

Probability Distribution Types

Prepared by Dr. Sami Alzahrani

PMG621
Risk and Decision Analysis

Lecture 10



Probability Distribution Types

Lecture 10

PMG621

Risk and Decision Analysis

Chapter
10

Introduction

Probability is the language of uncertainty. A common anxiety among people learning decision analysis is choosing the type of probability distribution to represent an uncertainty. Here we examine the most popular distributions and situations where and why they apply.

Chapter
10

Introduction

Representing judgments about uncertainty is key to using stochastic (probabilistic) project models. Usually, the most qualified people available are asked to provide their opinions about values that go into the model. When the expert does not know a significant parameter with confidence, a probability distribution can completely express his judgment about uncertainty.

Here we examine the most popular distributions and situations where and why they apply.

Chapter

10

In the example figures, the numbers in parentheses are the "arguments" of the distribution function. These are typically the way we specify the respective distributions to a **Monte Carlo simulation** (simulation) computer tool. In this chapter, I'll use *mean* Q instead of expected value (EV), though they are synonymous. Some distributions also use the *standard deviation* (**0**) statistic as an argument. We will use p to represent probability of success.

Chapter
10

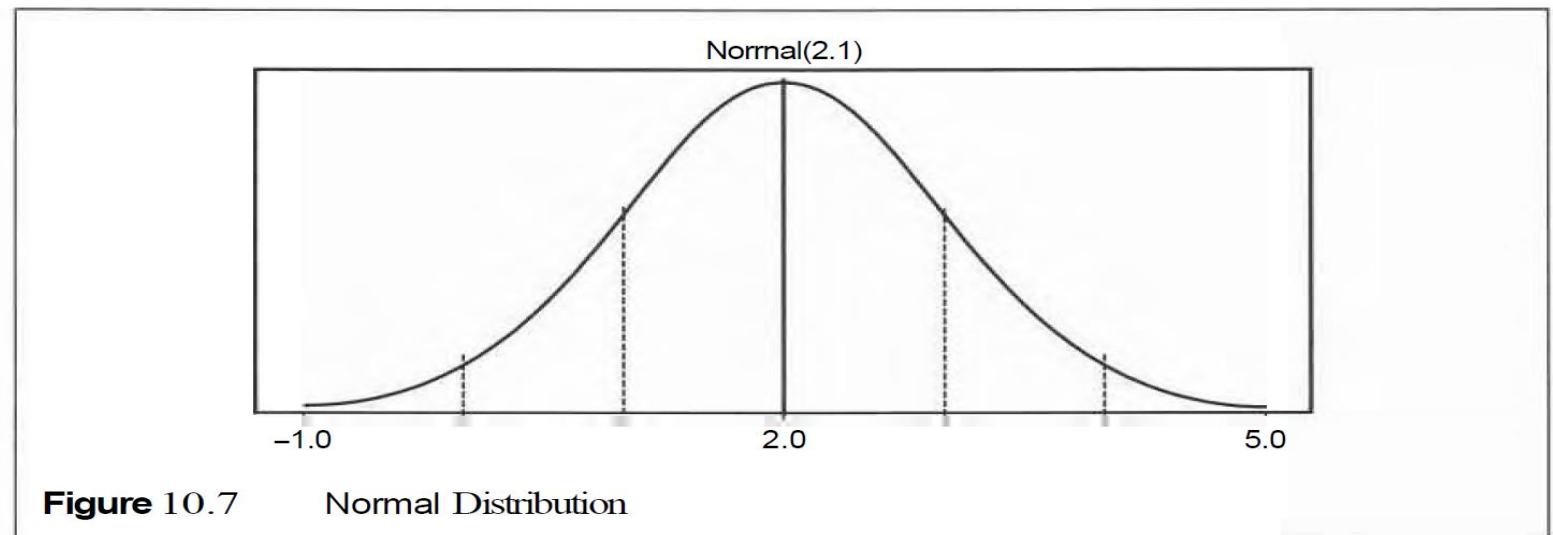
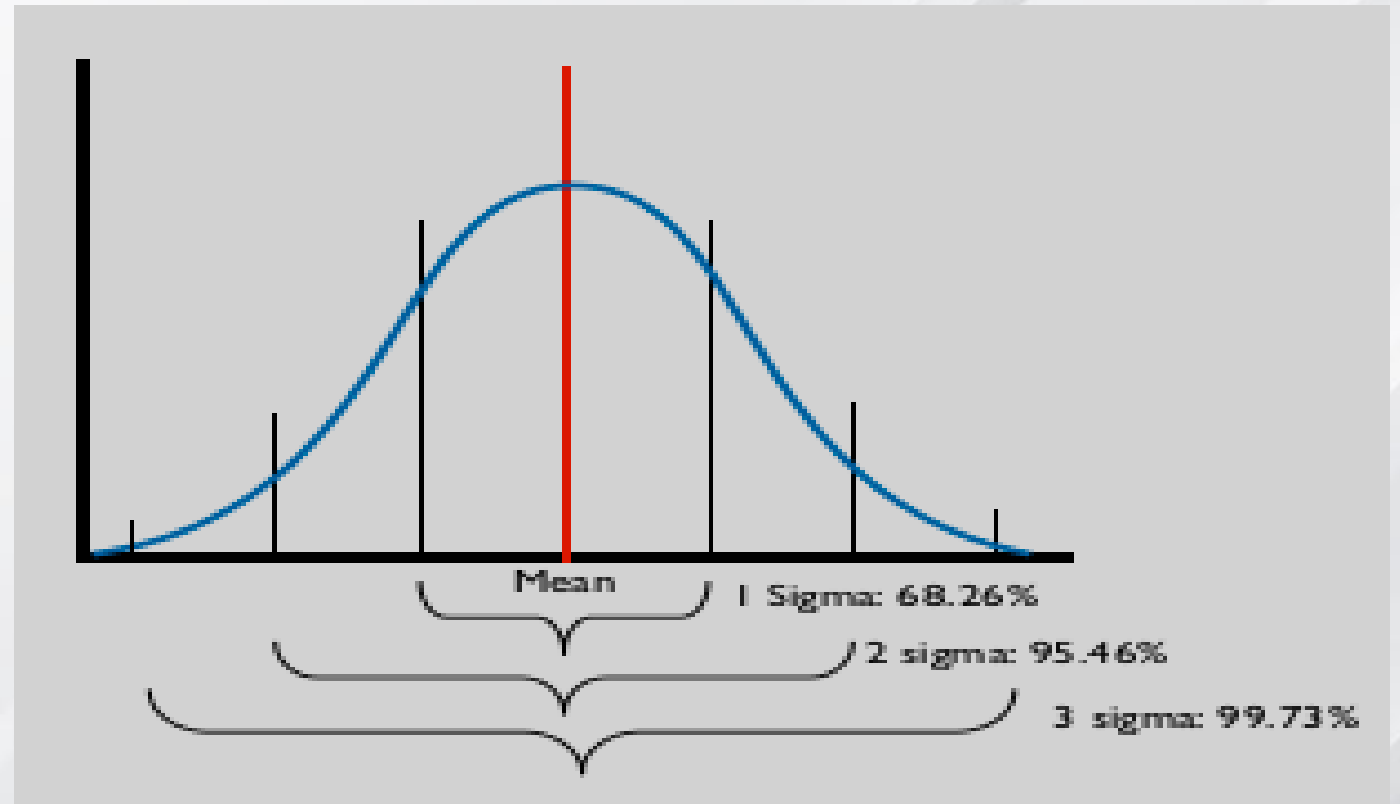
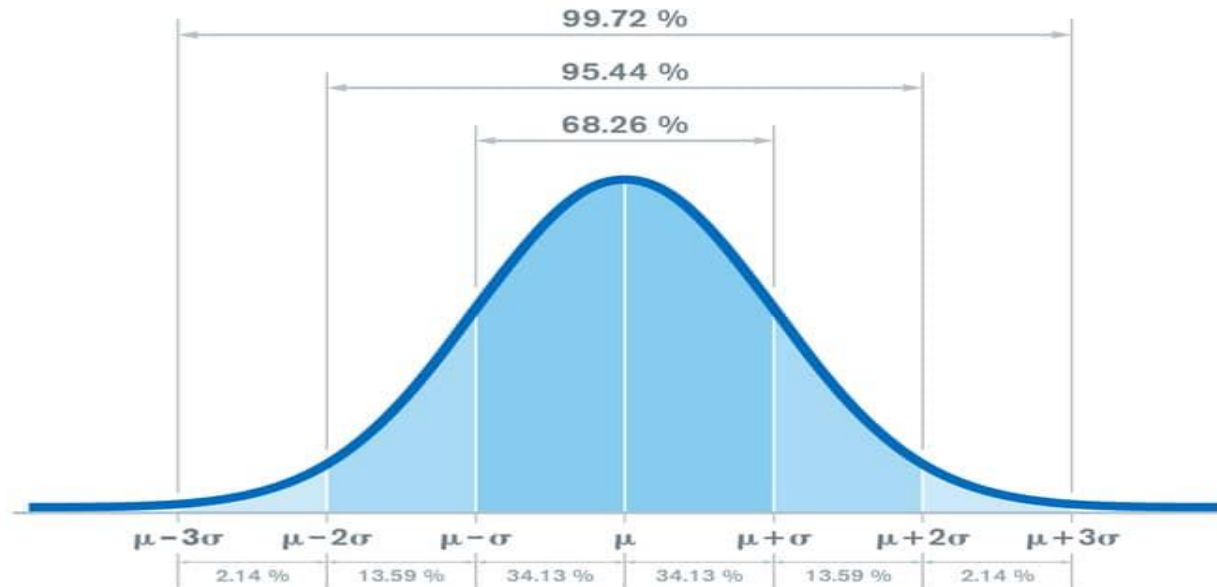


