

Review of Probability and Statistics

Introduction

Many of the simulation techniques presented in this course require some knowledge of probability and statistics. The reason for this paper is not to present the material in-depth but to introduce or review the concepts of probability theory and elements of statistics to students of construction engineering and project management.

In probability, properties of the population are assumed to be known, and can be modeled mathematically. Probability can be used to answer questions about the sample. Statistics is used to determine properties from a sample that can be applied to the general population. These relationships are in Figure 3-1.

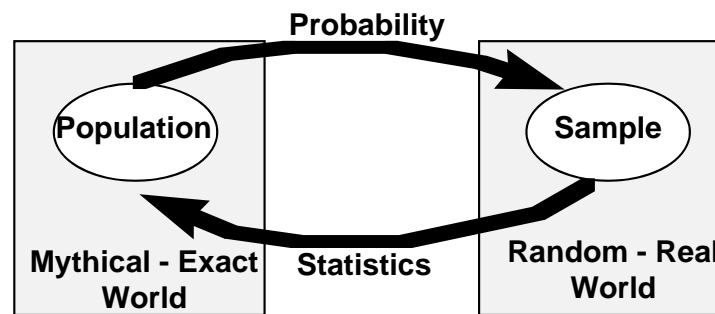


Figure 3-1. The relationship between probability and statistics

Descriptive Statistics

Frequently, we wish to infer information about a population. For example, concrete cylinders from a pour, average time to failure for a computer system, biological oxygen demand tests from an aquifer, etc. However, the information at our disposal consists of a portion or subset of the population. The part of statistics that deals with this process is called descriptive statistics. Descriptive statistics can be used when we have information about all the population, or when the data consists of a subset. In this paper we will use the following terms:

- μ = population mean
- \bar{x} = sample mean
- σ^2 = population variance
- s^2 = sample variance

Descriptive statistics are important for most engineers for a variety of reasons. The most of important of these is being all decisions in the planning process should be based upon data. Whether that data truly reflects public opinion, environmental consequences, program duration, etc., data is the most important input for the systems design process.

Measures of Location

Some features of a data set or representative sample are of interest to convey important information about that data set. We write this section to present methods for describing the location, in particular the center, of a data set. Section 3.2 addresses measures of the variability of the data set.

Suppose a sample space consists of a population of N elements that has a quantitative attribute (shear strength, cost, weight, yield stress, etc.). Usually it is impossible to observe or even know the number of N elements that comprise the sample space. If a

Chapter 3

sample size n of the population N is taken, then the sample mean or arithmetic average of the set is given by

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n} \quad (3-1)$$

The sample mean \bar{x} is the best estimate of the population mean or μ when a subset of the population is available. For small data sets, the sample mean can produce unsatisfactory results because its value can be influenced by one observation that lies significantly outside the range of the remaining data.

Example 3-1. Air temperature measurements were made at seven sites during November. Determine the sample mean for air temperature based upon those test results.

Sample	T (degrees)
1	38.7
2	42.5
3	34.7
4	42.3
5	37.6
6	43
7	37.3

$$\bar{x} = \frac{38.7 + 42.5 + 34.7 + 42.3 + 37.6 + 43 + 37.3}{7} = 39.4 \text{ degrees}$$

Measures of Variability

In addition to information about the location, you need some measure of the variability to give a meaningful description of the data set. The simplest measure of variability is the sample range. Unfortunately, the measure of variability reflects the two extreme observations and disregards those in the middle of the data set. A more effective measure might be the deviation from the sample mean. To obtain a meaningful measure of variability we need to use absolute deviations. Unfortunately, absolute deviations have many inherent theoretical difficulties. Thus, consider the average sum of the squared deviations that is often referred to as the sample variance, and is given by:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} \quad (3-2)$$

The sample standard deviation is the square root of the sample variance and is denoted by s . Variance and standard deviation are the most common descriptors used to describe sample variability.

Example 3-2. Using the data presented in Example 3-1, determine the sample variance and standard deviation for air temperature.

From Example 1, $\bar{x} = 39.4$ degrees

x	$(x - \bar{x})^2$
38.7	0.49
42.5	9.61
34.7	22.09
42.3	8.41
37.6	3.24
43.0	12.96
37.3	4.41

	61.21

Thus,

$$s^2 = \frac{61.21}{6} = 10.20$$

$$s = 3.19 \text{ degrees}$$

Regression

Regression analysis is used to answer questions about a sample and make predictions based upon that sample. By understanding how some phenomena depend on others we learn to predict the consequences of our actions. In simple terms, any method of fitting equations to data may be called regression. We will use the technique of least squares to fit data in this section. The least squares method of regression is the most widely used of any statistical data analysis technique.

The simplest deterministic relationship between two variables x and y is a linear relationship of the form

$$y = \beta_0 + \beta_1 x \quad (3-3)$$

where

β_0 = y intercept on a x-y plot

β_1 = slope of a straight line on a x-y plot

Note that the error term in predicting a value of the dependent variable at any point x_i , is given by $e = y_i - (\beta_0 + \beta_1 x_i)$. The principal of the least squares curve fitting technique is to minimize the square of the error with respect to the curve parameters, or

$$\text{minimize} \quad f(\beta_0, \beta_1) = \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]^2$$

Find the minimize values of the function by taking the partial derivatives of $f(\beta_0, \beta_1)$ with respect to β_0 and β_1 and solving the equations.

$$\frac{\partial f(\beta_0, \beta_1)}{\partial \beta_0} = \sum_{i=1}^n 2(y_i - \beta_0 - \beta_1 x_i)(-1) = 0$$

$$\frac{\partial f(\beta_0, \beta_1)}{\partial \beta_1} = \sum_{i=1}^n 2(y_i - \beta_0 - \beta_1 x_i)(-x_i) = 0$$

Canceling the 2's and rearranging produces what are called the normal equations

$$n\beta_0 + \left(\sum_{i=1}^n x_i\right)\beta_1 = \sum_{i=1}^n y_i$$

$$\left(\sum_{i=1}^n x_i\right)\beta_0 + \sum_{i=1}^n x_i^2 \beta_1 = \sum_{i=1}^n x_i y_i$$

These equations are linear in terms of two unknowns, β_0 and β_1 . Thus, the least squares estimates of the coefficients β_0 and β_1 of the true regression line are:

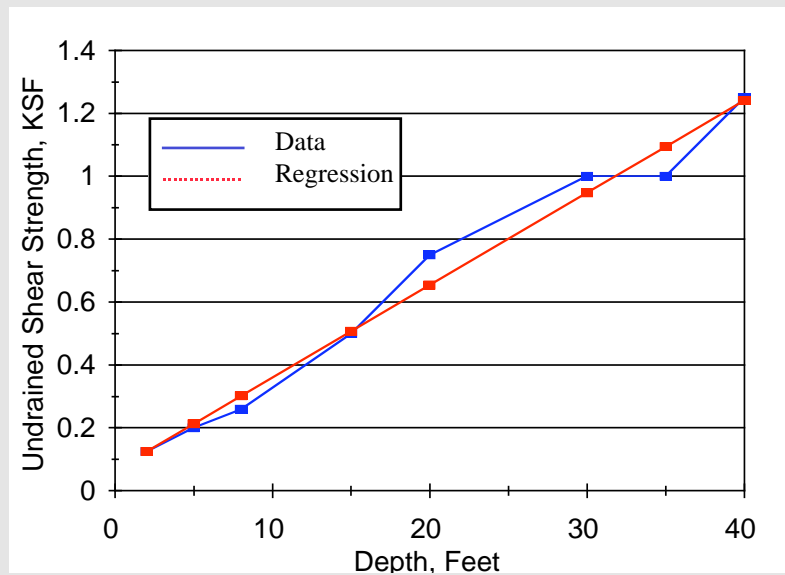
$$\beta_1 = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \quad (3-4)$$

$$\beta_0 = \frac{\sum_{i=1}^n y_i - \beta_1 \sum_{i=1}^n x_i}{n} \quad (3-5)$$

Example 3-3. A field investigation was conducted in the Gulf of Mexico to develop a detailed design to fasten a drilling platform. The table shows soil shear strengths. Perform a linear regression suitable for predicting shear strength to a depth of 40 feet.

Penetration Below Seafloor (ft)	Shear Strength (ksf)
2	0.125
5	0.20
8	0.26
15	0.50
20	0.75
30	1.0
35	1.0
40	1.25

Solution All spreadsheet programs contain some type of regression analysis capability. A spreadsheet software package was used to produce this output for the problem



$$\beta_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2} = \frac{8(140.83) - (155 * 5.085)}{8(4443) - (155)^2} = 0.0294$$

$$\beta_0 = \beta_0 = \frac{\sum y_i - \beta_1 \sum x_i}{n} = \frac{5.085 - .0294 * 155}{8} = 0.0663$$

Example 3-3. continued Thus, this linear equation could be used to describe undrained shear strength as a function of depth

$$y = \beta_0 + \beta_1 x = 0.0663 + 0.0294x$$

For a depth of 20 feet the model will predict a shear strength of 0.656 ksf.

Many relationships cannot be simply defined in terms of a single independent variable. For example, sales are often a function of price, advertising, competition, etc. If a relationship between those independent variables can be defined, better forecasts can be developed. Multiple regression, neural networks, and simulation are often used when more complicated techniques are required.

The general form of the multiple linear regression equation is

$$y = a_0 + \sum_{i=1}^n a_i x_i + b \quad (3-6)$$

where a_0 = value of y at $x_i = 0$ for all i

a_i = coefficient of the independent variable x_i

b = constant

x_i = independent variable

y = forecast variable

n = number of independent variables

Relationships may also be non-linear. Using the model

$$y = \sum_{i=0}^k a_i x^i \quad (3-7)$$

We may fit data with a polynomial of degree k . Coefficients of both multiple linear regression equation may be found using the least squares technique.

Correlation

There are many instances when analyzing the relationship between two variables is needed. The coefficient of correlation or r is used to measure how strongly two variables, x and y pairs, are correlated. The square of the sample correlation coefficient is referred to as the coefficient of determination. Both of these descriptive measures are used to describe how well we can predict the variation in y using a linear regression model utilizing the independent variable x . The coefficient of determination or r^2 can be calculated using:

$$r^2 = \left(\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \right)^2 \quad (3-8)$$

with $r = \pm \sqrt{r^2}$

The coefficient of determination has a range between -1 and 1 where a value of -1 implies perfect negative correlation and 1 a perfect positive correlation. The r^2 value is most often used when comparing how well a linear regression model predicts y . The smaller the value of r^2 the poorer the relationship.

Example 3-4. Determine r^2 for the data in Example 3-3. Based upon that value, are these data highly correlated?

Depth	Shear	$c=x-x_{\text{bar}}$	$d=y-y_{\text{bar}}$	$c*d$	$(x-x_{\text{bar}})^2$	$(y-y_{\text{bar}})^2$
2	0.125	-17.375	-0.510625	8.87210938	301.890625	0.26073789
5	0.2	-14.375	-0.435625	6.26210938	206.640625	0.18976914
8	0.26	-11.375	-0.375625	4.27273438	129.390625	0.14109414
15	0.5	-4.375	-0.135625	0.59335938	19.140625	0.01839414
20	0.75	0.625	0.114375	0.07148438	0.390625	0.01308164
30	1	10.625	0.364375	3.87148438	112.890625	0.13276914
35	1	15.625	0.364375	5.69335938	244.140625	0.13276914
40	1.25	20.625	0.614375	12.6714844	425.390625	0.37745664
Sum				42.308125	1439.875	1.26607188

x_{bar} 19.375

y_{bar} 0.635625

r 0.99090542

r^2 0.98189355

Probability

The term probability refers to the study of randomness and uncertainty. There are two ways that the term probability is usually defined. One is a classical definition; the other is a mathematical definition. We shall develop the definitions and present the concepts. These are definitions related to probability:

- **Experiment:** Any action or process that generates observations. In civil engineering, the taking of concrete cylinders during a concrete pour, breaking them in compression, and recording the stress at which they fail, may be considered an experiment. Measuring the time it takes to drive a number of steel piles, recording the times, and developing a production rate is another. One characteristic of an experiment is that it terminates with an outcome. If the experiment can be repeated under the same conditions, it is called a random experiment.
- **Sample Space of an Experiment:** The collection of all possible outcomes of a random experiment is called the sample space. A sample space may consist of a finite or infinite number of outcomes and is usually denoted by S . The probability of the entire sample space is 1. Thus $P(S)=1$.
- **Sample Point:** A single outcome in the sample space. The collection of all sample points comprises the sample space.
- **Event:** A collection (subset) of outcomes contained in the sample space S . Simple events consist of exactly one. Compound events consist of more than one outcome.
- **Random Variable:** Any rule that associates a number with each outcome in S for any experiment.

Example 3-5. Consider one throw of two dice, one red and the other white. There are 6^2 or 36 possible outcomes for this experiment, each equally likely. These 36 possible outcomes define the finite sample space for this two dice experiment.

Example 3-6. At every bid opening, a contractor records each competitor's bid and divides it by his own direct cost estimate. The results are then recorded. The number of all possible outcomes is countably infinite, each with a possible very different answer. This example demonstrates a random experiment with an infinite sample space.

Let us define an event A as an outcome of an experiment. Let n_a be the number of outcomes of event A in the sample space (frequency of occurrence). If n is the total of all possible outcomes in the sample space, then the classical definition of probability of the defined event occurring is:

$$P(A) = \frac{n_a}{n} \quad (3-9)$$

Example 3-7. What is the probability that the sum of two dice in one throw is equal to 7? If x is the outcome, then

$$P(X = 7) = \frac{6}{36} = \frac{1}{6}$$

Count the outcomes of the event $x = 7$ as (6,1), (5,2), (4,3), (3,4), (2,5), (1,6) for a total of 6 outcomes out of a possible 36 outcomes. Note that $P(X=7)$ is read as "the probability the random variable X assumes the value 7."

Consider the event that the ratio of a competitor's bid to a contractor's estimate, r , lies between 1.02 and 1.04 (e.g., $1.02 \leq r \leq 1.04$). The possible outcomes of r and the total number of possible outcomes of the experiment are impossible to count. This leads to the mathematical definition of a probability

$$P(A) = \lim_{n \rightarrow \infty} \frac{n_a}{n} \quad (3-10)$$

Therefore, in the bidding example, the probability of the ratio of a bid to a contractor's estimate falling between 1.02 and 1.04 is the limit of the number of times that the ratio could fall in that interval divided by the total possible outcomes as that number approaches infinity. This interpretation of probability depends upon the fact that an experiment can be repeated under essentially the same conditions. But, many persons extend probability to situations by treating it as a rational measure of belief.

Consider the Venn diagram in Figure 3-2.

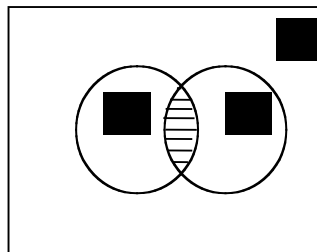


Figure 3-2. Venn diagram with two events

Chapter 3

Let the two circles represent events A and B and the rectangle C represent the sample space. These are rules or definitions:

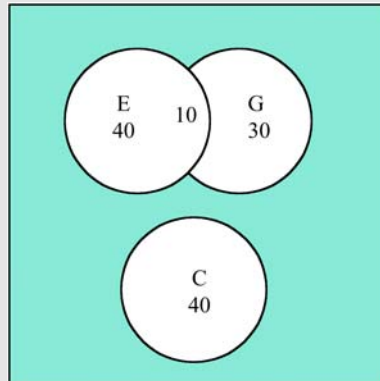
- **Union or \cup :** The union of two events A and B is the event consisting of all outcomes that are either in A or in B or in both events.
- **Intersection or \cap :** The intersection of two events A and B is the event consisting of all outcome that are both in A and B.
- **Complement:** The complement of an event A is the set of all outcomes in the sample space that are not contained in A.
- **Mutually Exclusive (or Disjoint):** When events A and B have no outcomes in common, they are said to be mutually exclusive or $A \cap B = \emptyset$
- **Additive Law:** The probability of either event A or B occurring is written, as $P(A \text{ or } B) = P(A) + P(B) - P(A \cap B)$ and $P(C) = 1$.
- **Conditional Probability:** The intersection of two events is derived from the definition of conditional probability

$$P(A/B) = \frac{P(A \cap B)}{P(B)} \quad (3-11)$$

Example 3-8. The Department of Systems Engineering at Stevens has a total of 120 undergraduate students and offers 3 majors with enrollments shown in the table.

