. Our two favorite are the Chi-Square and the Kolmogorov-Smirnov goodness-of-fit tests. The references in the prior paragraph are good places to start studying these classes of tests. Distributional testing will be covered later in this chapter after the Inverse Transform Technique is used to generate other statistical distributions from U(0, 1) numbers.

This section will address two particular independence tests for U(0, 1) random numbers: the Runs Up and Down test; and the Poker test. The independence class of tests involves statistical hypothesis testing. These tests are tools to aid in determining whether a given random number generator is performing adequately. The statistical hypothesis may be stated as follows:

H_0: U(0, 1) numbers generated are independent

H_a: numbers are not independent

The null hypothesis, H_0 assumes that the random number generator is generating what appears to be random U(0, 1) numbers. This is the equivalent to the United States legal system assumption of "Innocent until proven guilty beyond a reasonable doubt." The reasonable doubt legal analog is the setting of an α level that provides a degree of protection of mistakenly rejecting H_0 in favor of the alternative hypothesis H_a declaration that the random number generator is not performing adequately. Mathematically, $\alpha = P(rejecting\, H_0 \,|\, H_0\ true)$ which is the probably of making what statisticians call a Type I error. In other words, in testing the null hypothesis we allowing for the possibility of making a certain type of bad decision. We will not cover the Type II error here which is not rejecting H_0 when it is false - many statistical texts cover this. If α is set to 0.05, then 95% of the time when H_0 is true, it will not be rejected and 5% [or $(100 * \alpha)\%$] of the time we will mistakenly make a Type I error. Without delving into the depths of the statistical issues here, it is important to realize that a trade-off exists between making both kinds of errors.

1.3.1 Runs Up and Down Test for Independence of U(0, 1) random numbers

The runs up and down test is performed on the assumed random U(0,1) numbers by comparing the adjacent values in the order of their generation. A given random number is compared to the prior random number to see if it is larger or smaller. If it is larger than the prior random number, a "+" is used to illustrate this as seen in Figure 2-1 using the random numbers generated earlier. Likewise a "-" indicates that the given number is smaller than the prior number. The first random number is unlabeled as it cannot be compared to any prior number. A run is a succession of like signs. In Figure 2-1 are seen six runs up represented by a contiguous sequence of +'s and six runs down represented by a sequence of -'s. Let n represent the number of random numbers tested and a the number of runs found in the data. For this example, $n = 20$ and $a = 12$. A hypothesis test for randomness may be carried out as indicated below when $n \geq 20$. μ_a is the expected number of runs if the numbers were truly randomly generated (i.e., under the assumed null hypothesis H_0) while σ_a is the theoretical standard deviation of the number of runs. If the null hypothesis is true, then the test statistic Z_0 is asymptotically a standard normal distribution meaning that the larger the n gets, the more normal Z_0 becomes.

$$\mu_a = \frac{(2n)-1}{3}$$

$$\sigma_a = \left[\frac{(16n)-29}{90}\right]^{\frac{1}{2}}$$

$$Z_0 = \frac{a-\mu_a}{\sigma_a}$$

To test the twenty numbers generated at the 95% confidence level (implies that $\alpha = 0.05$) leads to a two-tailed critical value of $z_{\frac{\alpha}{2}} = 1.96$. Thus if H_0 is true, then 95% of the time the following will be true: $-z_{\frac{\alpha}{2}} \leq Z_0 \leq z_{\frac{\alpha}{2}}$. Carrying out the computations results in $\mu_a = 13.0$, $\sigma_a = 1.798$. As seen from Figure 2.1 where the runs are "boxed in", the number of runs is $a = 12$ which leads to $Z_0 = -0.556$. Since Z_0 falls in between -1.96 and 1.96, there is insufficient evidence to reject H_0 and thus we have no reason to disbelieve the randomness of the numbers generated from the runs up and down test. In practice, it is easy to generate many random numbers and perform the runs up and down test with a larger sample size.

1.3.2 Poker test of independence of U(0, 1) random numbers

There is a class of tests for independence called poker tests. We show only the 3 digit version. In this case the generated random numbers are rounded

R_i	Runs
0.006	
0.098	+
0.664	+
0.288	-
0.896	+
0.237	-
0.035	-
0.595	+
0.114	-
0.946	+
0.075	-
0.282	+
0.789	+
0.421	-
0.151	-
0.569	+
0.679	+
0.540	-
0.186	-
0.159	-

Figure 1: Figure 2-1. Runs Up and Down Test.

to 3 digits. Then each random number is classified into one of the following categories:

1. All digits are different from each other.
2. All digits are identical.
3. There is exactly one pair of identical digits.

In this test the sequence from one random number to another is not the focus. Rather this test focuses on the internal digits within a given number. If the digits are truly generated randomly, the following results hold.

P(all 3 digits different) = P(second digit different from the first)
* P(third digit different from both the first and second) = 0.9 * 0.8 = 0.72

P(all 3 digits identical) = 0.1 * 0.1 = 0.01

P(exactly one pair in the 3 digits) = 1 - (0.72 + 0.01) = 0.27

The Poker test uses the chi-square distribution to assess if statistically significant deviations of independence in the digits has occurred. The twenty random numbers used are too few for a fair comparison; however, their use is illustrative of how the poker test is performed. Figure 2-2 has a column called Poker Category with three potential values: 0 for zero matching digits; 2 for a pair of matching digits, and 3 for all three being identical. The chi-square test compares the observed frequency of each of these groupings to the theoretical or expected frequency which is obtained by multiplying the sample size by the corresponding theoretical probability. This is illustrated in Figure 2-3 below.

The column called x in Figure 2-3 represents the possible outcomes of the three digit random numbers as described in the prior paragraph. The column Pr(x) is the corresponding theoretical probability given earlier. The Obs Freq is the actual or observed frequency of each outcome while the Exp Freq is the product of the sample size 20 and the associated probability. For example the expected frequency for no matching digits (or row x = 0 above) is 20 * 0.72 = 14.4. The final column labeled Chi-sq is the $(ObsFreq - ExpFreq)^2/ExpFreq$ = 0.136 for the first row. The sum of this last column is compared to a chi-square distribution with 2 degrees of freedom. At the $\alpha = 0.05$ level, the critical value of such a chi-square is 5.99 as found in a chi-square table such as Appendix Table 3. If the observed value of the chi-square (0.810 in this case) exceeds the tabled or critical chi-square then the null hypothesis of independence is rejected. Otherwise there is insufficient evidence for rejection. For this particular case, there is no reason to disbelieve independence based on the poker test. This is a stretch because of the insufficient sample size. Statistician have various rules of thumb. One is that for a chi-square test, the expected frequency must be at least 3 (many say 5). Since the probability of having three identical digits is only 0.01, that implies we must have at least 300 (or perhaps 500) random numbers to have a sufficient poker test. Since it is easy to generate as many numbers as desired from a random number generator, this is easy to satisfy.