Figure 10.7 Normal Distribution

Chapter
10



Normal Distribution



Mean, Median, Mode & Rang

Mean

Add all the numbers then divide by the amount of numbers

9, 3, 1, 8, 3, 6

$$9 + 3 + 1 + 8 + 3 + 6 = 30$$

$$30 \div 6 = 5$$

The mean is 5

Median

Order the set of numbers, the median is the middle number

9, 3, 1, 8, 3, 6

1, 3, 3, 6, 8, 9

The median is 4.5

Mode

The most common number

9, 3, 1, 8, 3, 6

The mode is 3

Range

The difference between the highest number and lowest number

9, 3, 1, 8, 3, 6

$$9 - 1 = 8$$

The range is 8

Example

10, 8, 10, 8, 8, 4

MEAN = 8

VARIANCE = 4.8

STANDARD
DEVIATION = 2.19 

Step 1 : Find (n)

- You will need every number in your sample to calculate the mean. The mean is the average of all your data points.
- This is calculated by adding all of the numbers in your sample, then dividing this figure by the how many numbers there are in your sample (n).
- In the sample of test scores (10, 8, 10, 8, 8, 4) there are 6 numbers in the sample. Therefore $n = 6$.

Step 2: Find mean

- In the sample of test scores (10, 8, 10, 8, 8, and 4) there are six numbers, so $n = 6$.
- The sum of the test scores in the example was 48. So you would divide 48 by n to figure out the mean.
- $48 / 6 = 8$
- The mean test score in the sample is 8

Step 3: how much each data point differs from the mean.