<u>Major</u>	<u>Number of Students</u>
Systems	40
Engineering Management	30
Systems/Engineering Management Dual Major	10
Info Systems Engineering	40
Total	120

- 1) Draw the Venn Diagram that describes all majors and their populations



- 2) What is the probability that a student selected at random is an systems major? Engineering Management major? dual systems/Engineering Management major?

$$P(E) = \frac{50}{120} = 0.42; \quad P(G) = \frac{40}{120} = .33; \quad P(E \cap G) = \frac{10}{120} = .08$$

- 3) What is the probability that a student selected at random is either an systems or a Engineering Management major (use the additive law)?

$$P(E \cup G) = P(E) + P(G) - P(E \cap G) = .42 + .33 - .08 = .67 \text{ (i.e., } \frac{80}{120} \text{)}$$

- 4) What is the probability that a student selected at random is either an Systems or a Engineering Management major, but is not a dual major?

$$P(E \cup G) - P(E \cap G) = .67 - .08 = 0.58$$

- 5) Which majors are mutually exclusive?

Info Systems Engineering and systems, Info Systems Engineering and Engineering Management

Random Variables

A random variable is a special type of function critically important to the science of mathematical statistics. The concept of a random variable allows us to relate the exponential outcomes to a numerical function describing the outcome. It is a function that

- has as its domain the sample space,

Chapter 3

- assigns one and only one number to each point in the sample space (integer, 0 or 1, or real), and
- its value can be found only through an experiment.

The value assumed by a random variable associated with an experiment depends on the outcome of the experiment. As a convention, this text will use a capital letter (usually X or Y) to represent a random variable and its lower case to represent a specific numerical value of the random variable.

Example 3-9. You are trying to determine the probability of a contractor filing more than 2 claims for a given contract. You analyze the last five jobs, and the number of claims are:

Job	1	2	3	4	5
Claims	0	4	2	5	2

Let X be the number of claims. There are four possible outcomes: 0, 2, 4, and 5. Then

$$p(0) = P(X=0) = \frac{1}{5} = 0.2$$

$$p(2) = P(X=2) = \frac{2}{5} = 0.4$$

$$p(4) = P(X=4) = \frac{1}{5} = 0.2$$

$$p(5) = P(X=5) = \frac{1}{5} = 0.2$$

These values specify the probability distribution function. In simple terms, for every possible value of x, the probability distribution specifies the probability of observing that value when the experiment is performed. Thus,

$$P(X \geq 2) = 1 - [P(X=0) + P(X=1) + P(X=2)] = 1 - [0.2 + 0 + 0.4] = 0.4$$

Probability Density Functions

The *probability distribution (or density) function* (pdf) provides a complete description of a random variable. A pdf is an important part of civil engineering planning when trying to model complex processes (i.e., construction). Typically, the expected value, and some measure of scatter are used to summarize the important characteristics of a pdf.

Let X be a random variable such that $P(X = x) = f(x)$ is known for each x in the sample space. The term $f(x)$ is the pdf of X. It describes the probability for X over the entire sample space. There are two types of random variables, the discrete and the continuous, depending upon whether the results are countable or are infinite.

1) The Discrete Type of Random Variable. Let X be a special type of random variable defined on the set of real numbers in a way that for any finite interval there exists a finite number of values. Let $f(x)$ be a function such that

$$\sum_{i=1}^n f(x_i) = 1 \quad (3-12)$$

2) The Continuous Type of Random Variable. A random variable is said to be continuous when the value constitutes an infinite set between two numbers, say a and b. Let X be a random variable such that

$$P(a \leq X \leq b) = \int_a^b f(x) dx \quad (3-13)$$

and

- (1) $f(x) \geq 0$ for all x ,
- (2) $a < b$
- (3) $f(x)$ has at most, a finite number of discontinuities on every finite interval, and
- (4) $\int_a^b f(x) dx = 1$

then X is said to be a continuous random variable.

Example 3-10. As part of a major land development project, 125 construction sites are assigned a 1-5 rating, with 5 being the most desirable. The rating was achieved based upon view, trees, constructability, access, and lot size. Based upon a study of the area we obtained the values in the this table:

Lot Rating, x	1	2	3	4	5
Number	10	35	20	27	33
Probability, p(x)	0.08	0.28	0.16	0.216	0.264

The numbers 1 through 5 in these tables are values of the random variable. A pdf of discrete random variable is defined for every number x by $p(x) = P(X=x)$ or

$$\begin{aligned} p(1) &= P(X=1) = 0.08 \\ p(2) &= P(X=2) = 0.28 \\ p(3) &= P(X=3) = 0.16 \\ p(4) &= P(X=4) = 0.216 \\ p(5) &= P(X=5) = 0.264 \end{aligned}$$

and equivalent description and the most widely used for a pdf is

$$f(x) = \begin{cases} 0.08 & \text{if } x = 1, \\ 0.28 & x = 2, \\ 0.16 & x = 3, \\ 0.216 & x = 4, \\ 0.264 & x = 5, \\ 0 & \text{otherwise} \end{cases}$$

Figure 3-3 is a graph of the relative frequency histograms.

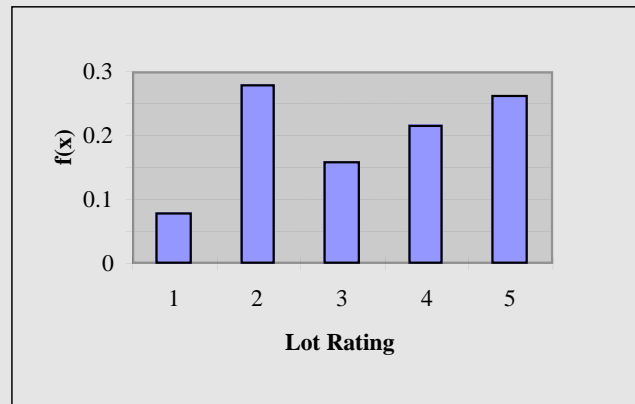


Figure 3-3. Probability histogram

Example 3-11. Let the discrete random variable X be the number of dump trucks that arrive at a rock crusher in a 60 minute period, and the pdf can be expressed as a Poisson distribution

$$f(x) = \frac{12^x e^{-12}}{x!}$$

The arrival rate of the truck has a great impact on the production rate, thus the cost and duration of a project. Say the project was bid based on the estimate that 10 trucks per hour would be the arrival rate. What is the probability that fewer than 10 trucks will arrive per hour?

$$\begin{aligned} P(X < 10) &= \sum_{x=0}^9 \frac{12^x e^{-12}}{x!} \\ &= 0.24 \end{aligned}$$

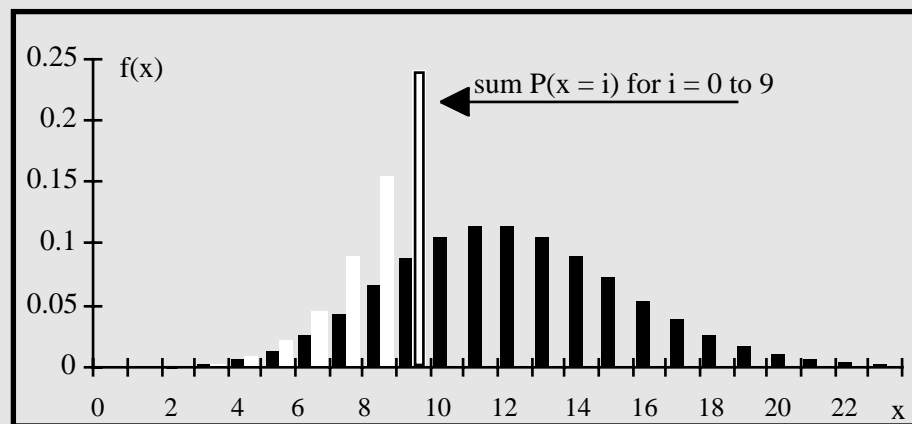


Figure3-4. Poisson pdf with a mean of 12

Example 3-12. Let the random variable X represent the strength of A36 steel. Let

$$f(x) = \frac{1}{\sqrt{2\pi(5)}} e^{-\frac{1}{2}\left(\frac{x-41}{5}\right)^2}$$

The probability that the strength of the steel is less than 36 ksi is

$$\begin{aligned} P(X \leq 36) &= \int_{-\infty}^{36} \frac{1}{\sqrt{2\pi(5)}} e^{-\frac{1}{2}\left(\frac{x-41}{5}\right)^2} dx \\ &= 0.16 \end{aligned}$$

Whether the random variable is continuous or discrete, the function $f(x)$ completely defines its probability properties. The function, $f(x)$, is called the *probability distribution function of x* .

Example 3-13. Let the random variable X have the pdf

$$f(x) = \begin{cases} 2x & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Find $P(1/2 \leq X \leq 3/4)$ and $P(-1/2 \leq X \leq 1/2)$

$$P(1/2 \leq X \leq 3/4) = \int_{1/2}^{3/4} 2x dx = \left[x^2 \right]_{1/2}^{3/4} = 5/16$$

$$P(-1/2 \leq X \leq 1/2) = \int_{-1/2}^0 0 dx + \int_0^{1/2} 2x dx = 1/4$$

Cumulative Distribution Functions

For some fixed value of x we often wish to compute the probability that the observed value of X will be at most x . Consider the random variable X . Take x to be a real number and consider the unbounded set $-\infty$ to x , including the point x itself. Let $F(x) = P(X \leq x)$. The function, $F(x)$, is called the *cumulative distribution function* (cdf) of the random variable X . Since $F(x) = P(X \leq x)$, then with $f(x)$ the pdf,

$$F(x) = \sum_{n=-\infty}^x f(n) \quad (3-14)$$

for the discrete type of random variable.

For the continuous type random variable, the cdf can be expressed as

$$F(x) = \int_{-\infty}^x f(u) du \quad (3-15)$$

In Figure 5, for each x , $F(x)$ is the area under the pdf curve to the left of x .

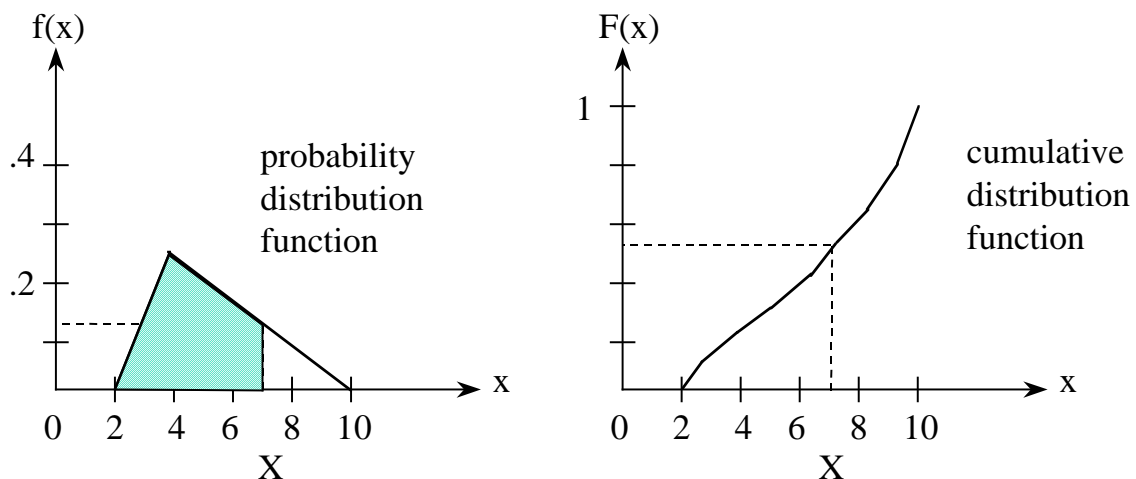


Figure 3-5. A pdf and associated cdf

Chapter 3

Note here that $F(x)$ is a function of the point, x , and

$$\frac{dF(x)}{dx} = f(x) \quad (3-16)$$

The probability $p(a \leq X \leq b) = F(b) - F(a)$ except in rare functions involving discontinuities.

Example 3-14. The pdf for the number of contractors that will bid on a minor construction jobs in your area is

x	1	2	3	4
$p(x)$.4	.3	.2	.1

Thus $P(X \leq 1) = 0.4 = p(1)$

$$P(X \leq 2) = 0.7 = p(1) + p(2)$$

$$P(X \leq 3) = 0.9 = p(1) + p(2) + p(3)$$

$$P(X \leq 4) = 1.0 = p(1) + p(2) + p(3) + p(4)$$

The cdf is

$$F(x) = \begin{cases} 0 & \text{otherwise} \\ .4 & x < 1 \\ .7 & x < 2 \\ .9 & x < 3 \\ 1.0 & x \leq 4 \end{cases}$$

Example 3-15. Let x possess the pdf

$$f(x) = \begin{cases} cx & 0 \leq x \leq 2 \\ 0 & \text{elsewhere} \end{cases}$$

1) Find c

For this to be a continuous function, $\int_a^b f(x)dx = 1$. Thus,

$$\int_0^2 cx dx = 1 = c \left[\frac{x^2}{2} \right]_0^2 = c[4 - 0] = 4$$

Thus, $c = 1/4$

2) Find $F(x)$

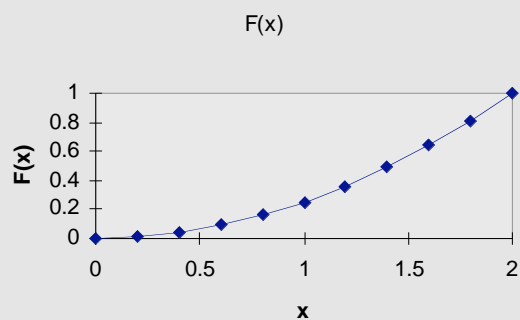
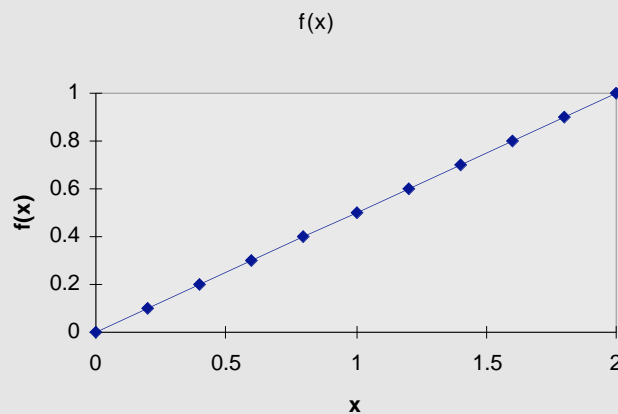
$$F(x) = \int_0^x \frac{1}{4} x dx = \frac{1}{4} \left[\frac{x^2}{2} \right]_0^x = \frac{x^2}{8}$$

Check the cdf @ $x=0$, $F(x) = 0$, @ $x=2$, $F(x) = 1$
Thus,

Example 3-15. continued

$$F(x) = \begin{cases} \frac{x^2}{4} & \text{for } 0 \leq x \leq 2 \\ 0 & \text{Elsewhere} \end{cases}$$

3) Graph $f(x)$ and $F(x)$



4) Determine $P(1 \leq X \leq 2)$

$$P(1 \leq X \leq 2) = F(2) - F(1) = 0.5 \left[\frac{2^2}{2} - \frac{1^2}{2} \right] = 0.75$$

Example 3-16. In a financial study, you estimate that a savings of \$60,000 was obtained by using a particular product of a research program. The estimate, while logical, has considerable uncertainty associated with it. Since this estimate will be added to others to produce a total sum of the benefits of the research program, you seek a suitable distribution to define the random variable X representing the true benefit of using the product. We desire the distribution to have three characteristics:

- (1) it cannot show a benefit less than zero;
- (2) its upper limit is bound-less; and
- (3) it can be considered skewed.