R_i	Poker Category
0.006	2
0.098	0
0.664	2
0.288	2
0.896	0
0.237	0
0.035	0
0.595	2
0.114	2
0.946	0
0.075	0
0.282	2
0.789	0
0.421	0
0.151	2
0.569	0
0.679	0
0.540	0
0.186	0
0.159	0

Figure 2: Figure 2-2. Classifications for Poker Test.

x	Pr(x)	Obs Freq	Exp Freq	Chi-Sq
0	0.72	13	14.4	0.136
2	0.27	7	5.4	0.474
3	0.01	0	0.2	0.200
		20	20	0.810

Figure 3: Figure 2-3. Poker Test Chi-Square Analysis Results.

1.4 Creating other statistical distributions from the Uniform(0,1)

Once it has been established that a good Uniform (0, 1) random number generator is available, then other statistical distributions may be generated from the stream of Uniform (0, 1) pseudo-random numbers. There are numerous ways to do this as documented in the references given earlier. In this chapter we will cover only one general method called the Inverse Transform Technique. For the Normal distribution we will also examine several specific ways of generating this important distribution. All of the methods of generating other distributions rely on having good Uniform (0, 1) random numbers available.

Once understood, the basic concept of the Inverse Transform Technique is seen to be clever, simple in basic approach, and also quite elegant. The more familiar you are with the cumulative distribution function, F(x), the sooner you will appreciate this general approach. Let's step back into probability and review f(x) and F(x). For discrete distributions, the random variable X (do not get a random variable mixed up with the random numbers we are generating!) takes on particular values x with probability f(x). Thus $0 \leq$ f(x) ≤ 1. The cumulative distribution function, F(x), sums f(k), for k $\leq$ x. Thus $0 \leq$ F(x) ≤ 1 also. But f(x) is only nonzero for particular values of x while F(x) is defined over the entire range $-\infty \leq x \leq \infty$. For continuous distributions f(x) is no longer bounded by 1.0 on the upper end. Additionally f(x) is defined for some continuous portion (if not all) of the real number line. Thus there are an infinite number of possible x values when the random variable X is continuous. Instead of summing f(x) to get F(x) as with discrete distributions, F(x) requires calculus to integrate f(x) over any desired range.

1.4.1 Discrete Distribution generation with the Inverse Transform Technique

A brief example of the above is a binomial distribution for the number of heads in three tosses of a fair coin (implying the probability of success of getting a head on any toss is p = 0.50). It can easily be shown (or looked up in Table 9 in the statistical appendix) that f(0) = 0.125, f(1) = 0.375, f(2) = 0.375, and f(3) = 0.125. For any value other than x = 0, 1, 2, or 3, f(x) = 0. The cumulative distribution function, F(x), for this has F(x) = 0 until x = 0 and then it takes a step jump at x = 0 to 0.125, i.e., F(0) = 0.125. F(x) stays at 0.125 until x = 1 and then it takes a step jump at x = 1 with F(1) = f(0) + f(1) = 0.125 + 0.375 = 0.500. Carrying this out over the interval -0.5 $\leq x \leq$ 3.5, the results are shown graphically in Figure 2-4. While f(0) may not be so clear since f(0) = F(0) = 0.125, f(x) has only four non-zero values at the discrete points x = 0, 1, 2, and 3. F(x) is non-zero for all x $\geq$ 0.0.

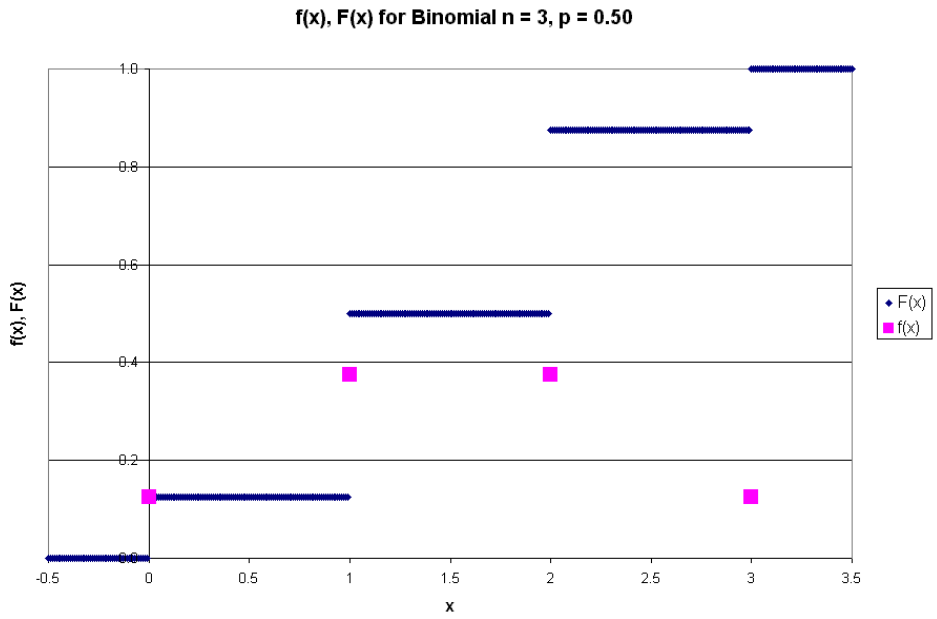

Figure 4: Figure 2-4. Binomial Distribution f(x) and F(x) for n = 3, p = 0.50.

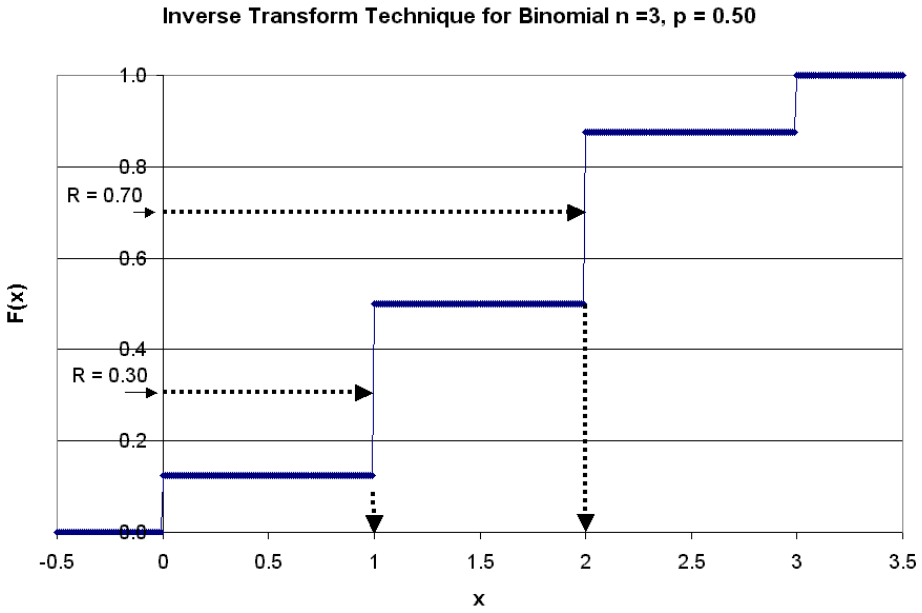

Figure 5: Figure 2-5. Inverse Transform for Binomial Distribution.