- **Subtract the mean from each of your numbers in your sample.** This will give you a figure of how much each data point differs from the mean.
- For example, in our sample of test scores (10, 8, 10, 8, 8, and 4) the mean or mathematical average was 8.
- $10 - 8 = 2$; $8 - 8 = 0$, $10 - 8 = 2$, $8 - 8 = 0$, $8 - 8 = 0$, and $4 - 8 = -4$.
- Do this procedure again to check each answer. It is very important you have each of these figures correct as you will need them for the next step.

.

Step 4: squaring results

- Remember, in the example of test scores we started by subtracting the mean from each of the scores and squaring these figures:
- $(10-8)^2 + (8-8)^2 + (10-8)^2 + (8-8)^2 + (8-8)^2 + (4-8)^2$
- $4 + 0 + 4 + 0 + 0 + 16 = 24.$
- The sum of squares is 24.
-

Step5: Find Variance

- In our sample of test scores (10, 8, 10, 8, 8, and 4) there are 6 numbers. Therefore, $n = 6$.
- $n - 1 = 5$.
- Remember the sum of squares for this sample was 24.
- $24 / 5 = 4.8$
- The variance in this sample is thus 4.8.

.

Final Step: find Standard Deviation (SD)

- Usually, at least 68% of all the samples will fall inside one standard deviation from the mean.
- Remember in our sample of test scores, the variance was 4.8.
- $\sqrt{4.8} = 2.19$. The standard deviation in our sample of test scores is therefore 2.19.
- 5 out of 6 (83%) of our sample of test scores (10, 8, 10, 8, 8, and 4) is within one standard deviation (2.19) from the mean (8).

.

10, 8, 10, 8, 8, 4

MEAN = 8

VARIANCE = 4.8

STANDARD
DEVIATION = 2.19 ✓

Probability Distributions

Probability refers to a number, 0-1, representing chance of occurrence. Most people are comfortable with this everyday idea. **A** phrase such as "a 40 percent (.4) chance of delivery delay" seldom presents confusion, so long as we are clear about what "delivery delay" means.

Often a risk is a binary event. It will be either true or false. **A** contingency event either happens or not. Often we assign values of 0 (false) and 1 (true) as outcome values of a binary distribution. Spreadsheet IF statements are a prime example of using binary chance events.

Discrete Distributions

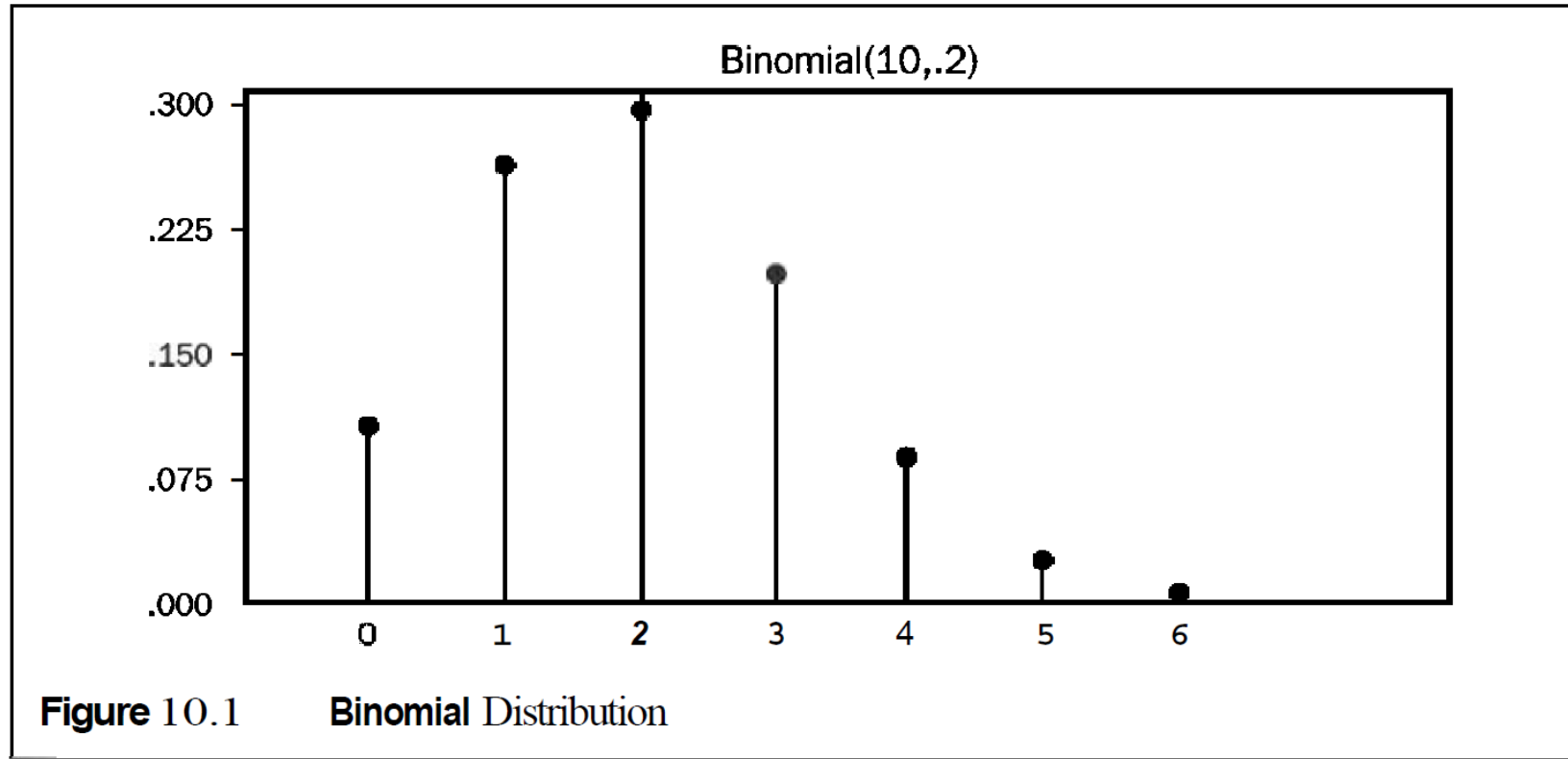
Does the chance event have two, or perhaps several, distinct possible outcomes? If so, we are talking about a discrete probability distribution. Frequently, the potential outcomes are integers, such as "number of work interruptions." The parameter name often begins with "number of." Examples include number of equipment breakdowns, number of people available, number of units needed, and so forth. Each possible outcome is assigned a probability, and these probabilities must sum to 1.