One such distribution function is the Weibull distribution that can be expressed in general form as

$$f(x) = \frac{\alpha x^{\alpha-1}}{\beta^\alpha} e^{-\left(\frac{x}{\beta}\right)^\alpha}$$

In this example

$$f(x) = \frac{1.5x^{\alpha-1}}{(60,000)^{1.5}} e^{-\left(\frac{x}{60,000}\right)^{1.5}}$$

and

$$F(x) = \int_0^x f(x)dx = 1 - e^{-\left(\frac{x}{60,000}\right)^{1.5}}$$

- 1) Determine the value of $X = x$ such that 50% of the values will fall to the left of x and 50% will fall to the right.

$$P(X \leq x) = 0.5 = 1 - e^{-\left(\frac{x}{60,000}\right)^{1.5}} \Rightarrow x = \$34,625$$

- 2) What is the probability that the benefits will fall between \$50,000 and \$70,000?

$$P(50,000 \leq X \leq 70,000) = 0.76 - 0.53 = 0.23$$

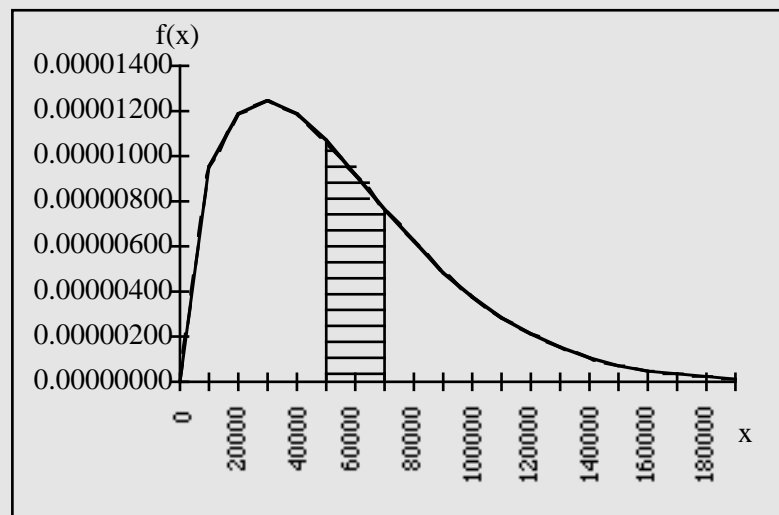


Figure 3-6. Weibull Distribution showing $P(50,000 \leq X \leq 70,000)$

Special Distributions

Several distributions are typically used to describe the properties of most variables. These special distributions (uniform, normal, and Poisson) all find applications in a wide class of civil engineering problems.

The Uniform Distribution

Let X be a random variable defined on an interval from a to b . Place a requirement on the pdf $f(x)$ that each number in the interval have an equal likelihood of selection. A model for this assertion is termed the *uniform distribution*. The equation for this distribution is in Figure 3-7.

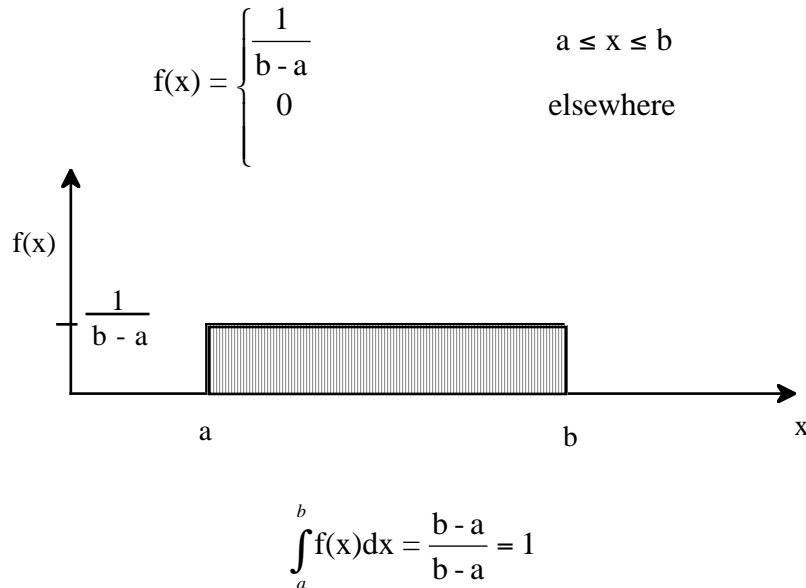


Figure 3-7. The uniform distribution

One of the most important uses of a uniform distribution is for the selection of a random number for a Monte Carlo simulation. Drawing a random number from a Uniform Distribution means that each number in the interval (0,1) has an equal likelihood of selection. Most computer languages and spreadsheets have an internal random number generator. One popular generator uses an overflow from a division process to find the number.

The Normal Probability Distribution

One of the most important continuous random variables is the normal (or Gaussian) distribution. It describes the behavior of many natural systems. For instance, consider the errors of a survey crew to measure the centerline of a highway. Using a stadia rod, a sonic measuring device, or a laser device, the distance will be made up of a series of shorter measurements. These measurements will be read to some discrete length, say to the nearest millimeter, nearest one-hundredth of an inch, etc. Then a distance of 100 meters will be between 99.99 and 100.01 meters if the accuracy is to the nearest millimeter.

The normal distribution theory describes the behavior of random variables in these situations:

1. The sum of random variables in which no single random variable dominates.
2. Errors in measurements.
3. Capacity of a system that fails after the saturation of redundant components have taken place for example,
 - 3.1 capacity of road, the sum of its lane capacities,

Chapter 3

- 3.2 collapse mechanism of a ductile, elastic-plastic frame,.
- 3.3 the sum of constants times the yield moments of specific joints, and
- 3.4 the distribution of total annual rainfall run-off over a specific area.

The normal distribution is also important because of the central limit theorem (CLT). Under very general conditions, it states that the average of a large number of random variables can be approximated by a normal distribution. Also, the normal distribution is often used as an approximation to other distributions under some situations to include the Poisson (when the mean is large), binomial, and Gamma.

The pdf for the normal distribution is of the form

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad -\infty < x < +\infty \quad (3-17)$$

A short hand notation of $N(\mu, \sigma^2)$ is often used. A plot of a normal distribution is in Figure 3-8.

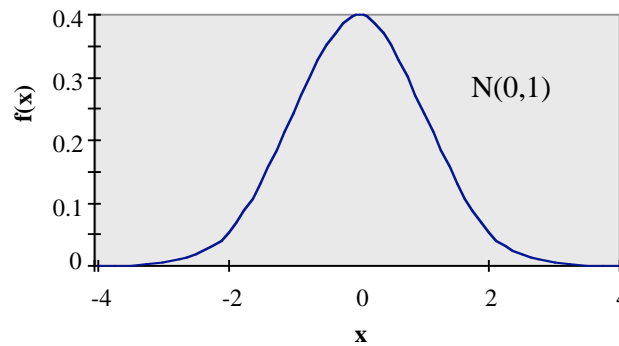


Figure 3-8. Graph of a normal pdf

A normal distribution with the parameters $\mu=0$ and $\sigma=1$ is called the standard normal distribution. Though not used for any existing problems, it has utility from which information about other normal distributions can be obtained. Because of its wide usage, a special notation $\Phi(z)$ is commonly used. The distribution function $N(0,1)$ is usually tabulated in tables (see Figure C.9 and Table C.1). However, with the development and proliferation of portable computers and most hand held calculators, the use of tables is obsolete technology. Using the standard normal pdf, properties of other normal pdfs can be determined based upon this methodology. Suppose for $N(\mu, \sigma^2)$ then

$$P(a \leq X \leq b) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \quad (3-18)$$

The solution to this integral does not exist in closed form. But, it can be evaluated indirectly by making this change of variable

$$z = \frac{x - \mu}{\sigma} \quad \text{and} \quad dx = \sigma dz$$

Then

$$P(a \leq X \leq b) = \frac{1}{\sqrt{2\pi}} \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} e^{-\frac{1}{2}z^2} dz \quad (3-19)$$

If you set $\mu = 0$ and $\sigma = 1$ in Equation 18 it will reduce to Equation 19. Thus, Equation 18 is the area of the standard normal pdf between $(a-\mu)/\sigma$ and $(b-\mu)/\sigma$. Thus

$$P(a \leq X \leq b) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right) \quad (3-20)$$

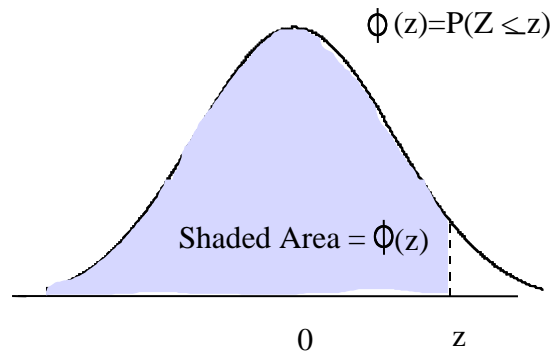


Figure 3-9. Standard normal curve areas

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.00034	0.00032	0.00031	0.00030	0.00029	0.00028	0.00027	0.00026	0.00025	0.00024
-3.3	0.00048	0.00047	0.00045	0.00043	0.00042	0.00040	0.00039	0.00038	0.00036	0.00035
-3.2	0.00069	0.00066	0.00064	0.00062	0.00060	0.00058	0.00056	0.00054	0.00052	0.00050
-3.1	0.00097	0.00094	0.00090	0.00087	0.00084	0.00082	0.00079	0.00076	0.00074	0.00071
-3	0.00135	0.00131	0.00126	0.00122	0.00118	0.00114	0.00111	0.00107	0.00104	0.00100
-2.9	0.00187	0.00181	0.00175	0.00169	0.00164	0.00159	0.00154	0.00149	0.00144	0.00139
-2.8	0.00256	0.00248	0.00240	0.00233	0.00226	0.00219	0.00212	0.00205	0.00199	0.00193
-2.7	0.00347	0.00336	0.00326	0.00317	0.00307	0.00298	0.00289	0.00280	0.00272	0.00264
-2.6	0.00466	0.00453	0.00440	0.00427	0.00415	0.00402	0.00391	0.00379	0.00368	0.00357
-2.5	0.00621	0.00604	0.00587	0.00570	0.00554	0.00539	0.00523	0.00508	0.00494	0.00480
-2.4	0.00820	0.00798	0.00776	0.00755	0.00734	0.00714	0.00695	0.00676	0.00657	0.00639
-2.3	0.01072	0.01044	0.01017	0.00990	0.00964	0.00939	0.00914	0.00889	0.00866	0.00842
-2.2	0.01390	0.01355	0.01321	0.01287	0.01255	0.01222	0.01191	0.01160	0.01130	0.01101
-2.1	0.01786	0.01743	0.01700	0.01659	0.01618	0.01578	0.01539	0.01500	0.01463	0.01426
-2	0.02275	0.02222	0.02169	0.02118	0.02068	0.02018	0.01970	0.01923	0.01876	0.01831
-1.9	0.02872	0.02807	0.02743	0.02680	0.02619	0.02559	0.02500	0.02442	0.02385	0.02330
-1.8	0.03593	0.03515	0.03438	0.03362	0.03288	0.03216	0.03144	0.03074	0.03005	0.02938
-1.7	0.04457	0.04363	0.04272	0.04182	0.04093	0.04006	0.03920	0.03836	0.03754	0.03673
-1.6	0.05480	0.05370	0.05262	0.05155	0.05050	0.04947	0.04846	0.04746	0.04648	0.04551
-1.5	0.06681	0.06552	0.06426	0.06301	0.06178	0.06057	0.05938	0.05821	0.05705	0.05592
-1.4	0.08076	0.07927	0.07780	0.07636	0.07493	0.07353	0.07215	0.07078	0.06944	0.06811
-1.3	0.09680	0.09510	0.09342	0.09176	0.09012	0.08851	0.08692	0.08534	0.08379	0.08226
-1.2	0.11507	0.11314	0.11123	0.10935	0.10749	0.10565	0.10383	0.10204	0.10027	0.09853
-1.1	0.13567	0.13350	0.13136	0.12924	0.12714	0.12507	0.12302	0.12100	0.11900	0.11702
-1	0.15866	0.15625	0.15386	0.15151	0.14917	0.14686	0.14457	0.14231	0.14007	0.13786
-0.9	0.18406	0.18141	0.17879	0.17619	0.17361	0.17106	0.16853	0.16602	0.16354	0.16109
-0.8	0.21186	0.20897	0.20611	0.20327	0.20045	0.19766	0.19489	0.19215	0.18943	0.18673
-0.7	0.24196	0.23885	0.23576	0.23270	0.22965	0.22663	0.22363	0.22065	0.21770	0.21476
-0.6	0.27425	0.27093	0.26763	0.26435	0.26109	0.25785	0.25463	0.25143	0.24825	0.24510
-0.5	0.30854	0.30503	0.30153	0.29806	0.29460	0.29116	0.28774	0.28434	0.28096	0.27760
-0.4	0.34458	0.34090	0.33724	0.33360	0.32997	0.32636	0.32276	0.31918	0.31561	0.31207
-0.3	0.38209	0.37828	0.37448	0.37070	0.36693	0.36317	0.35942	0.35569	0.35197	0.34827
-0.2	0.42074	0.41683	0.41294	0.40905	0.40517	0.40129	0.39743	0.39358	0.38974	0.38591
-0.1	0.46017	0.45620	0.45224	0.44828	0.44433	0.44038	0.43644	0.43251	0.42858	0.42465
0	0.50000	0.49601	0.49202	0.48803	0.48405	0.48006	0.47608	0.47210	0.46812	0.46414

Table 3-1. Normal distribution tables

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91308	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976

Table 3- 1. continued

Example 3-17. Suppose you are planning on developing a new shopping mall in an area that adjoins a flood plain. Data analysis of the gage readings from the small river that flows through the middle of the flood plain is $N(40,100)$. The proposed elevation for the parking lots and grade of the mall is 52 mean sea level (MSL).

1) What is the probability that the mall and parking lot will be flooded?

$$P(X \geq 52) = P(52 \leq X \leq \infty) = \Phi(\infty) - \Phi\left(\frac{52 - 40}{10}\right) = 1.0 - 0.88493 = 0.1106 \text{ or } 11.06\%$$

2) Your insurance company wants you to build the mall to height where it will flood only 1% of the time. Based upon $N(40,100)$ what is this height of the first floor finished grade?

$$P(a \leq X \leq \infty) = \Phi(\infty) - \Phi\left(\frac{a - 40}{10}\right) = 0.01$$

From Table C.1 for $\Phi\left(\frac{a - 40}{10}\right) = 0.99$, $\Phi(z) = 2.33$. Thus, $a = 63.3$ MSL.

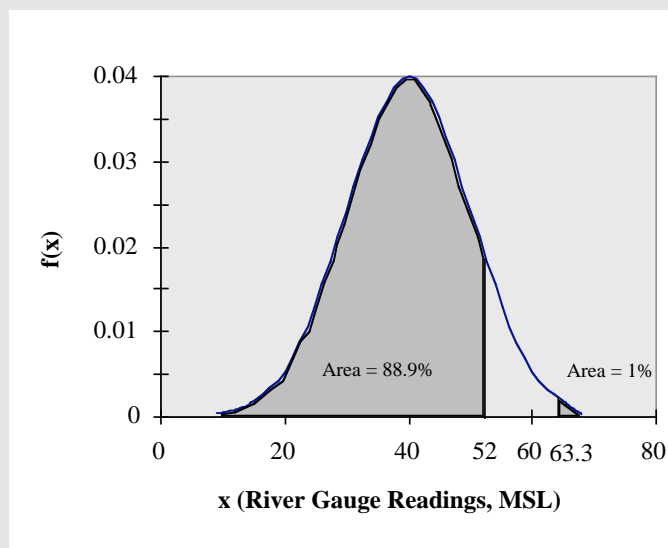


Figure 3-10. Normal pdf for the river gage readings

Example 3-18. Let X be $N(\mu, \sigma^2)$. Find $p(\mu - 2\sigma \leq X \leq \mu + 2\sigma)$.