Figure 2-5 is an illustration of how the Inverse Transform technique is applied to this particular situation. The stream of random U(0, 1) numbers are represented by the letter R in Figure 2-5 below. Two realizations are shown. For one, R = 0.70 is the random Uniform (0, 1) number. A horizontal line is drawn through the vertical axis at F(x) = 0.70. The horizontal line hits the vertical line representing the step function change of F(x) at x = 2. Imagine the random value 0.70 immediately falling perpendicularly to the horizontal x axis after encountering the vertical line. The place where it lands is at x = 2 which represents the resulting random binomial variate. In a similar vein, the random Uniform (0, 1) R = 0.30 is converted using the Inverse Transform technique to a random binomial x = 1.

The basic principle underlying the Inverse Transform technique is that F(x) is bounded on the range [0, 1] for any statistical distribution; whereas, f(x) is not for continuous distributions. Letting R represent the generated Uniform (0, 1) values, the Inverse Transform technique sets R = F(x) and then finds $x = F^{-1}(R)$ for the desired resulting distribution. This can be easily solved graphically for any discrete distribution as shown in Figure 2.5 for the binomial distribution. Continuous distributions are not as easy to handle and in some cases cannot be generated using this methodology.

Another way to examine this is in the form of Table 2-2. For the range in column R = F(x), the resulting x value for this Binomial distribution with n = 3 and p = 0.50 can be seen. If a set of cumulative probability tables are handy for a given discrete statistical distribution, then one can easily see what a given R value would generate for any tabulated cumulative discrete distribution. Caution must be taken to ensure that the table is giving cumulative not individual probabilities.

R = F(x)	Results in Binomial value x using the Inverse Transform Technique
[0, 0.125)	0
[0.125, 0.5)	1
[0.5, 0.875)	2
[0.875, 1.0]	3

Table 2-2. Tabular Inverse Transform Technique for Binomial example

1.4.2 Continuous Distribution generation with the Inverse Transform Technique

Continuous distributions generally require the use of calculus to use the Inverse Transform Technique. Whereas $F(x) = \sum_{k \le x} f(k)$ for discrete distributions, $F(x) = \int_{-\infty}^{x} f(t)\,dt$ for continuous distributions. To find F(x) it is often necessary to integrate f(x) unless it is readily available in a text book. Some distributions, e.g., the Normal distribution, cannot be integrated. Usually other approaches mentioned later are used to generate Normal random numbers. There are readily available polynomial approximations to the Normal that could be used to generate an approximate F(x) which can be made very accurate depending on the polynomial selected (see Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, National Bureau of Standards, John Wiley & Sons, New York, NY, 10th printing with corrections, 1972, ISBN 0-471-80007-4 edited by Milton Abramowitz and Irene A. Stegun). Another non-automated way of generating Normal random numbers from Uniform (0, 1) numbers would be to take advantage of the fact that the Normal F(x) is well tabulated for a standard Normal (mean of 0, standard deviation of 1). However we will focus on another distribution to illustrate the application of the Inverse Transform technique to continuous distributions.

The exponential distribution is a simple continuous statistical distribution for applying the Inverse Transform technique. The exponential distribution has only one parameter that is often represented in one of two different forms.

In the formulation below the mean of the distribution is $1/\lambda$ where the other formulation uses θ as the mean where $\theta = 1/\lambda$. The actual Greek letters used are immaterial, but the probability density function formulation [$f(x) = \lambda e^{-\lambda x}, x \geq 0$; 0; otherwise] is based on the mean being $1/\lambda$. The basic steps to create a continuous random number generator using the Inverse Transform technique are illustrated below for the exponential distribution. In these steps, R represents the pseudo-random U(0,1) number generated by a method such as the linear congruential method.

Step 1. Compute or look up in a book the cumulative distribution function (c.d.f.) for the distribution of interest. This often involves integrating the probability density function (p.d.f.) In this particular case integration of $f(x)$ results in $F(x) = 1 - e^{-\lambda x}, x \geq 0$.

Step 2. Set $R = F(x)$ for the applicable range of x. For the exponential distribution, this results in $R = 1 - e^{-\lambda x}$ for the range $x \geq 0$.

Step 3. Solve $R = F(x)$ for x or in other words find $x = F^{-1}(R)$ where F^{-1}means solving the equation for x and does not mean $1/F$. Solving for x below results in an equation providing a random number generator for an exponential distribution given the available U(0, 1) random R values. This formula for x can then be used to develop as many random exponential numbers as desired.

$R = 1 - e^{-\lambda x}$
$e^{-\lambda x} = 1 - R$
$-\lambda x = \ln(1 - R)$
$x = -\frac{1}{\lambda} \ln(1 - R)$

The steps above can get more complicated depending on the distribution involved. But the Inverse Technique is one of the most common methods to generate other random distributions from U(0, 1) random numbers. In concept, it is simple and quite elegant.

Suppose one wanted to manually use the Inverse Transform Technique to get a Normal distribution. Since there is no closed form equation for a normal F(x), one can use a normal table. When using a standard cumulative Normal table, caution is needed to understand how the particular normal table is formulated. It may also be necessary to take advantage of the fact the normal is symmetric about its mean. In the examples that follow the standard normal table used is Table 1 from Practical Geostatistics 2000 in which F(x) is formulated only for positive z values.

Example 1: Generate a normal ($\mu = 100$, $\sigma = 20$) for R = 0.9131. Search Table 1 in the statistical appendix until an F(x) value of 0.9131 is found or bounded by two adjacent F(x) values. For the normal distribution, F(x) = F(z) where $z = \frac{x-\mu}{\sigma}$. In this case, this corresponds to z = 1.36 or x = $\sigma z + \mu = 20 * 1.36 + 100 = 127.2$ as the random normal variable.

In some normal tables, only the probability of falling between 0.0 and some positive z value is given rather than F(z). If the table is so configured, then you would look up F(z) = 0.4131 and find z = 1.36. When using normal tables like this, keep in mind that 50% of the distribution is below the standard normal mean of 0.0.

Example 2: Generate a normal($\mu = 100$, $\sigma = 20$) for R = 0.9139. Since this F(x) or F(z) value falls in-between two values in the table, linear interpolation could be used to get a corresponding z value and then x value. In this case this R value falls exactly 50% of the way between F(x) = 0.9131 and F(x) = 0.9147 and thus z = 1.365 with a resulting x value of $x = \sigma z + \mu = 20 * 1.365 + 100 = 127.3$.