Simple systems often produce discrete distributions. Figure 10.1 and Figure 10.2 show the most popular discrete distributions, after the binary risk event. Though we don't have the distribution formulas printed here, because of space, most are simple enough. Software tools, even some hand calculators, are widely available with functions for the distributions described in this chapter.

Binomial Distributio

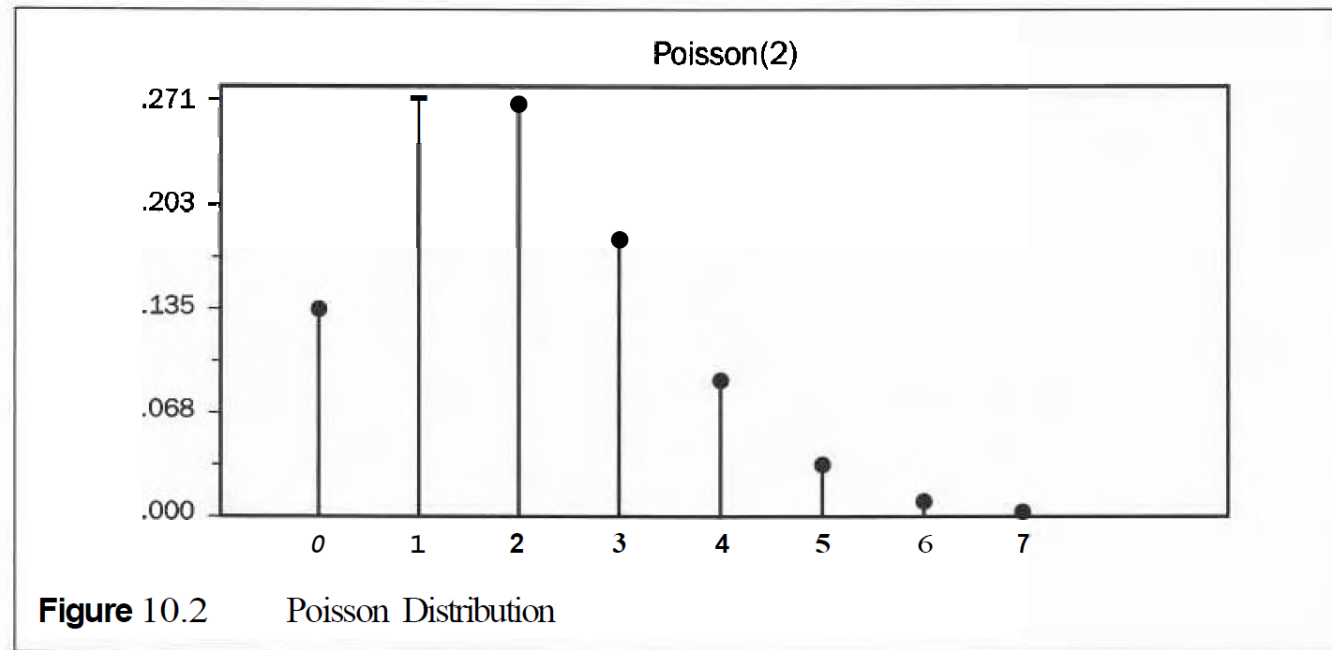
A collection of independent events, each with the same probability of success would exhibit a binomial distribution. For example, suppose that we have ten independent research projects, and we judge each as having a $p = 0.2$ chance of success. The binomial distribution characterizes this situation.

Though elements in actual systems may not be fully independent or have the same probability of success, the binomial distribution is often a reasonable approximation for a simple portfolio model. Figure 10.1 shows a binomial distribution with $n = 10$ trials with $p = .2$ probability of success on any trial. This distribution is often a starting point for representing simple portfolios, such as research projects.



Poisson Distribution

Sometimes events happen randomly, such as defects and work interruptions. The number of such events in a time or other interval is described by a Poisson distribution. The only parameter needed to specify this distribution is the average (μ) number of events occurring during an interval. Figure 10.2 shows a Poisson distribution with $\mu = 2$ occurrences. **This** distribution is widely used in simulating queuing (waiting-line) systems, such as representing the number of work requests during a time period.