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = P\left(\frac{\mu - 2\sigma - \mu}{\sigma} \leq X \leq \frac{\mu + 2\sigma - \mu}{\sigma}\right)$$

$$= \Phi(2) - \Phi(-2) = 0.977 - (1 - 0.977)$$

$$= 0.954$$

Thus, 95.4% of all x values will fall within $\pm 2\sigma$

The Poisson Distribution

There is no simple experiment on which the Poisson distribution is based. However, a number of random events that occur in a unit of time, space, or any other dimension follow this distribution.

A random variable X is said to have a Poisson distribution if the pdf of X is

$$f(x) = \frac{e^{-\lambda} \lambda^x}{x!} \quad x = 0, 1, 2, \dots \quad (3-21)$$

Like all pdfs, the Poisson probabilities do sum to unity. That is

$$\begin{aligned} F(x) &= \sum_{x=0}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} e^{\lambda} = 1 \end{aligned}$$

Example 3-19. The mean number of trucks hauling waste to a landfill, enters a preprocessing facility per 2 minute period, is one. An excessive number of trucks entering the facility during a brief period can cause major traffic and safety problems. Find the probability that the number of trucks entering the facility during a two-minute period exceeds the design amount of three?

$$\lambda = 1 \text{ truck/2 minutes}$$

$$p(X > 3) = 1 - (p(0) + p(1) + p(2) + p(3))$$

$$= 1 - \sum_{x=1}^3 \frac{e^{-\lambda} \lambda^x}{x!}$$

$$= 0.019$$

Mathematical Expectation

The concept of expected value is one of the most powerful in statistics. In a world of variation, we use averages or expected values to define what we are talking about. If for example the strength of a particular concrete mix is 3000 psi, it actually means that the expected strength of the mix is 3000 psi. It is a rare case indeed that a cylinder will fail at exactly 3000 psi.

Chapter 3

If X is a continuous random variable with a pdf of $f(x)$ then the expected value of X is defined as

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx \quad (3-22)$$

If X is a discrete random variable,

$$E(X) = \sum_{n=-\infty}^{\infty} xf(x)dx \quad (3-23)$$

Example 3-20. Let X have the pdf

$$f(x) = 5e^{-5x}$$

Find the expected value of the function.

$$E(x) = \int_{-\infty}^{\infty} 5xe^{-5x} dx = \frac{1}{5}$$

If the range of the pdf is discrete, then simply sum $xf(x)$ over all possible values of X . The expected value is also referred to as the first moment about zero and the *mean* of a pdf. If X has pdf $f(x)$ then

$$\mu = \bar{x} = E(x) = \int_{-\infty}^{\infty} xf(x)dx \quad (3-24)$$

which is a strong measure of the central tendency of the distribution.

The second moment about the mean is the *variance* of a pdf. If the mean of the pdf is μ then

$$\sigma^2 = E[(x - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx$$

This equation can be reduced to the more computational friendly version

$$= \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2 \quad (3-25)$$

$$= E(x^2) - \mu^2 \quad (3-26)$$

Like the expected value, if the range of the pdf is finite, then simply sum over all possible X values.

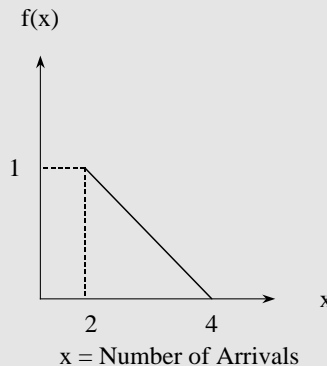
For the discrete case,

$$\mu = \bar{x} = E(x) = \sum_{n=-\infty}^{\infty} xf(x)dx \quad (3-27)$$

and

$$\begin{aligned} \sigma^2 &= E(x - \mu)^2 = \sum_{n=-\infty}^{\infty} (x - \mu)^2 f(x)dx \\ &= \sum_{n=-\infty}^{\infty} x^2 f(x)dx - \mu^2 \\ &= E(x^2) - \mu^2 \end{aligned} \quad (3-28)$$

Example 3-21. One of the most common uses of the cdf is to model data for computer simulation. For example, suppose you were developing an airport simulation. After collecting data you hypothesized that airplane arrivals per hour has this pdf.



1) Determine the pdf

The equation of a line with a negative slope is used to produce

$$f(x) = \begin{cases} 2 - \frac{x}{2} & 2 \leq x \leq 4 \\ 0 & \text{Otherwise} \end{cases}$$

One way to check that this is a valid pdf, is that $\int_a^b f(x)dx = 1$ or

$$\int_2^4 [2 - \frac{x}{2}]dx = 2x - \frac{x^2}{4} = 1$$

2) Determine F(x)

$$F(x) = \int_2^x (2 - \frac{x}{2})dx = 2x - \frac{x^2}{4} \Big|_2^x = 2x - \frac{x^2}{4} - 3 \text{ or}$$

$$F(x) = \begin{cases} 2x - \frac{x^2}{4} - 3 & 2 \leq x \leq 4 \\ 0 & \text{Otherwise} \end{cases}$$

We check to the cdf @x=2, $4 - 1 - 3 = 0$ and @x=4, $8 - 4 - 3 = 1$

3) Determine the expected value of the pdf

$$E(X) = \int_a^b xf(x)dx = \int_2^4 x[2 - \frac{x}{2}]dx = [x^2 - \frac{x^3}{6}] = 2\frac{2}{3}$$

4) Compute the variance and standard deviation

$$\sigma^2 = \int_a^b x^2 f(x)dx - \mu^2 = \int_2^4 x^2 [2 - \frac{x}{2}]dx = [2\frac{x^3}{3} - \frac{x^4}{8}]_2^4 - (2\frac{2}{3})^2 = \frac{110}{9}$$

Thus, $\sigma = 3.496$

Discrete Process Generators

Discrete Process Generators (DPG's) associate an outcome with a random number generated from a Uniform (0, 1) distribution. We use a U(0, 1) because this is the only random number distribution most software applications can approximate. Other random number generators in software applications are nothing more than functions of the U(0, 1) distribution. You will use a DPG when you wish to simulate an experiment that has a finite set of possible outcomes.

Before we discuss the actual development of a DPG, we will discuss the concept. Understanding DPG's is easy if you think of it in terms of a spinner that comes with a board game. Assume we had a game with 4 outcomes, and each outcome has a different chance of occurring. The outcomes and their associated probabilities are given below:

Outcome	Probability
<i>Go Forward 1</i>	.50
<i>Go Back 1</i>	.25
<i>Go Forward 2</i>	.20
<i>Go Back to Start</i>	.05

Table 3-2. Spinner Values

If we were to represent these outcomes and their probabilities on a game spinner, the spinner would look like Figure 3-11. The area of the circle allocated to each outcome is equal to the associated probability for each outcome. Therefore if we were to spin the spinner, 50% of the area where it could end up would indicate the outcome "*Go Forward 1*". Each of the outcomes has a portion of the circle that is equal to its probability of occurrence.

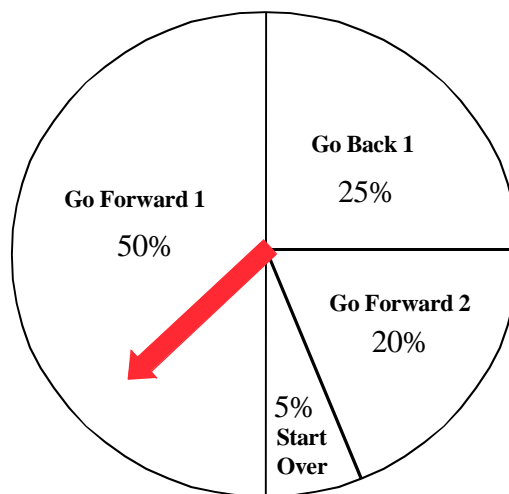


Figure 3-11. Spinner for Game

We could take this spinner analogy one step further and associate it with random numbers. You could also use 100 equally sized pieces of paper, number them from 0 to 99, and toss them into a hat. Then, you could randomly draw, with replacement, a piece of paper. The number on the paper would correspond to one of the possible outcomes. Because 50% of the area is associated with "*Go Forward 1*", we could let the numbers 0 through 49 indicate the outcome of moving forward 1. 25% of the area is associated with "*Go Back 1*" so we should let 25 of the numbers indicate the outcome of going back

Chapter 3

one. But we cannot use 0 through 24 because those numbers are included in "Go Forward 1". Therefore we will use the next 25 numbers that are 50 through 74. We continue this way until all of the numbered pieces of paper are associated with an outcome on the board. Table 3-3 shows which numbers corresponds with each outcome.

Outcome	Probability	CDF	Number Range
<i>Go Forward 1</i>	.50	.50	0 - 49
<i>Go Back 1</i>	.25	.75	50 - 74
<i>Go Forward 2</i>	.20	.95	75 - 94
<i>Go Back to Start</i>	.05	1.0	95 - 99

Table 3-12. Game Outcome Generator

For example if you pulled the number 64 out of the hat, you know you "Go Back 1". Likewise, drawing an 83 means you would "Go Forward 2".

Developing a DPG is very similar to drawing numbers out of a hat - it's just a lot faster and more useful in simulation. The difference is in the hat example we had only 100 numbers to draw. In a $U(0, 1)$ distribution there are an infinite number between 0 and 1 that can be drawn. We can handle that however when we set up the DPG. There are four steps to developing a DPG:

1. Classify and Count outcomes - From a sample set of historical data, identify all possible, relevant outcomes on the process you wish to model. Then based on the historical data, determine how many times each outcome occurred or was observed.
2. Determine relative frequency of occurrence - Divide the number of times each outcome was observed by the total number of observations. This is the Probability Mass Function (PMF) for the experiment.
3. Determine Cumulative Frequency Distribution (CDF)
4. Determine random number ranges - Identify the range of numbers, based on the CDF that will associate a $U(0, 1)$ random number with a specific outcome from the observed event. The CDF is used to set the lower and upper bounds for each range.

These 4 steps, when carried out in sequence, define a DPG that will return appropriate outcomes on the experiment, based on a series of random numbers from a $U(0, 1)$ generator. Each of these steps is demonstrated in the following scenario:

Chapter 3

You wish to simulate the business at a local call center so you can determine how many employees of certain types should be trained and hired. You observed the order counter during the lunch rush one day. One of the pieces of information you collected was the type of information ordered by each customer. The results of your observations are provided in the table below:

Call Center Information Demand				
Shipping Status	Payment Questions	Shipping Status	Payment Questions	Order Status
Order Status	Shipping Status	Order Status	Order Status	Shipping Status
Technical Support	Order Status	Technical Support	Order Status	Order Status
Payment Questions	Technical Support	Shipping Status	Technical Support	Shipping Status
Technical Support	Shipping Status	Technical Support	Payment Questions	Technical Support
Payment Questions	Technical Support	Request for Supervisor	Technical Support	Payment Questions

We see this table consists of a finite set of outcomes and that we could use it to develop a DPG in order to simulate the ordering process. The first step in developing the DPG is:

Classify and count outcomes - We see there are only 5 different choices to make for the customers. We identify and classify them by their type: Order Status, Shipping Status, Payment Questions, Technical Support, and the Request for Supervisor. We count the number of times each outcome was observed and we record it:

Information	Times Ordered
<i>Order Status</i>	7
<i>Shipping Status</i>	7
<i>Payment Questions</i>	6
<i>Technical Support</i>	9
<i>Request for Supervisor</i>	1

Determine relative frequency of occurrence - Divide each value by 30, the total number of observations, and record the resulting decimal value in the table.

Information Requested	Times Ordered	Frequency (PMF)
<i>Order Status</i>	7	$7/30 = .2333$
<i>Shipping Status</i>	7	$7/30 = .2333$
<i>Payment Questions</i>	6	$6/30 = .2000$
<i>Technical Support</i>	9	$9/30 = .3000$
<i>Request for Supervisor</i>	1	$1/30 = .0333$

Determine the CDF - Add each of the previous probabilities to derive a CDF value for each outcome in the experiment. We do this so we can avoid duplicate outcomes for some random numbers as discussed in the spinner example.

Information Requested	Times Ordered	Frequency (PMF)	CDF
<i>Order Status</i>	7	.2333	.2333
<i>Shipping Status</i>	7	.2333	.4666
<i>Payment Questions</i>	6	.2000	.6666
<i>Technical Support</i>	9	.3000	.9666
<i>Request for Supervisor</i>	1	.0333	1.000

Determine random number ranges - By developing a range from the CDF calculated above, and comparing a $U(0, 1)$ random number "r" to the set of ranges, we can simulate random information request orders that follow the probability distribution from the original observed data. We develop each range by allowing each outcome's CDF value to be the upper bound of its range and the previous outcome's CDF value to be the lower bound for the range.

Notice only the lower bound value has the "equality" associated with it. This is for two reasons. We cannot let the value .2333 represent both the *Order Status* and the *Shipping Status* outcomes, so we will let any value just up to but not including .2333 represent the Order Status. Secondly, while the random number generator can produce the value 0 (zero) it cannot produce the value 1 (one). Therefore we do not want to erroneously bias the DPG against the Request for Supervisor frequency by inadvertently decreasing its range.

Information Requested	Times Ordered	Frequency (PMF)	CDF	Random Number Range
<i>Order Status</i>	7	.2333	.2333	0.000 $\leq r_1 <$.2333
<i>Shipping Status</i>	7	.2333	.4666	.2333 $\leq r_1 <$.4666
<i>Payment Questions</i>	6	.2000	.6666	.4666 $\leq r_1 <$.6666
<i>Technical Support</i>	9	.3000	.9666	.6666 $\leq r_1 <$.9666
<i>Request for Supervisor</i>	1	.0333	1.000	.9666 $\leq r_1 <$ 1.000

We use the DPG by generating a series of random numbers. Each of these random numbers represents a customer's information requests. Therefore, from 5 different random numbers, we could generate 5 requests. Table 3-4 shows an application of the constructed information requested DPG. Just the DPG alone does not generate sufficient information to qualify as a simulation, generally a DPG will be used in conjunction with other process generators in order to create a model that returns interesting information.

A graphical representation of the DPG and how the information requested selection is related to the random number is presented in Figure 5. The random number r receives the value .50, its location is identified on the r axis (vertical axis). We extend a line from the axis to the CDF function. Where it touches the CDF, we drop a line and identify which information request was "generated" via the information request DPG and the U(0, 1) random number.