Example 3: Generate a normal($\mu = 100$, $\sigma = 20$) for R = 0.1038. If the normal table lists negative z values, then follow the steps in the earlier examples. However, many normal tables list only positive z values such as Table 1 in the statistical appendix. In this case, one must use symmetry to find the desired z values. A negative z value corresponding to having the probability of being below it of 0.1038 is symmetric about the standard normal mean of 0.0 with respect to a positive z value with the same magnitude having the probability 0.1038 of being exceeded. Once again, be careful of how the normal table is formulated. Being careful ensures obtaining the same correct answer. In our case z = 1.26 has a 0.8962 probability of a standard normal falling below it and thus a 1 - 0.8962 = 0.1038 probability of being exceeded. Using symmetry, this implies that z = - 1.26 corresponds to F(z) = 0.1038. Finally $x = \sigma z + \mu = 20 * (-1.26) + 100 = 74.8$.

Thus given a stream of U(0, 1) random values, one could generate random normal($\mu = 100$, $\sigma = 20$) values. This would be tedious to do by hand; however, it is easy in Excel using the following formula:

NORMINV(Cumulative Probability, Mean of Normal, Standard Deviation of Normal)

In this formula, the first argument is a cumulative probability. The second and third parameters are the population mean and standard deviation of the normal distribution. For example, NORMINV(0.1038, 100, 20) = 74.79617377 depending on how many decimal digits are formatted in Excel. Rounding this gives the 74.8 manually generated in Example 3 above. An example version of using the Inverse Transform Technique would be to generate a series of U(0, 1) numbers using the Excel RAND() function and then letting this be the Cumulative Probability argument in NORMINV, i.e., use NORMINV(rand(), 100, 20). Try generating 100 normal random numbers in Excel for a mean of 100 and standard deviation of 20. This will illustrate the ease of application of the Inverse Transform Technique for the normal distribution in Excel.

1.4.3 Special methods for Generating random Normal values

Note to ISE 521 Students: You are not responsible for this section on special ways of generating Normal pseudo-random values. You might have to know the subsequent remaining material starting in Section 1.5 depending on how the course is structured this quarter. ISE 704 students definitely have to know Section 1.5 through the end of the chapter.

This section of the chapter focuses on generating both univariate and multivariate normal pseudo-random values. The normal distribution is perhaps the most important distribution in probability and statistics and much research has been done on ways to efficiently generate random normal values. The first subsection below focuses on univariate generation in which a single distribution is being randomly generated, e.g., height of male college students. The subsequent multivariate section focuses on generating multiple possibly correlated normal distributions such as college male height and weight.

Univariate This section covers several ways to create random normal(0,1) values. This is the so called standard normal with mean 0 and standard deviation 1 that is usually in some tabulated form in most statistics texts. If good standard normal values can be generated, then these can easily be converted to a normal with any desired mean and standard deviation as was covered in the section on the generation of continuous distributions using the Inverse Transform Technique.

One of the original methods (e.g., it is the method used in *The GASP IV Simulation Language* by A. Alan B. Pritsker, John Wiley & Sons, New York, pp. 165-167) used to generate approximate standard normals used 12 pseudo-random Uniform (0, 1) values. This is still in use in many places though the academic simultion world now generally favors methods such as the ones covered below. This method relies on the central limit theorem (perhaps the most taught statistical theorem in existence) which roughly states that the mean of values from almost any distribution tends toward normality as the numbers of values summed increases. Questions such as are 12 values sufficient to approximate a normal might be issues, but also there are speed issues involved. To implement this procedure, one must first generate 12 Uniform (0, 1) values. Then it was necessary to compute their sum and subtract 6. The rationale for this is that the Uniform (0,1) distribution has a mean of 0.5 and a variance of $\frac{1}{12}$. The distribution from summing 12 Uniform (0,1) numbers (or R_i) would have a mean of 12 * 0.5 = 6, and a standard deviation of $\sqrt{(12 * \frac{1}{12})} = 1.0$. The resulting ratio $\frac{(\sum R_i) - 6}{1.0}$ under the central limit theorem would be tending toward normality with a mean of 0.0 and a standard deviation of 1.0. For many applications, this method has been found to perform adequately. However, unlike the following methods it is an approximation with maximum differences in the probability density function $f(x)$ of about 0.0050 and about 0.0023 in the cumulative distribution function $F(x)$ compared to an exact normal distribution.

The Box-Muller (Box, G.E.P., and Muller, M.E., 1958, A note on the generation of random normal deviates, Annals of Mathematical Statistics, Vol 29, pp 610-611) is one of the best known univariate normal generators today and like the following normal generators is an exact method. This method uses two Uniform (0, 1) and can generate two standard normals though in practice using

only one is generated at a time. Below are the steps in the Box-Muller standard normal generation.

1. Generate R_1(Uniform (0, 1) value), and set $\theta = 2\pi R_1$.

2. Generate R_2, set E = - ln R_2, $W = \sqrt{2E}$.

3. Compute $X = W \cos\theta$, and (if desired) compute $Y = W \sin\theta$. X, Y are independent standard normal deviates.

The Box-Muller method is derived theoretically by letting the pair (W, θ) be (X, Y) in polar coordinates for a bivariate standard normal probability density function. A criticism sometimes leveled at this algorithm is the use of the trignometric functions which slows down the computation considerably. Below is another polar method label derived by Marsaglia and Bray (Marsaglia, G., and Bray, T.A., 1964, A convenient method for generating normal variables, SIAM Review, Vol 6, pp 260-264).

The Marsaglia-Bray steps are as follows:

1. Repeat the generation of V_1, V_2 until $W = V_1^2 + V_2^2 < 1$. V_1, V_2 are Uniform(-1, 1) which can be generated from the Uniform(0, 1) R_i by setting $V_i = -1 + 2R_i$.

2. Compute C = $\sqrt{-2W^{-1} \ln W}$.

3. Let $X = CV_1$,and (if desired) $Y = CV_2$.

X and Y are two independent standard normal values. Often only X is computed. While this may require more iterations, it is much faster than the Box-Muller since it does not require the use of trignometric functions. When developing algorithms for the generation of random numbers, speed is more important than complexity. Marsaglia and others have developed more complex but speedier generators for normal deviates that are covered the book *Stochastic Simulation* by Brian D. Ripley (John Wiley & Sons, Inc., New York, 1987, ISBN 0-471-81884-4). This book is an excellent source for algorithms for the generation of random numbers.

Note: The Handbook of Simulation edited by Jerry Banks, chapter 5, page 154 mistakenly credits the Marsaglia-Bray method to Box and Muller.

WHAT SORT OF ENDING TO THIS SECTION WOULD MAKE SENSE?

Multivariate A multivariate normal p-dimensional random vector X has a mean vector $\boldsymbol{\mu}$ that is a $p \times 1$ vector and covariance matrix $\boldsymbol{\Sigma}$ that is $p \times p$. There are two general approaches to generate a multivariate normal ($\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$); however, the one based on a Cholesky decomposition of $\boldsymbol{\Sigma}$ is the most recommended method (Ripley, above book, Section 4.2, pp 98-99; Mark E. Johnson, Multivariate Statistical Simulation, section 4.1, pp. 49-54, John Wiley & Sons, Inc., New York, 1987).