Three-Level Estimates

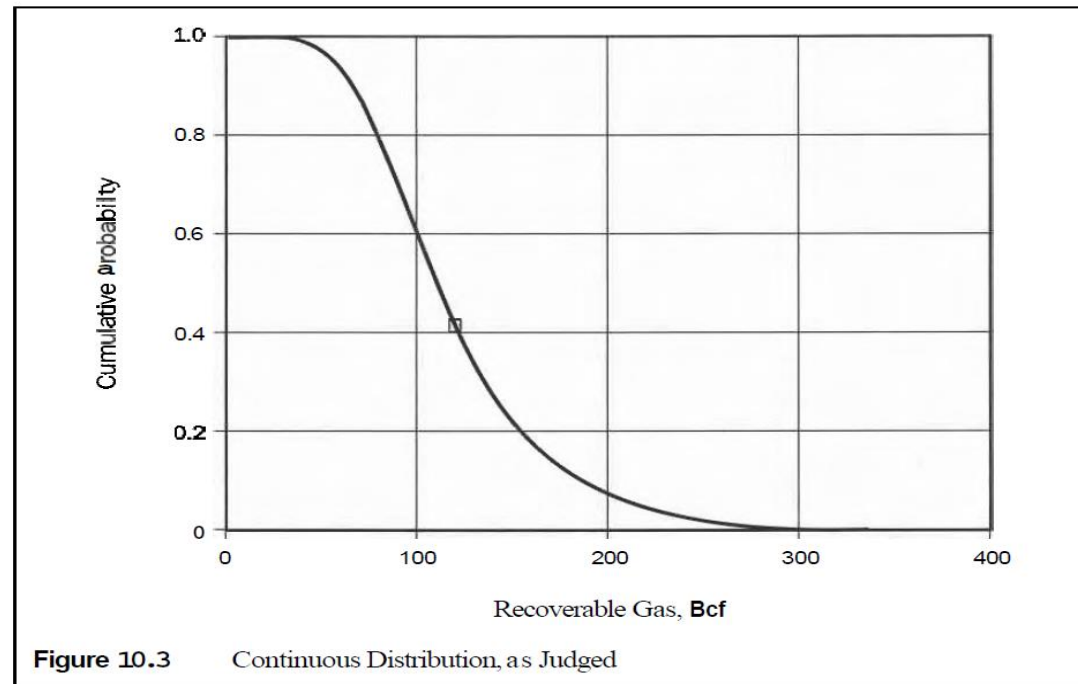
In decision tree analysis, we can use only discrete distributions. Thus, we must convert continuous chance events into discrete approximations. Typically, three branches are adequate: high, medium, and low case.

These are called "three-level estimates." We should choose these three values and their probability weights, so that the original continuous distribution's statistics are preserved. More than three outcomes can be used, if you want more resolution.

Suppose that your expert judges the uncertainty in volume of a natural gas reservoir as a lognormal distribution with $\mu = 120$ and σ cubic feet (Bcf). Figure 10.3 illustrates this judgment. With simulation, we would use the continuous distribution directly, as is. For a decision tree analysis, I **discretized** the distribution into the form in Figure 10.4. = 50 billion

Continuous Distributions

In project management and other types of planning and analysis, most uncertainties will be continuous. Cost, time, and quality metrics are prime examples. Figures **10.5** through **10.10** show the most popular continuous distribution types. Normally a scale is omitted from the y-axis. It is possible to show a scale, though the units are unnecessary. You should remember that the y-axis is scaled such that the total area under the curve-representing probability equals **1**. The probability of **an** x-axis value is proportional to the height of the curve above that point.



Uniform Distribution

The simplest continuous distribution is the *uniform*. This is often called a rectangle or boxcar distribution, because of its shape. If (a) one knows the upper and lower bounds of the range of possible values, and (b) if every value within that range is equally probable, we are describing a uniform distribution. Figure 10.5 illustrates a uniform distribution bounded at 0 and 2 units.

Many people incorrectly assign this distribution shape to uncertainties. However, it is a rare system that generates values that are uniformly distributed. More often, values in a central part of the range are more probable. Examples of uniform distributions include: position of a rotating shaft that stops randomly, position at which a cable breaks, and a small section of a broad distribution.

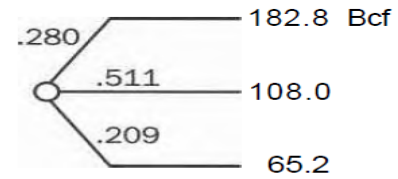


Figure 10.4 Three-Level Approximation

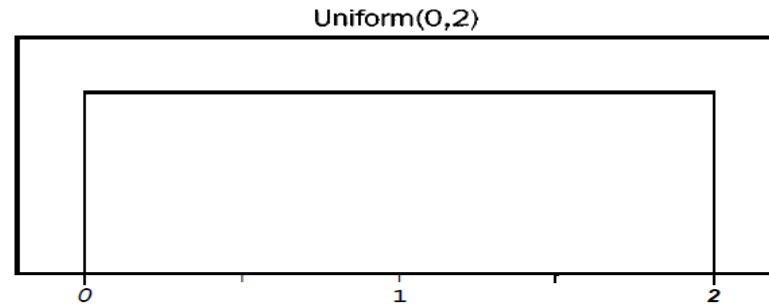
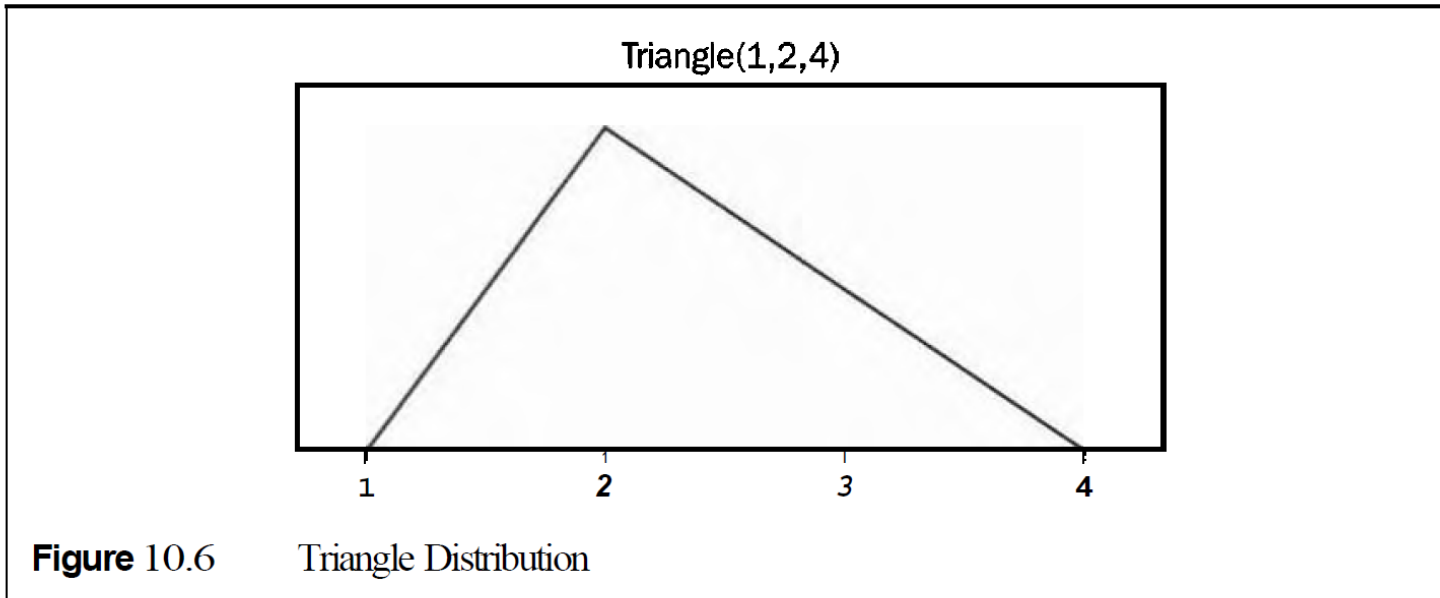