Table 3-4. Information Request DPG Being Used

Random Number Generated	Information Request Simulated
0.1531	Order Status
0.6937	Technical Support
0.8508	Technical Support
0.6297	Payment Questions
0.3118	Shipping Status

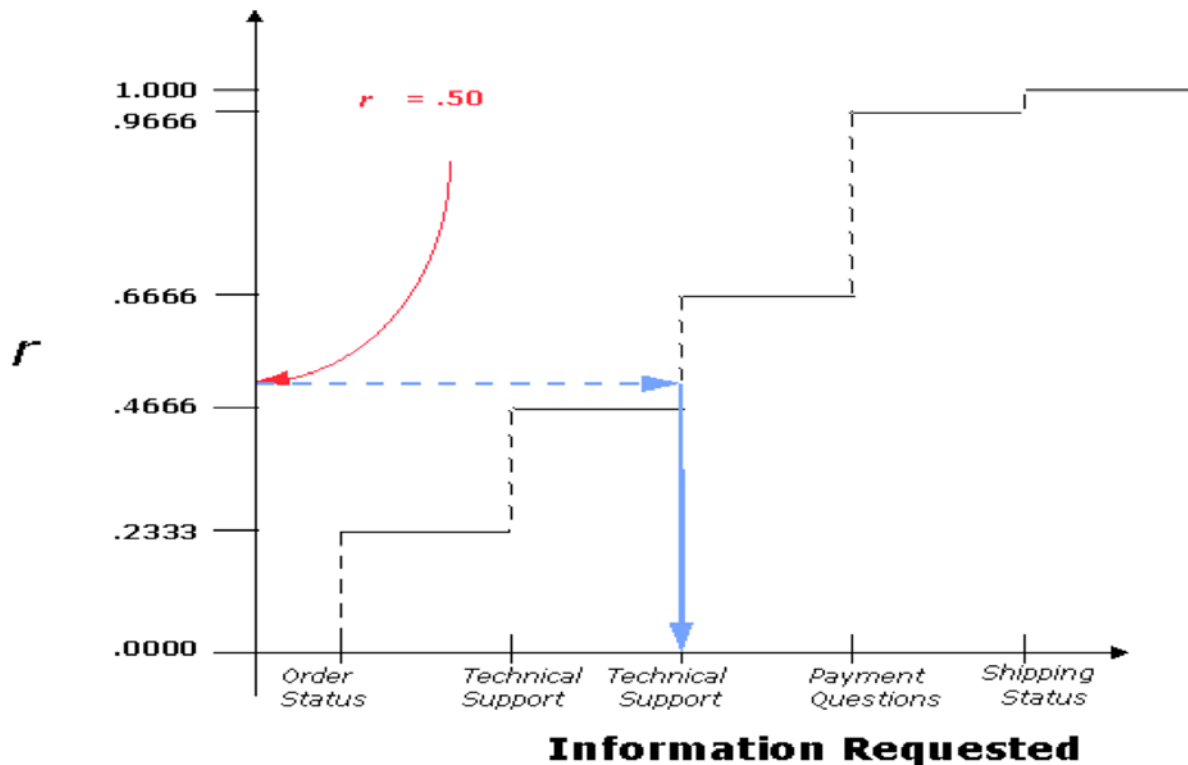


Figure 3-12. Graphical Representation of Information Request DPG

Continuous Process Generators

The discrete process generators you learned in the previous lesson are extremely useful, cover many applications, and you will find yourself using them frequently in simulation. But not all processes are discrete. Some of them have continuous results; the sample space is infinite. For example, the quantity of oil pumped from a well in one hour could be 34 barrels, it could be 34.12 barrels, it could be 34.1275 barrels, ... in fact, it could be any number of barrels between 0 and some reasonable number. There are an infinite

Chapter 3

number of possibilities and therefore the process must be simulated with a Continuous Process Generator (CPG). In this lesson you will be introduced to seven different CPG's. This set of seven is by no means all inclusive, but it does provide you with a variety of distributions that account for many of the random processes you will encounter when simulating both in this course and in the Army.

Inverse Transform Method

The method used to develop random numbers from a continuous distribution is called the Inverse Transform Method. The algorithm for using the inverse transform method is:

Step 1. Generate a random outcome r from the Uniform(0, 1) distribution.

Step 2. Substitute the number r into the CDF of the continuous distribution and solve for x using

$$x = F^{-1}(r).$$

You only need to select or develop the inverse of the CDF ($F^{-1}(r)$) once for each different distribution and then evaluate that expression as many times as you need a random number from that distribution. The Inverse Transform Method makes more sense if you look at it graphically. Figure 1 shows the CDF for an Exponential distribution with $\lambda = 2$, given by the expression:

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

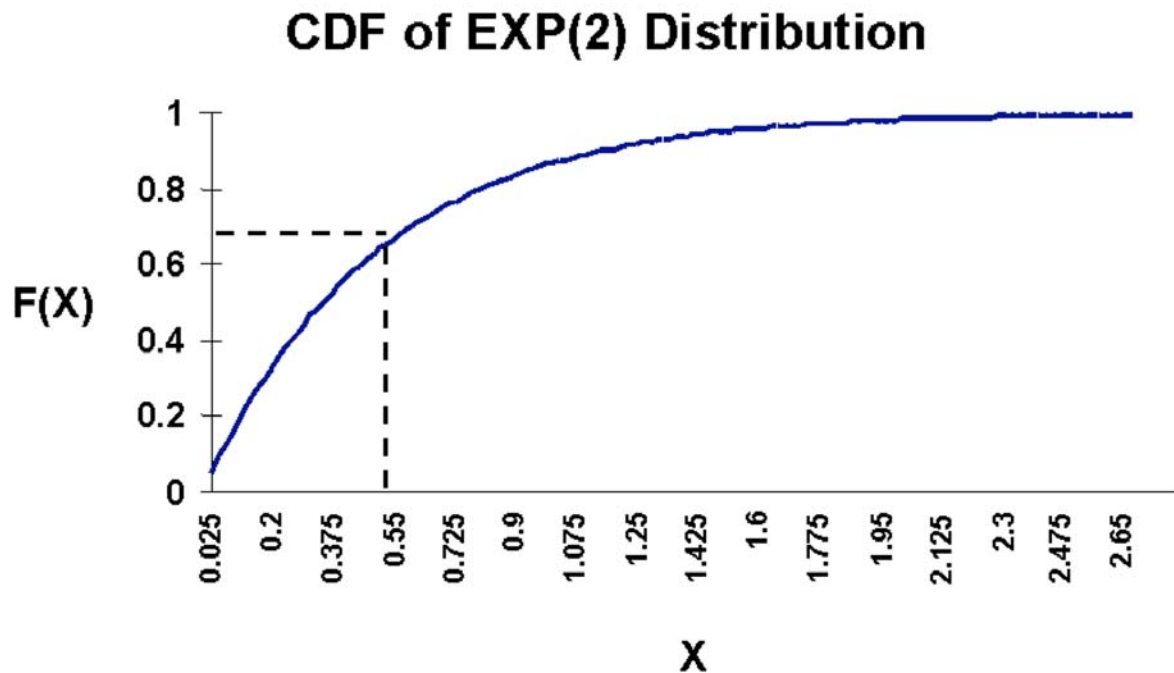


Figure 3-13. CDF for an Exponential Distribution

If you wanted to know the probability of getting a value of .55 or smaller, you could follow the dashed line up to where it intersected with the CDF curve, then read across the solid line to determine the probability to be approximately .667. Alternatively, you could use the equation for the CDF and the desired x value and determine the probability of getting .55 or less directly:

$$F(x) = 1 - e^{-2(.55)} = .667$$

Figure 3-14 shows the same CDF curve but instead shows what the inverse Transform Method does. You provide the probability, in this case .80, and read across the dotted line until you reach the CDF curve. Then read down the solid line and determine the x value for which the CDF has value .80. From Figure 2 we see it is approximately .80. Or we could use the equation for the inverse of the CDF of the exponential distribution and determine the X value is approximately

$$F(x) = 1 - e^{-2x} = .80$$

$$1 - .8 = e^{-2x}$$

$$\ln(.2) = -2x$$

$$\ln(.2) / -2 = x = .805$$

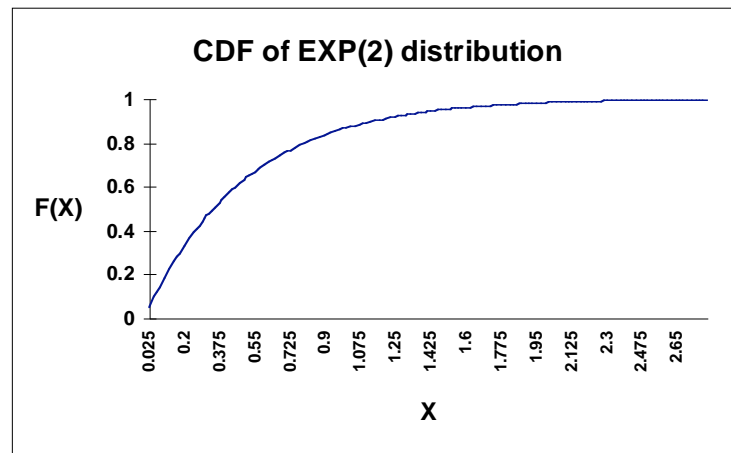


Figure 3-14. Same CDF as an Inverse Transform Function

The difficulty for some comes from how we can relate a $U(0, 1)$ random number to a probability from an exponential distribution. To show how this can be the case, assume we have the CDF of a continuous distribution, by definition the CDF is

$$F(x) = P[X \leq x]$$

Now we let the following hold true:

$$Y = F(X)$$

What this means is we will map each outcome x on the process to a value $y = F(x)$. Now Y is a random variable and its distribution is $U(0, 1)$. To see this, construct Y 's CDF, $F_Y(y)$, and see it is the CDF of a $U(0, 1)$ distribution (i.e., $F_Y(y) = y$ for $0 < y < 1$). Using the definition of CDF's, for any y between 0 and 1, we have:

$F_Y(y) = P[Y \leq y]$	(definition of y 's CDF)
$= P[F(X) \leq y]$	(because $Y = F(X)$ - we made a substitution)
$= P[X \leq F^{-1}(y)]$	(solved the equation inside the brackets $F(X) < y$ for X)
$= F(F^{-1}(y))$	(this is nothing more than the definition of the CDF again but this time, we have replaced the right hand side of the equation with the left hand side of the equation)
$= y$	(The function of its inverse is the identity - you get back what you put in)

Chapter 3

So we now have $F(y) = y$ for $0 < y < 1$, which is a uniform distribution, and is what we set out to prove in the first place. The significance of this is, for any random variable X with a continuous distribution and a CDF, we can generate outcomes on X by generating outcomes on $Y = F(X)$ (which can be done with a $U(0, 1)$ generator) and “inverse” transforming them to corresponding outcomes on X . This is the basis of almost all CPG's. This will provide us over the long run with a sample of x values that are exponentially distributed. Since probabilities are limited between 0 and 1 and they are evenly spaced across the $F(X)$ axis (the probability axis) in a CDF, in simulations we make the substitution of the $U(0, 1)$ number for $F(X)$ or $P(X \leq x)$ value and solve for the corresponding X value. Over the long run we could expect to evenly sample all probabilities from 0 to 1, this in turn would generate X values that cover the spectrum of the stated distribution. This is easier to perform than it is to understand.

A distinct disadvantage with the Inverse Transform Method is it does not work for continuous distributions which we do not have a closed form CDF - the Normal distribution should come to mind here¹. For these cases, special approximations have been created.

Exponential CPG Derived

There is no magic involved with generating the CPG functions provided in this block of instruction. They are nothing more than the inverse CDF's, with respect to the x value of the CDF's for certain distributions. Although you will not be required to derive CPG's from PDF's, you will be required to develop them from standard format equations, those provided on the formula cards. To demonstrate how these CPG's are developed (in case you find yourself needing to derive them for your design project or in case you really want to impress someone on a date!) Begin with the CDF for the Exponential distribution. Recall for an exponential distribution the CDF is $F(x) = 1 - e^{-\lambda x}$.

Let r be a $U(0, 1)$ outcome and substitute it for $F(x)$.

$$r = 1 - e^{-\lambda x}$$

Isolate the term with x in it:

$$r - 1 = -e^{-\lambda x}$$

Multiply through by negative 1:

$$1 - r = e^{-\lambda x}$$

Take the natural log of both sides:

$$\begin{aligned}\ln(1 - r) &= \ln(e^{-\lambda x}) \\ \ln(1 - r) &= -\lambda x\end{aligned}$$

Divide through by negative λ :

$$-\frac{\ln(1 - r)}{\lambda} = x$$

which gives the CPG for the exponential distribution. If you know λ and you have the means of producing $U(0, 1)$ random numbers, you can create $\text{Exponential}(\lambda)$ random numbers with this CPG.

Sample Application Exponential CPG

To demonstrate the application of the exponential CPG, let's take a situation from queuing theory and simulate the arrival rate of customers entering a system. Assume customers arrive at the "Quickie-Mart" for Squishies on average every 4 minutes (following an exponential distribution). We can simulate this process by using the Exponential CPG and substituting the appropriate values for r and λ .

First we develop the CPG:

We were given an average inter-arrival time ($1/\lambda$), not a rate (λ), so we must convert it to a rate. The arrival rate is:

$$\lambda = \frac{1}{\mu} = \frac{1}{4 \text{ minutes} / 1 \text{ customer}} = \frac{1 \text{ customer}}{4 \text{ minutes}} = \frac{.25 \text{ customers}}{\text{minute}}$$

Therefore the Exponential CPG for this situation is:

$$-\frac{\ln(1-r)}{\lambda} = x \Rightarrow -\frac{\ln(1-r)}{.25} = x$$

Using the following stream of random numbers r , obtained through calls to a $U(0, 1)$ generator, we can now determine the appropriate inter-arrival time between customers:

r	Time (min)	r	Time (min)
0.313283	1.503335	0.402831	2.062223
0.2114	0.949982	0.996252	22.3462
0.366326	1.82488	0.76487	5.790461
0.894921	9.012186	0.125705	0.537352
0.234296	1.067839	0.032982	0.134152
0.232669	1.05935	0.817108	6.795447
0.257893	1.193047	0.525263	2.979975
0.978617	15.3806	0.316599	1.522691
0.206963	0.927544	0.067519	0.279627
0.799827	6.434295	0.96842	13.82097

Table 3-5. Table of Generated Inter-Arrival Times

The first value is calculated in detail below:

$$-\frac{\ln(1-r)}{.25} = -\frac{\ln(1-.313283)}{.25} = -\frac{\ln(.686717)}{.25} = \frac{.37583}{.25} = 1.5033 \text{ minutes}$$

The average of the sample of inter-arrival times is 4.7811. You may feel this is not too close to the theoretical inter-arrival time of 4.0 minutes, but do not forget, we only have 20 observations. If you repeated this experiment with many observations you would find that averaging the data would give values asymptotically approaching the true mean. The histogram in Figure 3-15 shows the histogram of the data.

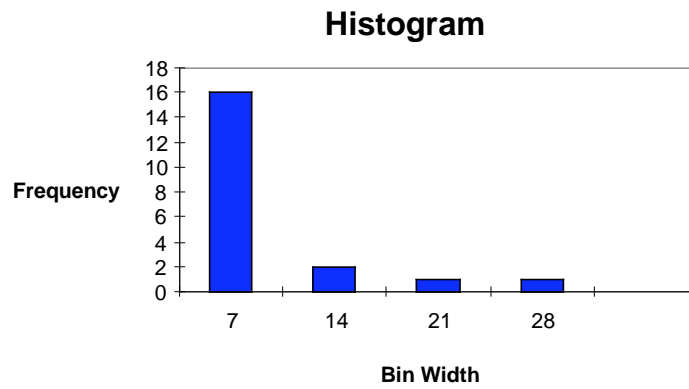


Figure 3-15. Histogram of Random Numbers

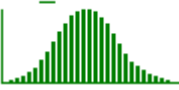
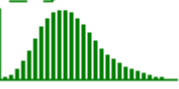
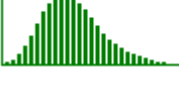







Conclusion

Obviously, a single discussion doesn't begin to develop the probabilistic and statistical concepts needed by an engineer. Statistical analysis must be persuasive in our professions. A better understanding of the randomness that affects complex systems, combined with better analytical tools, allows us the ability to incorporate statistical methods in the design process.

Discrete Process Generators allow us to simulate events with limited outcomes and control each outcome's probability of happening. The 4 steps explained in detail throughout this reading allow you to successfully create DPG's regardless of the discrete event being modeled. The concept and development of DPG's presented in this reading integrate seamlessly with spreadsheet simulations.

With the knowledge gained from this reading, you now have the ability to create simulations with either or both continuous and discrete distributions. You are not required to memorize the CPG formulas, you have formula cards that you can use, but you do need to know which CPG is appropriate to use and what the correct parameters are.