As with the univariate normal in which any univariate normal can easily be generated from a standard normal, a general multivariate normal $\mathbf{X}$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ can be generated by the following formula where $\mathbf{Z}$ is a multivariate normal with a mean vector $\mathbf{0}$, and a covariance matrix

$\mathbf{I}$ which is the identity matrix with 1's on the diagonals and 0's on the off diagonals.

$\mathbf{X} = \mathbf{AZ} + \boldsymbol{\mu}$ where $\mathbf{AA}' = \boldsymbol{\Sigma}$. There are always solutions for $\mathbf{A}$, but the preferred way to use Cholesky decompostion to obtain a unique solution $\mathbf{L}$ for $\mathbf{A}$ that is lower triangular. The Cholesky factorization also is readily calculated with recursion formulas.

WHAT ELSE TO ADD TO THIS SECTION? SOMETHING ABOUT LATER APPLICATIONS?

1.5 Assessing the Distribution Fit

Given a set of data and a theoretical distribution, how well does the data match the proposed distribution? In our context, we are concerned with data produced by a random number generator matching the distribution it is supposed to be pulled from. However, these techniques also apply anytime a set of data has been gathered and one wonders if a particular theoretical distribution might provide an adequate representation of the data.

There exist many techniques to help with this. Some are visual comparisons, some are quantitative but without a measure to assess if the fit passes any statistical measures, and finally some are hypothesis tests. In this section we examine two hypothesis test procedures but caution the reader not to rely solely on statistical significance as the determining factor. One should be as thorough as possible when comparing data to a given hypothetical distribution and this includes visual checks (e.g., see Chapter 3 of the Handbook of Simulation, edited by Jerry Banks).

The two tests examined here are typically termed goodness-of-fit tests. The chi-square and the Kolmogorov-Smirnov tests are both testing the null hypothesis that a particular proposed theoretical distribution fits the sample data. If the observed value of a test statistics exceeds a critical value (or analogously the p-value of the test is less than a specified α level), then it is may be declared that the data significantly deviates from the hypothesized statistical distribution.

Caution is needed as blind acceptance of the results of the goodness-of-fit test is not a rational thing to do. Small sample sizes often do not provide enough statistical power for the test to be rejected; therefore, acceptance of a distribution based solely on not rejecting the goodness-of-fit null hypothesis is inadequate study. On the other extreme, large sample sizes often reject the null hypothesis when otherwise the data look to match a given distribution closely. Statistics should be carefully used as a tool to aid the decision making process, but statistics is not an exact science.

1.5.1 Chi-Square Goodness-of-Fit

In 1900 Karl Pearson published the chi-square goodness-of-fit test (*On a Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such That it can be Reasonably Supposed to Have Arisen in Random Sampling,* Philosophical Magazine (5), 50: pp157-175, 1990). It is the oldest of all such tests and is the one most frequently taught in statistics classes. It in essence is a comparison between an observed histogram and a fitted density (for continuous data) or probability (for discrete data) function, i.e., f(x) or a collapsed f(x). Discrete distributions have natural intervals in which to consolidate the observed discrete data values. Consolidation of continuous data into histogram cells is somewhat arbitrary. In either case, the observed data is collected in k cells. Of the n total data points, n_i fall into interval i, $i = 1, ..., k$ such that $\sum n_i = n$.

The next step is to compute the theoretical number of the n data values that should fall into each interval given the assumed theoretical distribution. This expected number of observations is the product of the sample size n and the theoretical proportion of the proposed distribution that should fit into the corresponding interval. The expected proportion p_i may be calculated by summing over the appropriate discrete values for a discrete distribution. If the distribution is continuous, calculus is used to integrate over the range for each cell. In the case of the continuous normal distribution the values of the integral are available in a table in many statistics texts.

Determining a reasonable number of cells and cell sizes to collect the data is an art. In general, most statistics text books suggest that each cell have an expected number of observations of at least 5 though some authors go as low as 3. Many individuals take a first cut at establishing the cells by either trying to make the width of most cells approximately the same or to make each cell have approximately the same expected number of observations. Keep in mind that the expected number of observations is a based on the assumed statistical distribution and is not tied to the actual data. One advantage of having equiprobable cells is that each cell receives the same weighting division factor np_i so that perhaps a more fair comparison of the observed number of observations in the cell, n_i, to the expected or theoretical number of observations in the cell, np_i, is made. Otherwise a similar squared deviation between expected and actual observations can be considerably impacted by widely varying np_i. Below is the general formulation which is the sum of the squared deviations between observed and expected frequencies divided by the corresponding expected frequency.

$$\chi^2_{observed} = \sum_{i=1}^{k} [(n_i - np_i)^2]/np_i$$

If all the parameters of the assumed theoretical distribution are known, the sum above is approximately a chi-square distribution with $k - 1$ degrees of freedom. For each distribution parameter that must be estimated from the data, there is one less degree of freedom in the chi-square. In general if p parameters are estimated from the data, then the chi-square has $k-p-1$ degrees

of freedom. If one wanted to perform the hypothesis test at the 95% confidence level, then the critical tabled value, $\chi^2_{critical}$, for the chi-square distribution would be selected from a chi-square table with $k-p-1$ degrees of freedom and $\alpha = 0.05$. If the observed chi-square, $\chi^2_{observed}$, based on the data is greater than the critical theoretical chi-square tabled value, $\chi^2_{critical}$, then the null hypothesis that the observed data is adequately fit by the given theoretical distribution may be rejected (keep earlier caveats in mind). With much of today's software, there is no longer a need to look up any critical values. Instead an observed level of significance called a p-value is computed. If the p-value is less than α, then the null hypothesis may be rejected.

Below is an example based on Chapter 5 of Practical Geostatistics 2000. In this example the null hypothesis is that a Poisson distribution fits a geologic data set of Zircon counts in Oswego Graywacke (Sichel 1971; Griffiths 1960). This data comes from counting the number of Zircon particles per unit area in "thin sections" of rock that are examined under a microscope. This is a common evaluation method in geology and mineralogy.

In the original published analysis, Griffiths fits a negative binomial distribution to the data and finds that it fails the chi-square goodness-of-fit test. Sichel (1971) compares the negative binomial and Sichel Compound Poisson distributions using chi-square goodness-of-fit tests. Sichel replicates Griffiths's finding for the negative binomial, but finds a good fit for the Sichel Compound Poisson. The analysis in Chapter 5 of Practical Geostatistics 2000 illustrates the use of the chi-square goodness-of-fit for Sichel's Compound Poisson as well as the standard Poisson found in most statistics books. Only the chi-square analysis for the standard Poisson distribution is shown below.