Figure 10.5 Uniform Distribution

The triangle distribution

The triangle distribution is very popular, though I've never seen a system in business or in nature generate values with a triangular shape. Simplicity is the appeal. Just three values fully describe the triangle distribution: minimum (L), most likely (M , mode), and maximum (E). Figure 10.6 shows an example.

Even people without training in probability can relate to these points. The shape of a particular distribution is usually less important than its position Q and width (which we might characterize with σ). Let me offer two cautions when using the triangle distribution:





Chapter 10

People tend to judge distributions too narrowly. The low and high parameters are the actual extremes, and not just pessimistic and optimistic cases. Many people will confuse "most likely" (the mode) with the best singlepoint estimate. It is not, unless the triangle happens to be symmetric. Drawing the distribution will help eliminate the confusion. The best single-point estimate is the mean Cu , Ev), and this happens to have a simple formula: $y = (L + M + q/3$.

Chapter
10

Normal Distribution

The most commonly observed distribution is the normal. Galileo observed the familiar bell-shaped distribution when measuring the positions of stars and planets. Figure 10.7 shows a normal distribution with a $\mu = 2$ and a $\sigma = 1$ unit. The normal distribution occurs often in nature and in business. When we sum many independent, continuous chance events together, the total tends to be normally distributed.

Simple project models often assume that activities are independent. Then if the activities have similar duration, the distribution for time to complete a chain of activities is approximately normal. Amazingly, this is true regardless of the shape of the component distributions. In statistics, the central **limit** theorem--one of the fundamental principles in statistics--describes this key behavior.



Chapter 10

Most every statistics book includes a table of the standard normal distribution ($\mu = 0, \sigma = 1$). You can obtain the same values from the NORMDIST function in Excel. From a table or the function, one can determine confidence intervals. For example, an interval $\bar{y} \pm 1\sigma$ represents a 68 percent confidence interval for a normal distribution. The vertical dashed lines in Figure 10.7 show $\pm 2\sigma$ intervals. In theory (and equation), the normal distribution is boundless. As a practical matter, for planning and evaluation work, there is virtually no chance (0.3 percent) of being outside the range of $\bar{y} \pm 3\sigma$.



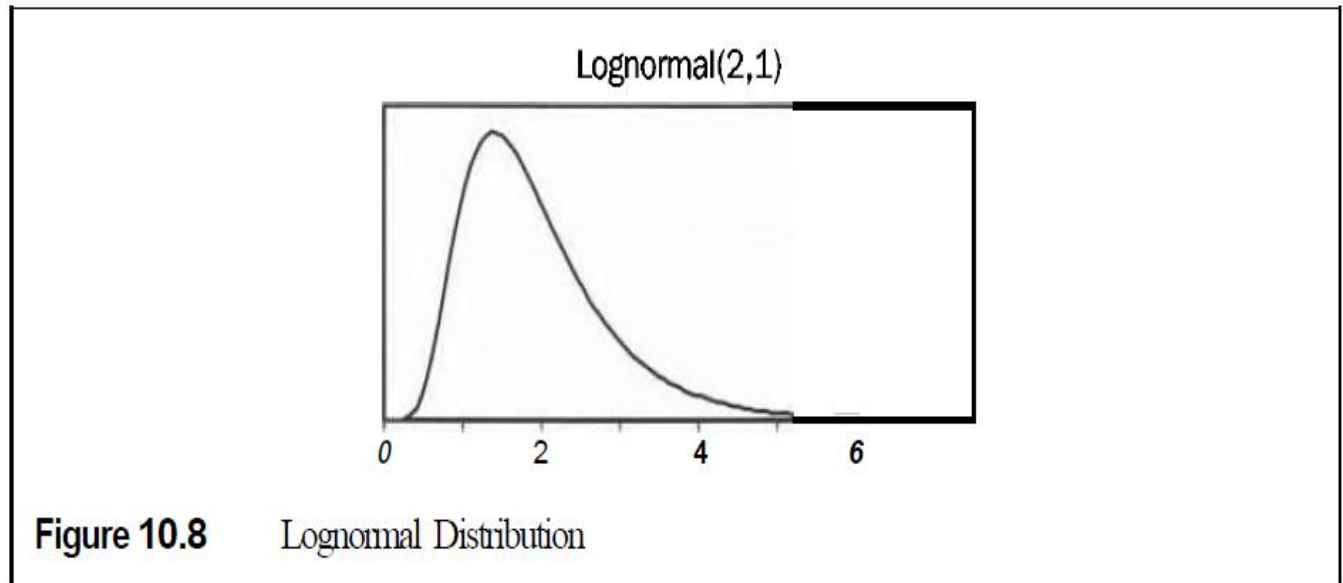
Chapter 10

Lognormal Distribution

We often observe data where the frequency distribution is positively skewed—that is, having a longer tail to the right. When two or more distributions are multiplied, such as time $\times$ hourly cost, the product is typically positively skewed. As we multiply more independent components together, the product becomes more lognormal in shape. *Ws*, too, is a result of the *central limit theorem*. Some older readers may remember adding logarithms on slide rules to multiply.)