Appendix 3-A Functions from Crystal Ball

Distribution	Description ¹
	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.
	Models the number of trials required to achieve k successes; for example the number of computer chip that we must inspect to find 4 defective chips.
	Models the number of independent events that occur in a fixed amount of time or space; for example the number of customers that arrive at a store during a 1 hour period or the number of defects found in 30 square meters of sheet metal.
	Models the distribution of a process that can though of as the product of a number of component processes – for example the rate on an investment, when interest is compounded, is the product of the returns for number of periods.
	Models the time between independent events, or a process time this memoryless – for example, the times between the arrivals from a large population of potential customers who act indepently of each other. The exponential is a highly variable distribution that is often overused because it leads to mathematically tractable models.
	An extremely flexible distribution used to model nonnegative random variables.
	An extremely flexible distribution used to model bounded random variables.
	Models the time to failure for components – for example the time to failure for a disk drive.
	Models complete uncertainty. All outcomes are equally likely. This distribution is often used inappropriately when there are no data.
	Models a process for which only the minimum, most likely, and the maximum values of the distribution are known. For example the minimum, most likely, and maximum time to test a product. It is a marked improvement over the uniform distribution.

¹ Taken from Discrete-Event System Simulation by Jerry Banks, John S. Carson, II, Barry L. Nelson, and David M. Nicol, 4th Edition, Prentice Hall International Series in Industrial and Systems Engineering, 2005.

Appendix 3-B

K-S Goodness of Fit Test

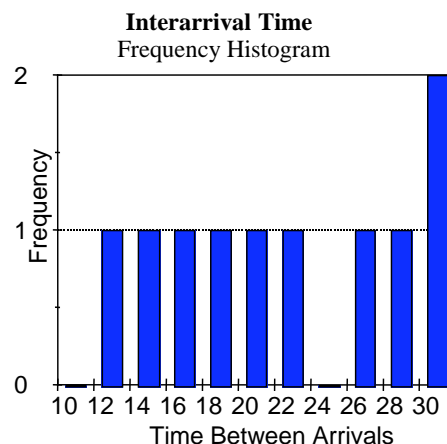
We often make several assumptions when we construct a model of a system. In queuing theory, for example, the basic single server model assumes service times and interarrival times are exponentially distributed. If the interarrival times of customers to a system are not exponentially distributed, the output of the basic single server model will probably not match the behavior of the real system. In other words, if the assumptions of the model are not valid, then the model is not valid. Therefore, we need a method to test the validity of our assumptions.

The Kolmogorov-Smirnov Goodness of Fit Test (KS Test) is used to compare a continuous theoretical cumulative distribution function (TCD) to an empirical (observed) cumulative distribution (OCD). **If the observed sample distribution is similar (as measured by a test statistic) to the theoretical distribution, then we can state with some level of significance that the population is distributed like the theoretical distribution.** In general, use the Chi-squared goodness of fit test on data from a discrete distribution (refer to any statistics book for a discussion of the Chi-squared test).

For example, let's record the interarrival times (time between arrivals) of students coming to class. The interarrival times of 10 students is shown in the table below.

Student	1	2	3	4	5	6	7	8	9	10
Inter-arrival time (sec)	21.7	11.8	29.9	16.0	13.8	27.8	17.7	28.6	19.5	25.2

The table provides no information about the shape of the probability distribution function, so let's put the data into a frequency histogram using Excel (see the figure below).



This figure contains a histogram of student interarrival times. A bin size interval of 2 seconds was used to create the histogram. The distribution looks very uniform. Remember, this is a small sample, so the absence of a point at the "10" and "24" spots and 2 observations at the "30" spot are not unusual. The sample distribution looks like a uniform distribution with 10 to 30 seconds between arrivals. Our task is to compare this sample distribution to the theoretical uniform probability distribution function. We consider this uniform example in this handout. The uniform example is used here because it is an easy problem for your first introduction to the K-S test.

Chapter 3

The K-S test compares the observed CDF to the theoretical CDF that we are hypothesizing. Statistically speaking, our null and alternative hypothesis is the following:

Ho: Sample comes from a Uniform (10, 30) distribution
Ha: Sample does not come from a Uniform (10, 30) distribution

The PDF of our hypothesized distribution is:

$$f(x) = \begin{cases} \frac{1}{20} & \text{for } 10 \leq x \leq 30 \text{ (seconds)} \\ 0 & \text{otherwise} \end{cases}$$

Recall from your probability and statistics course that the cumulative distribution function, $F(x)$, gives you $P(X \leq x)$; that is, the probability that you get an interarrival time of x seconds or less. This is the area under the PDF curve.

$$F(x) = P(X \leq x) \\ = \int_{-\infty}^x f(t) dt$$

For this example:

$$F(x) = \int_{10}^x \frac{1}{20} dt \\ F(x) = \frac{x}{20} - \frac{10}{20} \quad \text{for } 10 \leq x \leq 30$$

This is the Theoretical Cumulative Distribution (TCD) function for a uniform (10, 30) distribution. Substitute the interarrival times from the sample into this equation to determine the theoretical distribution value at each interarrival time t . For example, with the data for Student #2,

$$F(11.8) = \frac{11.8}{20} - \frac{10}{20} = .090$$

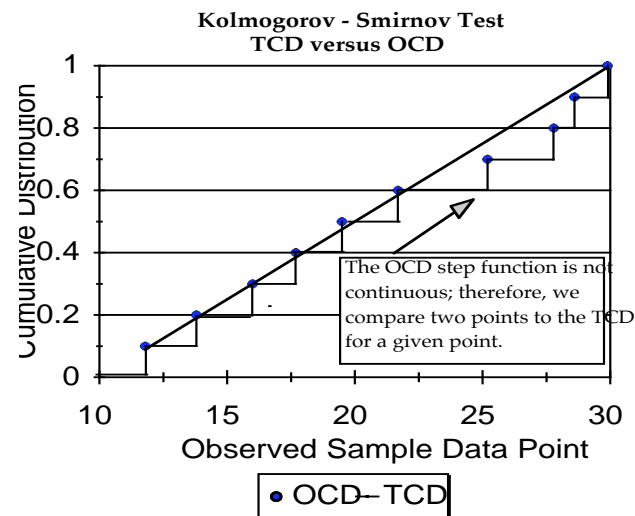
If the sample of 10 points is from a uniform population, we would expect to see the smallest value of our data set equate to 1/10th of the range of the theoretical cumulative distribution (TCD). In this case, the smallest value in the data, 11.8, results in a TCD value very close to .10. The greatest value in the data set should equate to a value 10/10ths of the range, or 1.0. In our sample, $F(29.9) = .995$.

The Observed Cumulative Distribution (OCD) is calculated in 2 steps. First, sort the data from the smallest to the largest observation. Second, determine the OCD by dividing the number (not the value) of the observation by the total number of points in the sample. For our example, the sorted values and the OCD are shown in the second two columns of the table below. Since there are 10 sample data points, the OCD will range from .10 to 1.0 with a step size of .10.

Sample Pt.	Sorted Data	OCD	TCD	Max Δ
21.7	11.8	0.1	0.090	0.090
11.8	13.8	0.2	0.190	0.090
29.9	16.0	0.3	0.300	0.100
16.0	17.7	0.4	0.385	0.085
13.8	19.5	0.5	0.475	0.075
27.8	21.7	0.6	0.585	0.085
17.7	25.2	0.7	0.760	0.160
28.6	27.8	0.8	0.890	0.190
19.5	28.6	0.9	0.930	0.130
25.2	29.9	1	0.995	0.095

The fourth column in the table above shows the TCD for all 10-sample points using equation (1.0).

The basic question is: "How similar is the Observed Cumulative Distribution (OCD) to the Theoretical Cumulative Distribution (TCD)?" One would be inclined to simply take the difference between the OCD and TCD rows in the table to find the greatest deviation. However, this is a test of a continuous distribution, not a discrete distribution.



The OCD is a step function with discrete values of 0, .1, .2, ..., 1.0. The TCD is a continuous function; therefore, we must compare the distance from the TCD to 2 points (steps) on the OCD. For example, compare the first TCD value (.090) to the 2 possible values for the step function at data point 11.8. The 2 values are 0 and .1 at the bottom and top of the step, respectively. The greatest distance between .090 and 0 versus .090 and .1 is .090. In mathematical notation, the operation is as follows:

$$K - S \text{ statistic} = \max_{\text{all } x_i} \left\{ |TCD(x_i) - OCD(x_i)|, |TCD(x_i) - OCD(x_{i-1})| \right\}$$

where x_1, x_2, \dots, x_n is the observed sample sorted into ascending order

In words, compare the absolute distance from the TCD at each x_i data point to the OCD (step function) at the previous point (x_{i-1}) and the current point (x_i). The maximum absolute differences for this example are shown in the last column in the table above.

Chapter 3

The greatest maximum difference is .190 and is used as the K - S statistic in the hypothesis test (see the shaded cell in the table).

Is .190 too great a difference between the TCD and the OCD? The OCD will never equal the TCD at all sample points, but if it is approximately the same, we can substantiate our model assumption concerning interarrival times. The K - S Test compares the largest deviation to a K - S critical value based on a chosen level of significance. The critical value is obtained from the K - S table in the appendix of your text by Cook and Russell. If the sample K - S test statistic (the maximum difference between the TCD and the OCD) is less than the K - S critical value, we fail to reject the null hypothesis with a given level of significance. Remember,

Ho: Sample comes from a Uniform (10, 30) distribution
Ha: Sample does not come from a Uniform (10, 30) distribution

Therefore, if we fail to reject Ho, we say the sample data is consistent with a population that is distributed uniformly from 10 to 30 seconds (with a level of significance associated with the K - S statistic). For example, the K - S critical value for 10 sample data points at a .05 level of significance is .410. Since .190 is less than .410, we fail to reject the null hypothesis.

In summary, there are 4 steps to take when testing the goodness of fit of sample data suspected to follow a continuous distribution. They are:

1. Collect Data.

- a. Beware of error and bias while collecting data.
- b. Decide on an appropriate sample size.

2. Construct a histogram.

- a. Choose several bin sizes. If the histogram shape differs greatly using different bin sizes, you should collect more data.
- b. Hypothesize a distribution based on the histogram's shape.

3. Use the K-S test to test the hypothesis the data comes from distribution X with parameters a, b, c, etc.

- a. Sort the data into increasing magnitudes (table, column 2).
- b. Form the observed cumulative distribution function, OCD (table, column 3).
- c. Form the hypothesized cumulative distribution function (the theoretical cumulative distribution, TCD in column 4).
- d. Determine the largest vertical distance between the OCD and the TCD (this is the largest value in column 5).
- e. Compare the maximum vertical distance with the critical value from the K - S Test Critical Values table (see Appendix A).

4. Conclude whether the hypothesis can be rejected based on step (3e) above. Reject Ho if the K - S statistic \leq K - S critical value.

The Kolmogorov-Smirnov Goodness of Fit Test is an excellent tool to use to test the validity of assumptions in a basic queuing model. In the basic single server model, service and interarrival times are assumed to be exponential. If you take data on service and arrival times, say during the queuing design exercise in this course, you can test a hypothesis that the observed service and arrival of customers do indeed follow an exponential distribution. The test can also be used to show why the basic queuing models may not be useful for your system. If a K - S test shows sample interarrival

Chapter 3

times to be consistent with a uniform (not exponential) distribution, the test helps explain why model output is not similar to the actual system.

Table 3-B1 One Tailed Table of Critical Values for the Kolmogorov-Smirnov (KS) Test

Values of $d^*(N)$ such that $\Pr[\max|S_N(x) - F_0(x)| > d^*(N)] = \alpha$, where $F_0(x)$ is the theoretical cumulative distribution and $S_N(x)$ is an observed cumulative distribution for a sample of N .

Sample Size (n)	Level of Significance (α)				
	0.20	0.15	0.10	0.05	0.01
1	0.900	0.925	0.950	0.975	0.995
2	0.664	0.726	0.776	0.842	0.929
3	0.565	0.597	0.642	0.708	0.828
4	0.494	0.525	0.564	0.624	0.733
5	0.446	0.474	0.510	0.565	0.669
6	0.410	0.436	0.470	0.521	0.618
7	0.381	0.405	0.438	0.486	0.577
8	0.358	0.381	0.411	0.457	0.543
9	0.339	0.360	0.388	0.432	0.514
10	0.322	0.342	0.368	0.410	0.490
11	0.307	0.326	0.352	0.391	0.468
12	0.295	0.313	0.338	0.375	0.450
13	0.284	0.302	0.325	0.361	0.433
14	0.274	0.292	0.314	0.349	0.418
15	0.266	0.283	0.304	0.338	0.404
16	0.258	0.274	0.295	0.328	0.392
17	0.250	0.266	0.286	0.318	0.381
18	0.244	0.259	0.278	0.309	0.371
19	0.237	0.252	0.272	0.301	0.363
20	0.231	0.246	0.264	0.294	0.356
25	0.21	0.22	0.24	0.27	0.32
30	0.19	0.20	0.22	0.24	0.29
35	0.18	0.19	0.21	0.23	0.27
Over 35	$\frac{1.07}{\sqrt{N}}$	$\frac{1.14}{\sqrt{N}}$	$\frac{1.22}{\sqrt{N}}$	$\frac{1.35}{\sqrt{N}}$	$\frac{1.63}{\sqrt{N}}$

Example 3-22. K-S Test for Normal Distribution

The data below are a sample of 15 interarrival times. Use the K-S test to determine whether the interarrival times are normally distributed with a mean of 1.96 and a standard deviation of .73.

The first step is to formulate your hypothesis test. Remember, in hypothesis testing, H_0 is what you are testing for, H_1 is the opposite of what you are testing. So our Hypothesis test is as follows:

H_0 = Data is from a distribution that is $N(1.96, .73)$

H_1 = Data is not from a distribution that is $N(1.96, .73)$

(Keep in mind that $N(1.96, .73)$ is standard notation for a normal distribution with a mean of 1.96 and a standard deviation of .73)

Now that the test is set up, time to start filling out the table. The first two columns you are given. The third is just the data that was collected sorted in ascending order. The third column is the observed Cumulative Distribution Function (CDF). We get this from the following equation:

$$\text{OCD}(t) = (\# \text{ of data point } t) / (\# \text{ of points observed})$$

In this case, a sample calculation for the first point is: $\text{OCD}(1) = 1/15 = .067$ *(three digits is sufficient)*

The Z value column is necessary if you are calculating this table by hand. Since you will probably see this on a WPR, it would be good for you to do it this way at least once. Use the equation for converting any normal distribution to the Standard Normal distribution:

$$z = \frac{x - \mu}{\sigma}$$

Where z is the value we are looking for, x is the interarrival time, μ is the mean of the distribution, and σ is the standard deviation of the distribution.

A sample calculation for the first term is:

$$z(\text{1st point}) = \frac{x - \mu}{\sigma} = \frac{.84 - 1.96}{.73} = \frac{-1.12}{.73} = -1.534$$

To get the value for the next column, you either have to use your calculator and get the CDF for each Z value, or look it up in the Normal table in your textbook on page 873. Using your text, the CDF value for the normal distribution having a z value of -1.53 is found this way:

The CDF for 1.53 is .9370 therefore the value for -1.53 = $1 - .9370 = .063$

(we had to solve for the negative Z value this way because the table in the book does not have values for z terms less than 0)

The final column is determined by taking the maximum of two values. As indicated in your supplemental reading, we must do this because we are comparing a point distribution against a continuous distribution. *(Go back to the supplemental reading starting with the last paragraph on page 3 and read through the second paragraph on page 4).* The equation is as follows:

$$\text{MAX}\{|OCD_{t-1} - TCD_t|, |OCD_t - TCD_t|\}$$

A sample calculation done for the first term follows:

$$\text{MAX}\{|0 - .063|, |.067 - .063|\} = \text{MAX}\{.063, .004\} = .063$$

Because the first term has no observed values before it, $OCD_{t-1} = 0$, this is only true for the first term.