Table 5.1(b) of PG2000: χ^2 analysis for Zircon data

g	*Obs*	Standard Poisson Exp.	Standard Poisson Rev. Exp.	Rev. Obs	Standard Poisson χ^2
0	227	99.4836	99.4836	227	163.448
1	66	127.9470	127.9470	66	29.992
2	18	82.2770	82.2770	18	50.215
3	13	35.2725	35.2725	13	14.064
4	10	11.3411	11.3411	10	0.159
5	3	2.9172			
6	5	0.6253			
7	0	0.1149			
8	3	0.0185			
9	2	0.0026			
10	1	0.0003			
11	4	0.0000			
12	1	0.0000			
13	1	0.0000			
14	1	0.0000			
16	1	0.0000			
19	1	0.0000			
20	1	0.0000			
21	1	0.0000			
26	1	0.0000	3.6789	26	135.431
	360	360.0	360.0	360	χ^2 =393.309
				p =	7.76×10^{-84}
				df =	$6-1-1$
			Poisson λ =		1.286111

Table 2-3. Chi-Square Analysis of Poisson Distribution.

Note to ISE 521 students: if it makes you feel more comfortable, you may mentally think of x and $\bar{x}$ instead of the g and $\bar{g}$ below. In mining the random variable is often shown as g instead of x where g represents the grade of the ore sample.

Table 5.1(b) of *Practical Geostatistics 2000* (Clark and Harper, 2000, Ecosse North America, LLC) provides the calculations for an assumed Poisson distribution. The first column in this table represents the actual Zircon counts found in the samples while the second column gives the frequency of the counts listed in the first column. For example there were 227 counts with 0 Zircon particles. The only Poisson parameter is λ which equals $\bar{g} = 1.2861$. The third column calculates the expected number of observations assuming the Poisson fits the data and is the product of the total sample size of 360 times the associated Poisson

probability. This probability may be computed using the Poisson formula, using Excel, or possibly looking values up in a Poisson table if the desired λ value is given in the table. The fourth column collapses the expected frequencies for $g \geq 5$ into the value 3.6789 so that all cells used in the χ^2 analysis have expected frequencies of at least 3. The fifth column collapses the actual observations to match the expected. The final column computes the individual contributions to the χ^2 and results in an overall χ^2 value of 393.31 with a very small p-value of 7.76×10^{-84} indicating that the null hypothesis assumption that a Poisson adequately fits this data set may be rejected for very small values of α.

If one wanted to compare the computed chi-square value of 393.31 to a table critical value, the degrees of freedom must first be computed for the chi-square. There are 6 cells in the collapsed table. The parameter λ of the Poisson was estimated from the data. Thus the chi-square degrees of freedom = 6 - 1 -1 = 4. For four degrees of freedom, the critical value for an $\alpha = 0.05$ level is 9.488 as given on the second page of Table 3 in the statistical appendix. Even an α level of 0.001 corresponds to a critical value of 10.83 which is still much smaller than the observed chi-square value of 393.31 indicating the p-value is < 0.001 which agrees with the prior paragraph.

In performing the chi-square goodness-of-fit test, both the expected and observed counts were collapsed so that the expected frequency for the collapsed or consolidated cells is at least 3. As an example, the row corresponding to $g = 26$ under the standard Poisson has a revised expected value of 3.6789 and a revised number of observations of 26. The value 3.6789 is the sum of the expected values for $g \geq 5$. Note that the sum of this column is 360 which is equal to the number of actual observations - this is the type of check that is important to make sure that one's logic has been properly applied when collapsing cells. Similarly the value of 26 in the column for the revised number of observations represents the sum of the actual observations for $g \geq 5$.

1.5.2 Kolmogorov-Smirnov Goodness-of-Fit Test

The Kolmogorov-Smirnov (K-S) test is often a more powerful statistical test than the chi-square test. Rather than a comparison of an observed histogram to the corresponding theoretical f(x), the K-S test compares the theoretical distribution function, F(x), to its empirical equivalent. The K-S does not require grouping of the observed data as does the chi-square. This is especially an issue when working with continuous distributions that have no natural grouping of the observations. The major drawback to the K-S is a lack of exact critical values for many distributions thus making the tests approximate.

For our purposes here the empirical distribution function $F_n(x)$ is defined below where $x_{(i)}$are the ordered sample values and are sometime referred to as the order statistics. The $x_{(i)}$ are ordered from smallest to largest, i.e., $x_{(i)} \leq x_{(i+1)}$.

$F_n(x) = \frac{number\ of\ x_{(i)} \leq x}{n}$; therefore, $F_n(x_{(i)}) = \frac{i}{n}$.

The K-S test compares the empirical distribution function $F_n(x)$ based on n observations to the hypothesized theoretical distribution function $\hat{F}(x)$. $\hat{F}(x)$ may assume the distributional parameters are known or may fit the parameters needed from the available data. In this latter case, the K-S test is an approximate test as the necessary critical values are not known exactly. The formal definition of the test statistic D_n used for the K-S test is given below; however, in applications the subsequent definitions are used to apply the test. Then the observed D_n is compared to the critical values given in Table **xxx** (Massey, Frank J. Jr., "The Kolmogorov-Smirnov Test for Goodness of Fit", Journal of the American Statistical Association, Vol 46, Issue 253, March 1951, pp 68-78). If D_n is larger than the appropriate critical value then the null hypothesis that the proposed theoretical distribution adequately fits the available data may be rejected at the α level used.

$D_n = \sup_x[|F_n(x) - \hat{F}(x)|]$

In the above definition, sup is the supremum which in practice for us is the maximum. Thus we are looking for the largest (maximum) absolute vertical difference between $F_n(x)$ and $\hat{F}(x)$. Defining D_n^+ and D_n^- as given below, the D_n used in practice follows below. D_n^+ and D_n^- are vertical deviations between $F_n(x)$ and $\hat{F}(x)$ with D_n^+ representing the maximum deviation where $\frac{i}{n}$ is larger than $\hat{F}(x)$. Likewise D_n^- is the maximum deviation between the empirical and theoretical values when the theoretical distribution function $\hat{F}(x)$ is higher than $\frac{i-1}{n}$. Finally D_n is the max of D_n^+ and D_n^-.

$D_n^+ = \max_{1 \leq i \leq n}\{\frac{i}{n} - \hat{F}(x_{(i)})\}$, $\quad D_n^- = \max_{1 \leq i \leq n}\{\hat{F}(x_{(i)}) - \frac{i-1}{n}\}$, $D_n = \max\{D_n^+, D_n^-\}$.

Figure 2-6 gives a general graphical picture of how the K-S test is performed where the solid curve represents the assumed theoretical cumulative distribution function and the horizontal lines represent step function jumps in the empirical cumulative distribution function for a sample of size $n = 4$. Specific graphs for data analyzed with the K-S test are presented later illustrating visually D_n^+ and D_n^-.

Below are small sample applications of the Kolmogorov-Smirnov test. The first is for an exponential distribution with a mean of 0.185 or in terms of the exponential formulation given in the Inverse Transform section $\lambda = 1/0.185$. The first column provides the index "i" for the sorted data while the second column lists the eight data values in ascending order. The third column is the cumulative distribution function for the assumed theoretical model. The final four columns follow from the needed calculations for D_n^+, D_n^- as defined above. Keep in mind that D_n^+, D_n^- are both maximums and not the largest absolute value. Thus $D_n^+ = 0.115$ and not |-0.161| = 0.161. $D_n^- = 0.286$, and thus $D_n = \max\{0.115, 0.286\} = 0.286$.