Figure 10.8 shows an example lognormal distribution, with $p = 2$ and $\sigma = 1$. This is a common distribution shape for uncertainties that are positively skewed.

Chapter
10





Chapter 10

The log of values obtained from a lognormal population plot as a normal distribution. Special "probability graph paper" is available for plotting the cumulative frequency distribution.

This has one axis linear in standard deviations of a normal distribution. If the cumulative frequency of the data plots as a line on a linear value scale, then the data are normally distributed. If the data plot as a line on log value scale, then the data are lognormally distributed.



Chapter 10

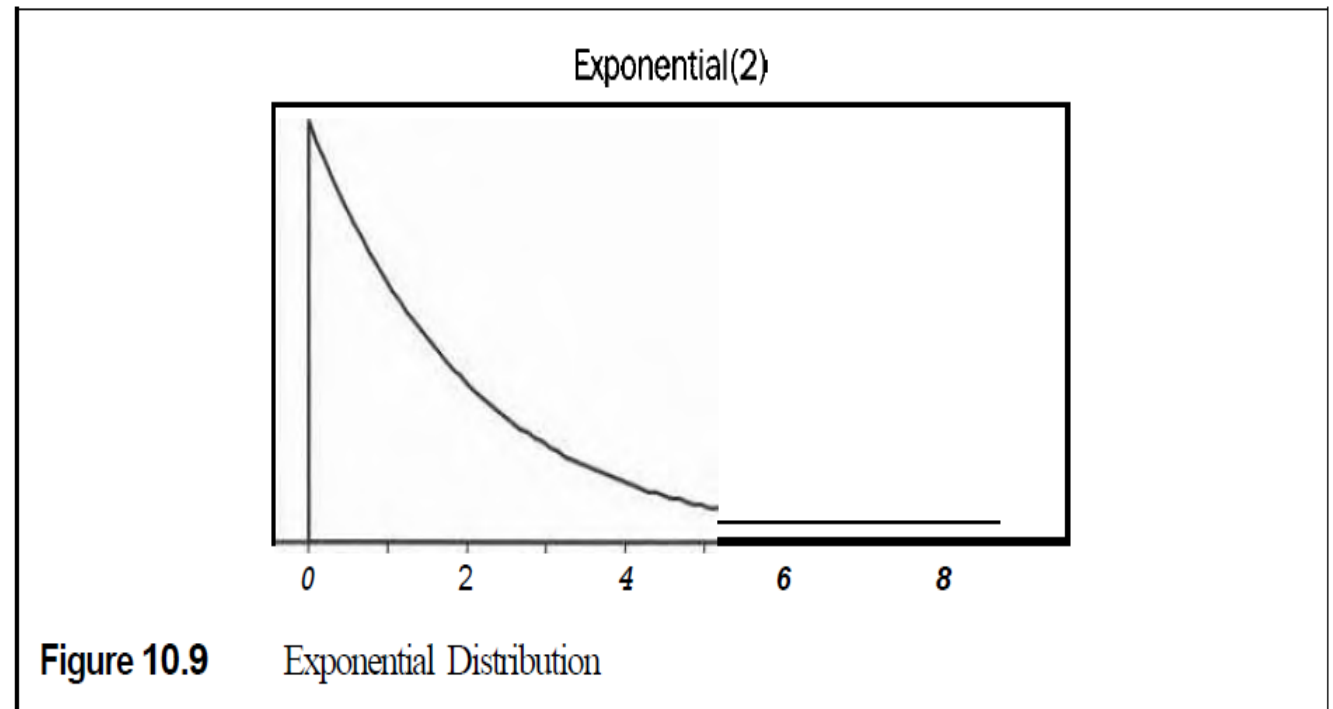
Both the normal and lognormal distributions are completely described by μ and σ . Some computer tools accept alternate statistics to specify the distribution, such as the 10 percent and 90 percent confidence points.

Chapter
10

Exponential Distribution

The *exponential* distribution is a counterpart to the discrete Poisson distribution. The exponential distribution is most commonly used to represent time between arrivals of random events. Figure 10.9 shows an exponential distribution with $p = 2$ units:

Chapter
10





Chapter 10

Like the Poisson, we use the exponential distribution in discrete-event simulations, such as of queuing systems. Common uses include representing the time until the next arrival of a service request or unscheduled work interruption.