The completed table follows:

Point	Inter-arrival Time	Sorted Value	Observed CDF	Z Value	Theoretical CDF	Max Absolute Difference
1	0.84	0.84	0.067	-1.534	0.063	0.063
2	1.69	1.18	0.133	-1.068	0.143	0.076
3	3.36	1.23	0.2	-1	0.159	0.041
4	1.84	1.27	0.267	-0.945	0.172	0.095
5	2.15	1.67	0.333	-0.397	0.346	0.079
6	1.18	1.69	0.4	-0.37	0.356	0.044
7	1.67	1.84	0.467	-0.164	0.435	0.035
8	2.37	1.84	0.533	-0.164	0.435	0.098
9	1.27	1.89	0.6	-0.096	0.462	0.138
10	2.7	2.09	0.667	0.178	0.571	0.096
11	2.09	2.15	0.733	0.26	0.603	0.13
12	3.23	2.37	0.8	0.562	0.713	0.087
13	1.84	2.7	0.867	1.014	0.845	0.045
14	1.23	3.23	0.933	1.74	0.959	0.092
15	1.89	3.36	1	1.918	0.972	0.039

The maximum difference is the term we are looking for. In this problem it is the difference associated with the 9th sorted point. Absolute difference is .138. This is the statistic we will use to compare against our K-S test statistic. At this point you must determine what level of significance you wish to prove your H_0 hypothesis. We will select an $\alpha = .05$. Looking at the tables provided for the K-S test, we go down the left most column until we find the row that corresponds to a sample size equal to 15. We read across the top row until we find the column that corresponds to a level of significance $\alpha = .05$. At the intersection of that row and column we get the K-S statistic 0.338. Because our max value from our calculated table above (.138) is less than the K-S table value (.338), we can say we **fail to reject the null hypothesis**. In fact, since we fail to reject the null hypothesis at all levels of α , we have a strong argument that the distribution is in fact $N(1.96, .73)$.

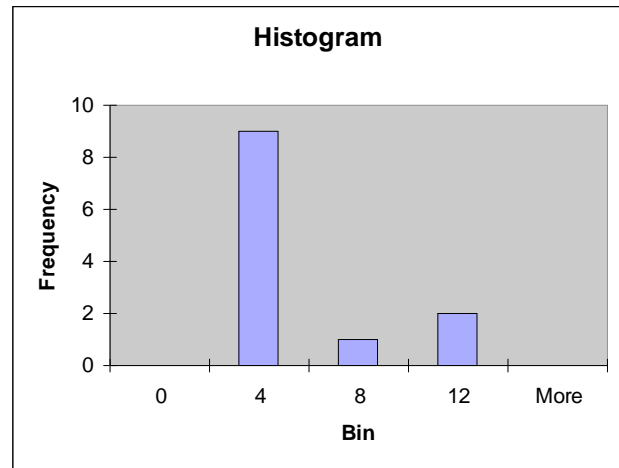
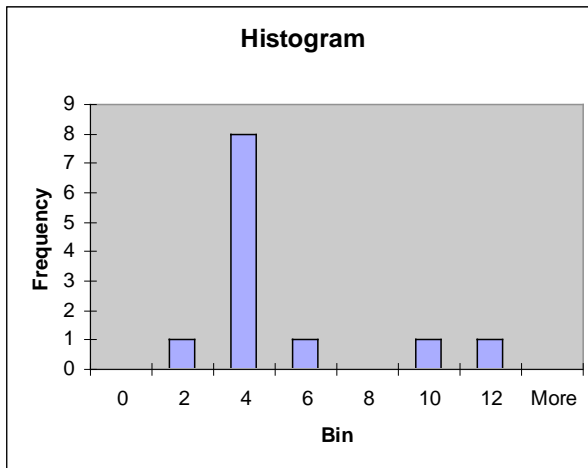
Example Problem 3 -23.

The data below shows the service times for 12 trucks at a warehousing facility.

Customer	1	2	3	4	5	6
Time (min)	1.85	10.04	3.57	3.79	9.10	2.31
Customer	7	8	9	10	11	12
Time (min)	3.67	4.89	2.56	2.31	2.40	2.61

REQUIREMENT: Use the 4 step process discussed in this handout to hypothesize a distribution for the data in the table. The table provided is **Step 1 (Collect Data)**.

2. Construct a histogram. Below are two histograms with two different bin widths.



3. Use the K-S test to test the hypothesis the data comes from distribution X with parameters a, b, c, etc.

Based upon these distributions, we hypothesis test that:

H_0 = Data is from a distribution that is exponential with a mean of 1/4.1

H_1 = Data is not from a distribution that is exponential with a mean of 1/4.1

The CDF of an exponential distribution is:

$$F(x; \lambda) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\beta x} & x \geq 0 \end{cases}$$

The mean of an exponential distribution is $\beta = 1/\lambda$

Point	Time	Sorted	OCD	TCD	Difference
1	1.85	1.85	0.083	0.363	0.363
2	10.04	2.31	0.167	0.431	0.347
3	3.57	2.31	0.250	0.431	0.264
4	3.79	2.4	0.333	0.443	0.193
5	9.1	2.56	0.417	0.464	0.131
6	2.31	2.61	0.500	0.471	0.054
7	3.67	3.57	0.583	0.581	0.081
8	4.89	3.67	0.667	0.591	0.075
9	2.56	3.79	0.750	0.603	0.147
10	2.31	4.89	0.833	0.697	0.137
11	2.4	9.1	0.917	0.891	0.058
12	2.61	10.04	1.000	0.914	0.086

4. Conclude whether the hypothesis can be rejected based on step (3e) above.

Reject H_0 if the K - S statistic \leq K - S critical value.

The maximum difference is the term we are looking for. In this problem it is the difference associated with the 1st sorted point. Absolute difference is .363. This is the statistic we will use to compare against our K-S test statistic. At this point you must determine what level of significance you wish to prove your H_0 hypothesis. We will select an $\alpha = .05$. Looking at the tables provided for the K-S test, we go down the left most column until we find the row that corresponds to a sample size equal to 12. We read across the top row until we find the column that corresponds to a level of significance $\alpha = .05$. At the intersection of that row and column we get the K-S statistic 0.375. Because our max value from our calculated table above (.363) is less than the K-S table value (.375), we can say we **fail to reject the null hypothesis**. In fact, since we fail to reject the null hypothesis at all levels of α , we have a strong argument that the distribution is in fact exponential with a mean of 4.1.

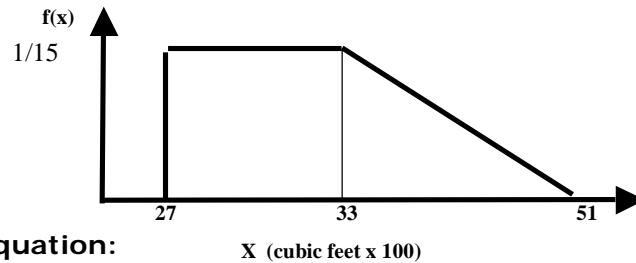
Appendix 3-C

Inverse Transformation Examples for Combination Functions

Example 3-23.

Using the Inverse Transformation Method, determine the continuous process generators that will generate random variables with the following triangle distribution:

The sketch of this pdf:



Determine the pdf equation:

$$\begin{aligned} \text{Uniform: } f(x) &= \frac{1}{15} & \Rightarrow \frac{1}{15} &, \text{ for } 27 \leq x \leq 33 \\ \text{Ramp: } f(x) &\Rightarrow (y - 0) = \frac{-1/15}{18}(x - 51) \Rightarrow \frac{-x}{270} + \frac{51}{270}, & \text{ for } 33 \leq x \leq 51 \\ &\Rightarrow 0 & , \text{ otherwise} \end{aligned}$$

Develop a process generator for these interarrival times using the inverse transformation method.

$$\begin{aligned} F(t) &= \int f(t) dt = \int_{27}^x \frac{1}{15} dt & F(t) &= .40 + \int_{33}^x f(t) dt = .40 + \int_{33}^x \left(\frac{-x}{270} + \frac{51}{270} \right) dt \\ \Rightarrow \left[\frac{t}{15} \right]_{27}^x &\Rightarrow \left[\frac{x}{15} - \frac{27}{15} \right] & \Rightarrow .40 + \left[\frac{-t^2}{540} + \frac{51t}{270} \right]_{33}^x \\ \Rightarrow \frac{x}{15} - \frac{27}{15} &= r & \Rightarrow .40 + \left[\left(\frac{-x^2}{540} + \frac{51x}{270} \right) - \left(\frac{-33^2}{540} + \frac{51(33)}{270} \right) \right] \\ \Rightarrow x - 27 &= 15r & \Rightarrow \\ \Rightarrow x &= 15r + 27 & \Rightarrow x^2 - 101.887x = \frac{-r}{.001855} - 2057.682 \\ & & \text{Complete the Square} \\ \Rightarrow x^2 - 101.887x + (50.944)^2 &= \frac{-r}{.001855} - 2057.682 + 2595.291 \\ \Rightarrow (x - 50.944)^2 &= \frac{-r}{.001855} + 537.609 \\ \Rightarrow x &= -\sqrt{-539.08r + 537.609} + 50.944 \\ & \text{Adjust for Rounding} \\ \Rightarrow x &= 15r + 27 \end{aligned}$$

$$x = -\sqrt{-540r + 540} + 51$$

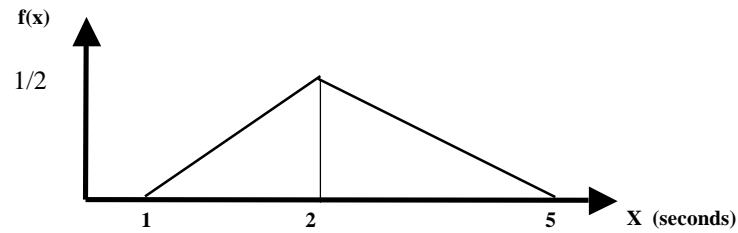
VERIFY:	r = 0 x = 27 ✓	r = .40 x = 33 ✓
	r = .40 x = 33 ✓	r = 1 x = 51 ✓

Chapter 3

Example 3-24.

Using the Inverse Transformation Method, determine the continuous process generators that will generate random variables with the following triangle distribution:

The sketch of this pdf:



Determine the pdf equation:

$$\begin{aligned} \text{Ramp: } f(x) &\Rightarrow (y - y_i) = m(x - x_i) \\ \Rightarrow (y - 0) &= +\frac{1}{2}(x - 1) \Rightarrow \frac{x}{2} - \frac{1}{2}, \text{ for } 1 \leq x \leq 2 \\ \Rightarrow (y - 0) &= -\frac{1}{6}(x - 5) \Rightarrow -\frac{x}{6} + \frac{5}{6}, \text{ for } 2 \leq x \leq 5 \\ &\Rightarrow 0, \text{ otherwise} \end{aligned}$$

Develop a process generator for these interarrival times using the inverse transformation method.

$$\begin{aligned} F(x) &= \int f(t) dt = \int_1^x \left(\frac{t}{2} - \frac{1}{2} \right) dt & F(x) &= \frac{1}{4} + \int_2^x f(t) dt = \frac{1}{4} + \int_2^x \left(-\frac{t}{6} + \frac{5}{6} \right) dt \\ \Rightarrow \left[\frac{t^2}{4} - \frac{t}{2} \right]_1^x & & \Rightarrow \frac{1}{4} - \left[-\frac{t^2}{12} \right]_2^x + \left[\frac{5t}{6} \right]_2^x \\ \Rightarrow \left[\frac{x^2}{4} - \frac{1}{4} \right] - \left[\frac{1}{2} - \frac{1}{2} \right] & & \Rightarrow \frac{1}{2} - \left[\frac{x^2}{12} - \frac{4}{12} \right] + \left[\frac{5x}{6} - \frac{10}{6} \right] \\ \Rightarrow \frac{x^2}{4} - \frac{x}{2} + \frac{1}{4} &= r & \Rightarrow -\frac{x^2}{12} + \frac{5x}{6} - \frac{13}{12} &= r \\ \Rightarrow x^2 - 2x &= 4r - 1 & \Rightarrow x^2 - 10x &= -12r - 13 \\ \text{Complete the Square} & & \text{Complete the Square} & \\ \Rightarrow x^2 - 2x + 1 &= 4r - 1 + 1 & \Rightarrow x^2 - 10x + 25 &= -12r - 13 + 25 \\ \Rightarrow (x - 1)^2 &= 4r & \Rightarrow (x - 5)^2 &= 12 - 12r \\ \Rightarrow x &= 1 + \sqrt{4r} & \Rightarrow x &= 5 - \sqrt{12 - 12r} \end{aligned}$$

VERIFY:	$r = 0$	$x = 1 \checkmark$	$r = 1/4$	$x = 2 \checkmark$
	$r = 1/4$	$x = 5 \checkmark$	$r = 1$	$x = 1 \checkmark$

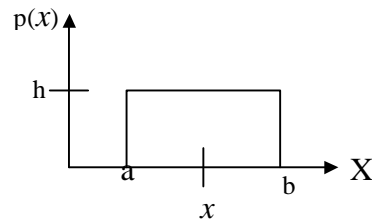
Appendix 3-D

Continuous Process Generator Formulas

Terms: X = The random variable being generated
 r = A Uniform (0,1) random number
 λ = The arrival or service RATE associated with an Exponential random variable. Remember: a RATE is units/time period such as cars/hr., customers/min., etc.

UNIFORM [a,b]

$$x = a + r(b - a)$$



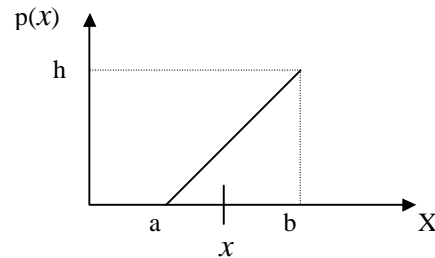
EXPONENTIAL [λ]

$$x = \frac{-\ln(1-r)}{\lambda} \quad \text{or} \quad X = \frac{-\ln(r)}{\lambda}$$

UPWARD RAMP [a,b]

$$h = \frac{2}{b-a}$$

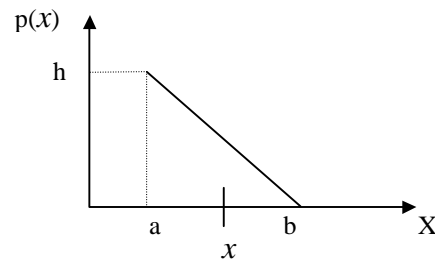
$$x = a + \sqrt{\frac{2r(b-a)}{h}} \quad \text{for } 0 \leq r \leq 1$$



DOWNWARD RAMP [a,b]

$$h = \frac{2}{b-a}$$

$$x = b - \sqrt{(b-a)^2(1-r)} \quad \text{for } 0 \leq r \leq 1$$



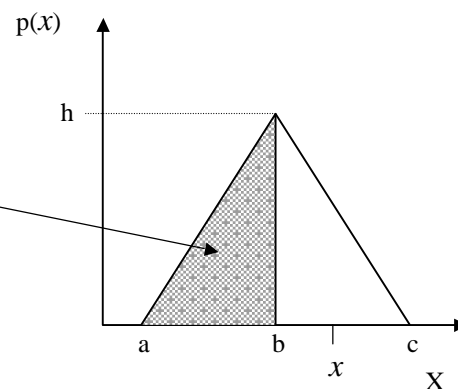
TRIANGLE [a,b,c]

$$h = \frac{2}{c - a}$$

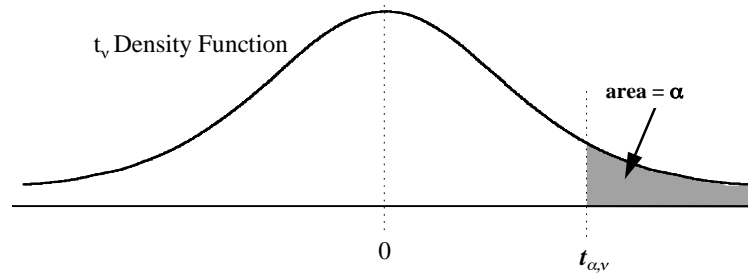
$$\mathbf{AREA} = \text{Area of left triangle} = \frac{h(b - a)}{2}$$

$$x = a + \sqrt{r(b - a)(c - a)} \quad \text{for } 0 \leq r \leq \mathbf{AREA}$$

$$x = c - \sqrt{(1 - r)(c - a)(c - b)} \quad \text{for } \mathbf{AREA} \leq r \leq 1$$



Appendix 3-E t-Distribution Table



ν	.10	.05	.025	α .01	.005	.001	.0005
1	3.078	6.314	12.706	31.821	63.657	318.310	636.620
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Bibliography

Benjamin, J. R. and Cornell, C. A., *Probability, Statistics, and Decision for Civil Engineers*, McGraw-Hill, Inc., New York, NY, 1970.