From Table **xx** of critical values for the K-S test, the $\alpha = 0.05$ critical value for n = 8 is 0.457 which is not exceeded by the observed K-S $D_n = 0.286$;

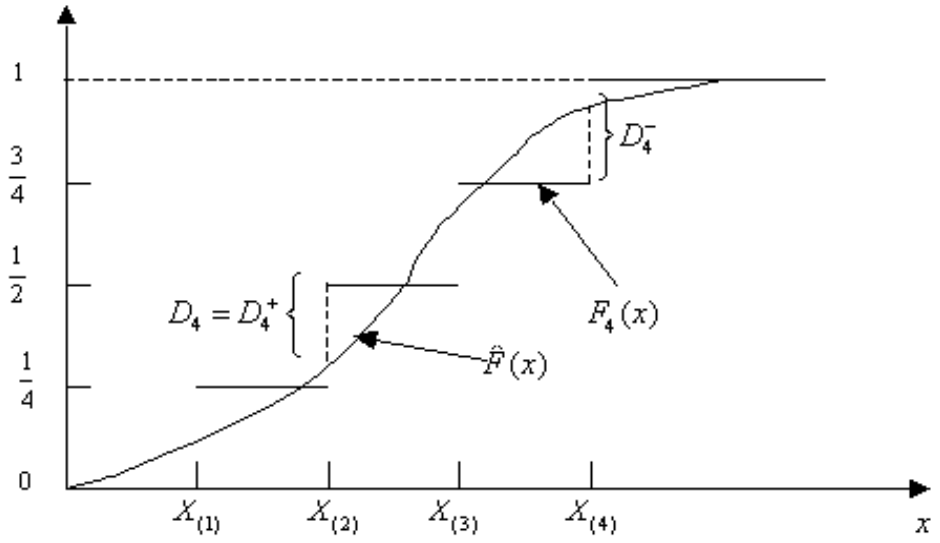

Figure 6: Figure 2-6. General K-S Test Graph.

therefore, there is no solid evidence that an exponential distribution should be rejected. Only 8 data values is not a large sample; however the K-S test is able to pick up distributional fits better than a chi-square test for small sample sizes. Generally the most difficult part of performing the K-S test is computing the column titled $Fhat(x_{(i)})$ below which has been referred to as $\hat{F}(x_{(i)})$ earlier in the mathematical formulation. From the earlier section on the Inverse Transform technique, $F(x) = 1 - e^{-\lambda x}, x \geq 0$ for an exponential distribution. For our assumed theoretical exponential model with $\lambda = 1/\,0.185$, $\hat{F}(x_{(i)}) = 1 - \exp(-\frac{x_{(i)}}{0.185})$ where exp is the natural base e. This can easily be done in a spreadsheet, but it is worthwhile to perform a few of the calculations by hand to make sure the process is understood and that the spreadsheet has been properly setup. As one example, take the lowest data value of 0.01 and compute $\hat{F}(x_{(i)}) = 1 - \exp(-\frac{0.01}{0.185}) = 0.053$ as seen in the $Fhat(x_{(i)})$ column corresponding to $x_{(i)} = 0.01$ in Figure 2-7.

Figure 2-8 graphically depicts the key aspects of the K-S test for the eight data values. Since two of the data values are the same (i.e., two are 0.20), it is not as easy to see as if all the values were different. The diamond shaped points represent $\frac{i}{n}$, the solid curve is the assumed exponential distribution, and the circular points are $\frac{i-1}{n}$. The largest positive deviation of $\frac{i}{n}$ minus the theoretical curve is 0.115 and thus $D_n^+ = 0.115$. $D_n^- = 0.286$ as this is the maximum deviation between the theoretical curve and $\frac{i-1}{n}$. Then $D_n = \max\{D_n^+, D_n^-\}$

i	$x_{(i)}$	Fhat($x_{(i)}$)	i/n	i/n-Fhat($x_{(i)}$)	(i-1)/n	Fhat($x_{(i)}$)-(i-1)/n
1	0.01	0.053	0.125	0.072	0.000	0.053
2	0.05	0.237	0.250	0.013	0.125	0.112
3	0.07	0.315	0.375	0.060	0.250	0.065
4	0.20	0.661	0.500	-0.161	0.375	**0.286**
5	0.20	0.661	0.625	-0.036	0.500	0.161
6	0.25	0.741	0.750	0.009	0.625	0.116
7	0.30	0.802	0.875	0.073	0.750	0.052
8	0.40	0.885	1.000	**0.115**	0.875	0.010

Figure 7: Figure 2-7. K-S example for the exponential distribution.

= 0.286.

The example in Figure 2-9 assumes that a normal distribution with $\mu = 25$, and $\sigma = 5$ fits the five sorted data values seen in the second column below. The columns are the same as given in the exponential example above. Computing $Fhat(x_{(i)})$ below [or $\hat{F}(x_{(i)})$] may be done easily using a normal table and keeping in mind the relationship between a standard normal with $\mu = 0$, and $\sigma = 1$ and any other normal, i.e., $z = \frac{x-\mu}{\sigma}$, where z is the standard normal and x is any normal. The Excel function NORMDIST($x_{(i)}$,25,5) can be used to compute $Fhat(x_{(i)})$.

As can be pulled from Figure 2-9 $D_n^+ = 0.441$ and $D_n^- = 0.164$. From this $D_n = \max\{D_n^+, D_n^-\} = 0.441$. At the $\alpha = 0.05$ level, the K-S critical value is 0.565 which is larger than the observed K-S test statistic D_n. Thus there is no strong evidence to reject a normal distribution with $\mu = 25$, and $\sigma = 5$ fitting the small data set. Try doing this by hand, but be careful how the normal table you use is setup. For example, take $x_{(i)} = 26$ that corresponds to a standard normal $z = \frac{x-25}{5} = \frac{26-25}{5} = 0.2$. Table 2 in the statistical appendix shows that $F(.2) = 0.5793$ or 0.579 to 3 decimal places as shown in Figure 2-9. Figure 2-10 graphically shows the K-S test for a normal distribution with $\mu = 25$, and $\sigma = 5$ applied to this data set.

The Kolmogorov-Smirnov is a powerful goodness-of-fit test. We suggest (if sufficient data exists) using both the K-S and chi-square goodness-of-fit tests in addition to visual checks (e.g., probability plots). One of the nicest aspects of the K-S test is its ability to work with small samples. It is also nice not to have to create bins or cells for the data as is required for the chi-square goodness-of-fit test.

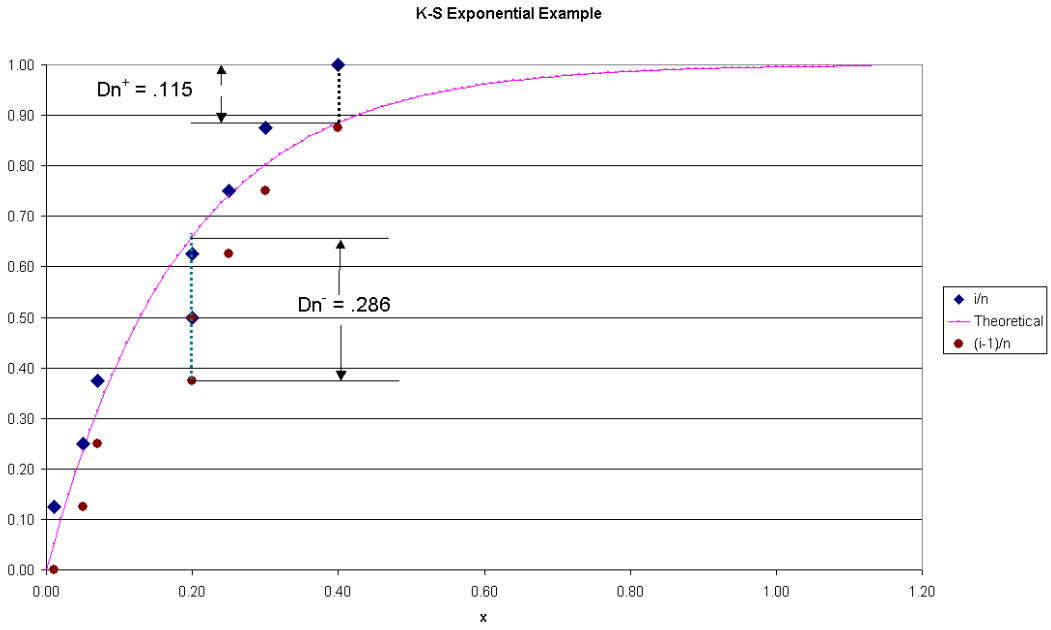

Figure 8: Figure 2-8. Graphical View of K-S test appliced to Exponential distribution.

i	$x_{(i)}$	i/n	$Fhat(x_{(i)})$	i/n-$Fhat(x_{(i)})$	(i-1)/n	$Fhat(x_{(i)})$-(i-1)/n
1	17.5	0.2	0.067	0.133	0	0.067
2	19	0.4	0.115	0.285	0.2	-0.085
3	20	0.6	0.159	0.441	0.4	-0.241
4	26	0.8	0.579	0.221	0.6	-0.021
5	34	1	0.964	0.036	0.8	0.164

Figure 9: Figure 2-9. K-S Example for the Normal distribution.

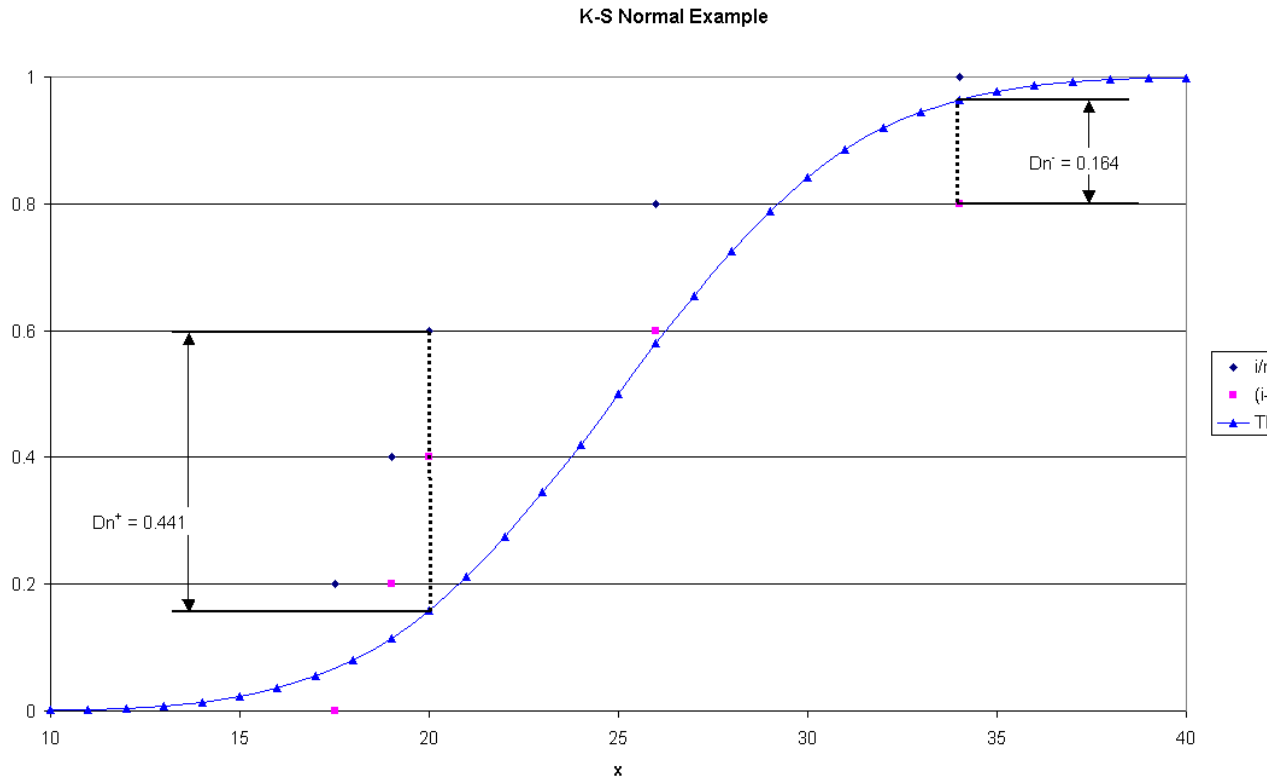

Figure 10: Figure 2-10. K-S Normal Example

2 Summary

This chapter introduced many aspects and concepts of simulation. One can easily get lost in the mathematics. As fun as the math can be, it is important to step back and question if the model is sufficient for the type of answers sought. Keep in mind that the model is an approximation of the real world. Besides introducing various simulation jargon, this chapter has endeavored to peel away the mystery sometimes associated with "random" numbers. The linear congruential method was used to create pseudo-random Uniform(0, 1) numbers. The Inverse Transform Technique illustrated one avenue for the generation of other statistical distributions from the Uniform(0, 1) numbers. For the normal distribution, some additional procedures were illustrated. Finally the issue of goodness-of-fit reared its head. As with all endeavors in life, believe in yourself and your ability to see if things pass a TLAR (That Looks About Right) test in addition to using statistical tools as an aid to decision making. Ultimately the decisions are yours and you must feel comfortable with them.

Need other methods of Normal generation, MVN.

Volume II: normal or Normal? Change x to g; use z instead of x. I wish we stayed with classical statistical notation in our geostat books.