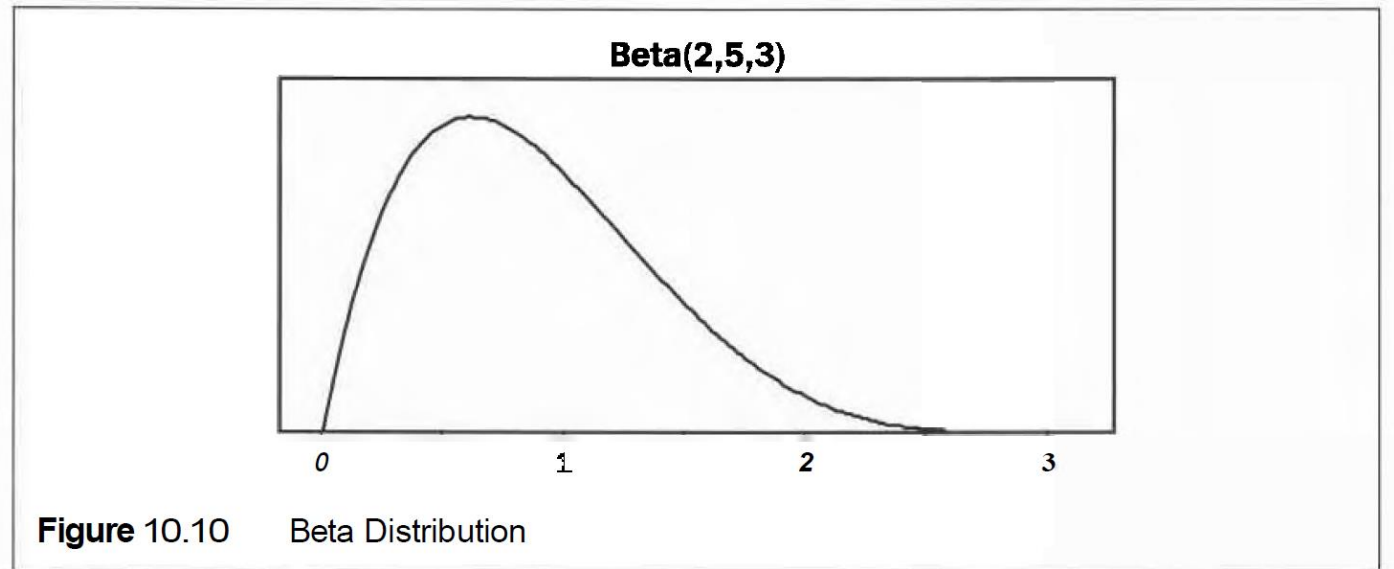


Chapter 10

Beta distribution

The beta distribution is a mathematical function that assumes a variety of profiles, depending upon two shape parameters. The basic distribution is bounded by 0 and 1. In the distribution illustrated in Figure 10.10, the shape parameters are 2 and 5, and **3** is a scaling factor. With the shape factors, we can make many different symmetric and asymmetric shapes. We can reposition and rescale the distribution as needed. Project management professionals have used beta distributions since the 1950s to represent activity times-to-complete in PERT project models.

Chapter
10



Chapter
10

Expected Activity Duration (Triangular Distribution)	Expected Activity Duration (Beta Distribution)	Beta Activity Standard Deviation
$\frac{P+M+O}{3}$	$\frac{P+4M+O}{6}$	$\frac{P-O}{6}$

Legend: P = Pessimistic, M = Most Likely, O = Optimistic

Figure 6.6: Formulas for Three-Point Estimating and Activity Standard Deviation

حيث يمثل Mean = حساب متوسط الوقت للنشاط بطريقة PERT
ويمثل Standard Deviation الانحراف المعياري ويشار له بالرمز σ او
سيقما Sigma
ويمثل Variance السماحية في الوقت وهو حاصل تربيع الانحراف المعياري

Chapter
10

بناء على معلومات المشروع التالية :

$T_p=32$ days $T_m=26$ days $T_o= 22$ days

ماهو الوقت المتوقع لانتهاء النشاط؟ في حالة ان نسبة الانحراف المعياري المتبعه في المنظمة هي 95.46% = 2 سجا

الاجابة

الوقت المتوقع لانتهاء النشاط حسب قانون PERT: $(T_p+4T_m+T_o)/6 =$
 $26.33 = (22+(4*26)+32)/6 =$ يوم

ولحساب الوقت بدقة اكثر
 في حالة ان نسبة الانحراف المعياري المتبعه في المنظمة هي 95.46%

نبدأ اولاً بحساب الانحراف المعياري للأيام:

الانحراف المعياري (سيجما) $(\sigma) = (t_p - t_o) / 6 = (32 - 22) / 6 = 1.66$ يوم

من الشكل المبين يتضح ان قيمة 95.46% يقابله انحراف معياري بمقدار 26

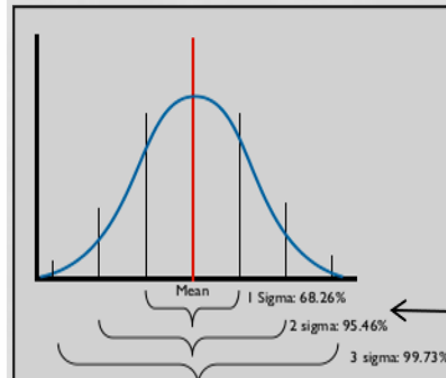
وهذا يعني معدل الفرق في الوقت = $2 * 1.66 = 3.32$ يوم
 فيصبح الوقت المتوقع لانتهاء النشاط هو 26.33 ± 3.32 يوم
 اي ما بين 23 يوم الى 29.64 يوم

PERT:

$$\text{Mean} = \frac{t_o + 4 \cdot t_{ML} + t_p}{6}$$

$$\text{Standard Deviation} = \frac{t_p - t_o}{6}$$

$$\text{Variance} = \sigma^2$$





Chapter 10

If you are the expert, the best distribution is the one that completely expresses your belief about the uncertainty. Often we have data that suggest that a system is producing a common distribution shape. Be aware that mathematicians have described hundreds of distributions, and your system need not produce a distribution with a recognizable mathematical function. Fortunately, the distributions listed earlier usually suffice. Importantly, the popular software tools allow users to specify custom distribution shapes. Without data, perhaps the best way to obtain a distribution is to model the subsystem that causes the uncertainty. If the expert can explain how the process works, we can model it and use the model to generate a distribution shape. The amount of effort in detailed modeling should be consistent with the importance of the variable to the decision at hand.



Chapter 10

Usually our best available experts provide judgments that go into the project models. Delegating responsibility for these inputs is a responsibility of the decision maker. **A** probability distribution completely describes a judgment about a single variable. There is an additional dimension of expert opinion: Are there influencing relationships among the variables? That is the topic of Chapter 12. First, we will discuss the source of these probability distributions in Chapter 11, Judgments and Biases.



Thank You