Cheremisinoff, N. P., *Practical Statistics for Engineers and Scientists*, Technomic Publishing Co., Inc., Lancaster, PA, 1987.

Freund, J. E., *Mathematical Statistics*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1962.

Devore, J. L., *Probability and Statistics for Engineering and the Sciences*, 3rd Edition, Brooks/Cole Publishing, Pacific Grove, CA, 1991.

Harr, M. E., *Reliability-Based Design in Civil Engineering*, McGraw Hill, Inc., New York, NY, 1987.

Hoel, P. G., *Introduction the Mathematical Statistics*, 3rd Edition, John Wiley and Sons, Inc., New York, NY, 1962.

Hogg, R. V. and Craig, A. T., *Introduction to Mathematical Statistics*, 4th Edition, Macmillan Publishing Co., Inc., New York, NY, 1978.

Mann, P. S., *Introductory Statistics*, 2nd Edition, John Wiley and Sons, Inc., New York, NY, 1995.

Matloff, N. S., *Probability Modeling and Computer Simulation*, PWS-Kent Publishing Co., Boston, MA, 1988.

Neter, J., Wasserman, W., and Kutner, M. H., *Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs*, 3rd Edition, Irwin Publishing, Homewood, IL, 1990

Ostle, B., *Statistics in Research*, 2nd Edition, The Iowa State University Press, Ames, IA, 1964.

Ross, S., *A First Course in Probability*, 3rd Edition, Macmillan Publishing Co., Inc., New York, NY, 1988.

Scheaffer, R. L., and McClave J. T., *Probability and Statistics for Engineers*, 3rd Edition, PWS-Kent Publishing Company, Boston, MA, 1990.

Sen, A., and Srivastave, M., *Regression Analysis, Theory, Methods, and Applications*, Springer-Verlag, New York, NY, 1990.

Siegel, A. F., *Statistics and Data Analysis: An Introduction*, John Wiley and Sons, Inc., New York, NY, 1988.

Discussion Question 3-1

Simulations are essential models for processes with a mathematical representation of the entity behavior. Which is more difficult to model: the process or the entity behavior?

Class Problem 3-1

As part of a business process reengineering study, you collected the following times that calls were received at your information technology (IT) support desk. You would like to build a simulation to investigate bottlenecks and the return on investment of new technology. Your IT department provided you with following interarrival times in minutes. Using Crystal Ball fit a continuous process distribution to the data for us in a simulation.

Interarrival Times (mins)		
	1.108568755	0.732264692
0.588814749	0.547610383	0.242657492
0.397494588	1.538859923	0.352512396
1.673683718	2.719185953	0.112583034
0.084404664	3.242295179	0.115828188
0.230179077	0.118464846	3.459217465
0.587504092	0.583542146	0.711006448
1.37812423	0.389833717	0.610918746
0.327181262	0.746621891	0.478771272
0.535358812	0.658668746	0.291238526
1.064741685	0.03681801	0.018324774
2.761979873	0.187436268	0.621802042
0.56351202	0.382481155	0.817091787
3.29436124	0.990546057	0.290123951
0.366030742	0.771284964	0.335684761
0.073726122	1.694452517	3.786707383
1.780062427	0.562536755	0.46974402
0.544917667	0.039699342	0.926288378
0.182478972	0.669752309	0.189516512
3.550765671	1.590598732	0.737572538
0.592049529	0.162172788	0.662148766
1.932861976	0.109502818	0.835056714
0.331807512	0.874628264	2.907609454
0.39967384	0.25855252	1.18275267
0.523997406	0.330339127	0.067442005
1.413733145	0.069066114	0.403250281
0.492974166	0.365648322	0.043291821
0.909592149	3.288387525	1.49132871
0.471302103	2.225006862	1.309829203
0.722365606	0.632092026	0.22914525
1.887436095	1.475870936	0.891470352
0.197749859	1.172587523	0.576193114
1.82730058	3.223637925	0.867377053
	1.374153606	2.127784231

b. Using the K-S statistics from Crystal Ball, how confident can we be that this data behaves according to the hypothesized distribution?

c. Using histograms² develop a discrete process generator representative of this data.

² Typically, we use the $n^{1/2}$ rule for determining bin sizes for histograms – see page 12 of course text for details.

Chapter 3

d. Under what circumstances is using a discrete process generator preferable to a continuous process generator? Give an example.

Class Problem 3-2

Steven's students drink a lot of coffee at the recently remodeled Café Java. The manager has gathered data on arrival, coffee demand, and the service time using a single cashier. The relative frequency for the demand of coffee size for the 8:00 to 9:00 AM hours (peak consumption period) of operation is shown in the table below. Develop a discrete process generator that you can use in simulating operations:

Demand	Frequency	P(x)	Cum Prob	Random Number Range
0	8			
Small	10			
Medium	22			
Large	10			

Class Problem 3-3

The restaurant manager contacted one of their best coffee customers, Dr. Farr, to develop the other data needed for the simulation. Based upon his extensive experience as a patron, he determined that the interarrival time of students on average is 2 minutes (closely approximating a Poisson process). Also the service times were uniformly distributed with a minimum time of 10 seconds and a maximum time of 55 seconds. Develop the continuous process generators needed for the simulation.

- a. Arrival Time Process Generator
- b. Service Time Process Generator

Class Problem 3-4

Use the following random number streams:

Stream 1) .11, .36, .45, .98, .78, .67, .51, .23, .15, .89

Stream 2) .91, .39, .72, .14, .42, .78, .55, .38, .60, .71

The following information from the Petroco Service Station is given:

Time Between Arrivals	Probability
1	.15
2	.3
3	.4
4	.15
Total	1.0

Chapter 3

We can now start the Monte Carlo simulation process. First, we must construct a table and determine the random number ranges.

Time Between Arrivals	Probability f(x)	Cummulative Probability F(x)	Random Number Ranges

a. Now we must simulate the arrival of cars using the first stream of random numbers.

Arrival	Random Number	Time between Arrivals
1	.11	
2	.36	
3	.45	
4	.98	
5	.78	
6	.67	
7	.51	
8	.23	
9	.15	
10	.89	
Average		

b. Next we simulate the arrival of cars using the second stream of random numbers.

Arrival	Random Number	Time between Arrivals
1	.91	
2	.39	
3	.72	
4	.14	
5	.42	
6	.78	
7	.55	
8	.38	
9	.60	
10	.01	
Average		

c. Is there a difference in the results? If yes, then why?

Class Problem 3-5

Consider a continuous random variable with the following pdf that represents service time:

$$f(x) = \begin{cases} \frac{1}{3}, & \text{for } 2 \leq x \leq 5 \\ 0 & \text{otherwise} \end{cases}$$

- a) Sketch the pdf,
- b) Develop a process generator for these service times using the inverse transformation method.

Class Problem 3-6

Consider a continuous random variable with the following pdf that represents service time:

$$f(x) = \begin{cases} 2 - \frac{x}{2}, & \text{for } 2 \leq x \leq 4 \\ 0 & \text{otherwise} \end{cases}$$

- a) Sketch the pdf,
- b) Develop a process generator for these service times using the inverse transformation method.

Class Problem 3-7

The COTS PM is also considering a refinement to the simulation model to study another possible change in inventory policy of the active sonar geophones. The number of days the Navy has to wait from the time they place the order until the time the geophones arrive at the depot is not constant. The lead-time (a random variable X) is described by the following probability:

$$f(x) = \begin{cases} 8x - 4 & \frac{1}{2} \leq x \leq 1 \text{ weeks} \\ 0 & \text{otherwise} \end{cases}$$

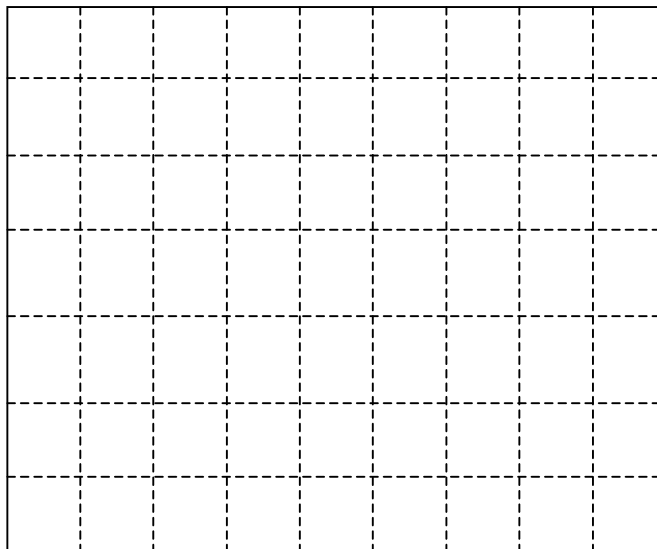
- Graph the function
- Use the inverse transformation method to derive the continuous process generator for the lead time to receive the geophones.

Class Problem 3-8

Given the continuous probability distribution below, develop a process generator by using the inverse transformation technique.

$$f(x) = \begin{cases} \frac{x}{18}, 0 \leq x \leq 6 \\ 0, \text{otherwise} \end{cases}$$

- Graph the function



- Use the inverse transformation method to derive the continuous process generator.

Class Problem 3-9

To demonstrate the application of the exponential CPG, let's take a situation from queuing theory and simulate the arrival rate of customers entering a system. Assume customers arrive at the 7 Eleven for Squishies on average every 4 minutes (following an exponential distribution). Simulate this process by using the Exponential CPG. Using

Chapter 3

Excel® develop a histogram using 50 data points and verify that you CPG is indeed replicating the exponential behavior.

Class Problem 3-10

The data below are a sample of 15 interarrival times.

- a. Use the K-S test to determine whether the interarrival times are exponentially distributed with a mean arrival rate equal to .51 customers per minute.

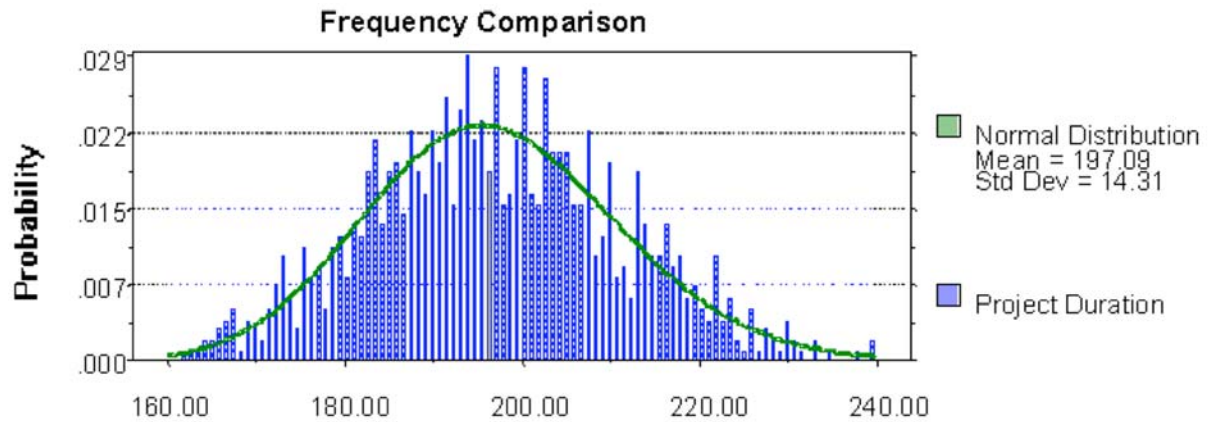
Point	Interarrival Time	Sorted Value	Observed CDF	Theoretical CDF	Max Absolute Difference
1	0.84	0.84			
2	1.69	1.18			
3	3.36	1.23			
4	1.84	1.27			
5	2.15	1.67			
6	1.18	1.69			
7	1.67	1.84			
8	2.37	1.84			
9	1.27	1.89			
10	2.7	2.09			
11	2.09	2.15			
12	3.23	2.37			
13	1.84	2.7			
14	1.23	3.23			
15	1.89	3.36			

- b. Use the K-S test to determine whether the interarrival times are uniformly distributed from 0 to 3.5 minutes per customer.

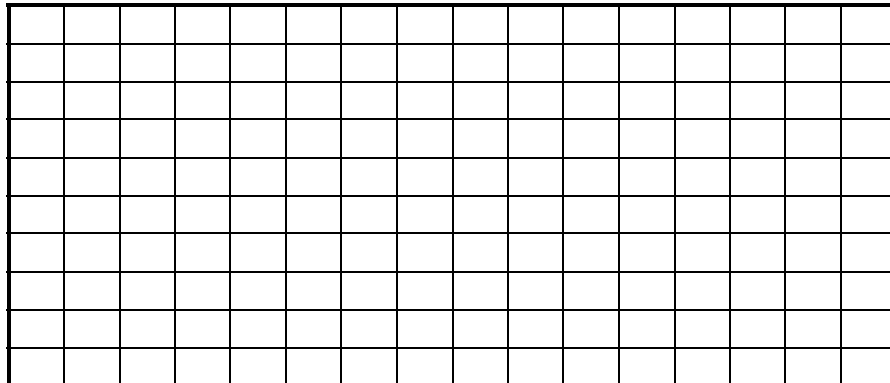
Point	Interarrival Time	Sorted Value	Observed CDF	Theoretical CDF	Max Absolute Difference
1	0.84	0.84			
2	1.69	1.18			
3	3.36	1.23			
4	1.84	1.27			
5	2.15	1.67			
6	1.18	1.69			
7	1.67	1.84			
8	2.37	1.84			
9	1.27	1.89			
10	2.7	2.09			
11	2.09	2.15			
12	3.23	2.37			
13	1.84	2.7			
14	1.23	3.23			
15	1.89	3.36			

Class Problem 3-11

You run a Crystal Ball simulation of project completion time as shown below:



- You company bid this contract on completing the contract within 185 days. What is the probability of completing this project in less than 185 days (show your work)?
- You would like to develop a chart for the Vice President of Engineering to justify additional resources. On the graph below, develop any chart that would be of interest using the information provided and explain the chart in simple terms.

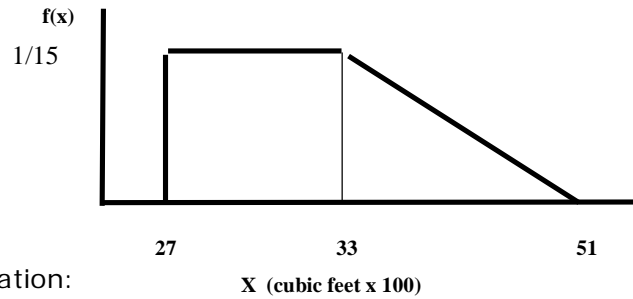


Explain this chart in simple terms:

Class Problem 3-12

Using the Inverse Transformation Method, determine the continuous process generators that will generate random variables with the following triangle distribution:

The sketch of this pdf:

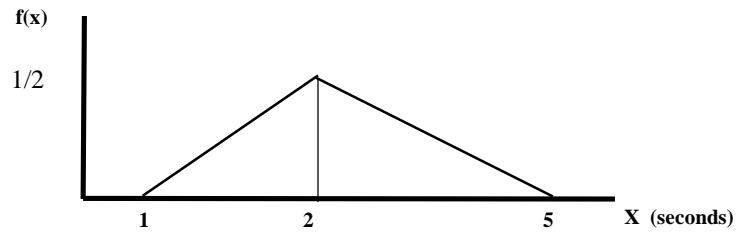


Determine the pdf equation:

Class Problem 3-12

Using the Inverse Transformation Method, determine the continuous process generators that will generate random variables with the following triangle distribution:

The sketch of this pdf:



Determine the pdf equation: