Open-Ended Problems: A Future Chemical Engineering Education Approach. J. Patrick Abulencia and Louis Theodore.

19

Probability and Statistics

This chapter is concerned with process probability and statistics. As with all the chapters in Part II, there are several sections: Overview, several technical topics, illustrative open-ended problems, and open-ended problems. The purpose of the first section is to introduce the reader to the subject of probability and statistics. As one might suppose, a comprehensive treatment is not provided although several technical topics are also included. The next section contains three open-ended problems; the authors' solution (there may be other solutions) are also provided. The final section contains 42 problems; *no* solutions are provided here.

19.1 Overview

This overview section is concerned with—as can be noted from the chapter title—probability and statistics. As one might suppose, it was not possible to address all topics directly or indirectly related to probability and statistics. However, additional details may be obtained from either the references provided at the end of this Overview and/or at the end of the chapter.

Note: Those readers already familiar with the details associated with this topic may choose to bypass this Overview.

The title of this chapter is Probability and Statistics. Webster [1] defines *probability* as "the quality or state of being probable; likelihood; something probable; the number of times something will probably occur over the range of possible occurrences, expressed as a ratio" and the term *statistics* as "facts or data of a numerical kind, assembled, classified, and tabulated so as to present significant information about a given subject; the science of assembling, classifying, tabulating, and analyzing such facts or data". There are obviously many other definitions.

The key area of interest to chemical engineers is however, statistics. The problem often encountered is usually related to interpreting limited data and/or information. This can entail any one of several topics.

1. Obtaining additional data and/or information
2. Deciding which data and/or information to use
3. Generating a mathematical model (generally an equation) to represent the data and/or information
4. Generating information about unknowns, a process often referred to as *inference*

The material in this chapter is divided into 6 sections. Titles are provided below.

1. Probability Definitions and Interpretations
2. Introduction to Probability Distributions
3. Discrete and Continuous Probability Distributions
4. Contemporary Statistics
5. Regression Analysis
6. Analysis of Variance

Topic (5) receives the bulk of the treatment in the presentation to follow.

Finally, the reader should note that much of the material in this chapter was adopted from L. Theodore and F. Taylor, "Probability and Statistics," Theodore Tutorials, East Williston, NY, originally published by USEPA/APTI, RTP, NC, 1993 [2].

19.2 Probability Definitions and Interpretations [2,3]

Probabilities are nonnegative numbers associated with the outcomes of so-called *random* experiments. A random experiment is an experiment whose outcome is uncertain. Examples include throwing a pair of dice, tossing a coin, counting the number of defectives in a sample from a lot of manufactured items, and observing the time to failure of a tube in a heat exchanger or a seal in a pump or a bus section in an electrostatic precipitator. The set of possible outcomes of a random experiment is called the *sample space* and is usually designated by S. Then P(A), the probability of an event A, is the sum of the probabilities assigned to the outcomes constituting the subset A of the space S. A *population* is a collection of *objects* having observable or measureable characteristics defined as a *variate* while a *sample* is a group of "objects drawn from a population (usually random) where each is likely to be drawn."

Consider, for example, tossing a coin twice. The sample space can be described as

$$S = [HH, HT, TH, TT] \tag{19.1}$$

If probability ¼ is assigned to each element of S and A is the event of at least one head, then

$$A = [HH, HT, TH] \tag{19.2}$$

The sum of the probabilities assigned to the elements of A is ¾. Therefore, P(A) = ¾. The description of the sample space is not unique. The sample space S in the case of tossing a coin twice could be describe in terms of the number of heads obtained. Then

$$S = [0, 1, 2] \tag{19.3}$$

Suppose probabilities ¼, ½, and ¼ are assigned to the outcomes 0, 1, and 2, respectively. Then A, the event of at least one head, would have for its probability,

$$P(A) = P(1,2) = 3/4 \tag{19.4}$$

Probability P(A) can also be interpreted *subjectively* as a measure of degree of belief, on a scale from 0 to 1, that the event A occurs. This interpretation is frequently used in ordinary conversation. For example, if someone says, "The probability that I will go the race track tonight is 90%," then 90% is a measure of the person's belief that he or she will go to a race track (a site regularly visited by one of the authors). This interpretation is also used when, in the absence of concrete data needed to estimate an unknown probability on the basis of observed relative frequency, the personal opinion of an expert is sought. For example, a chemical engineer might be asked to estimate the probability that the seals in a newly designed pump will leak at high pressures. The estimate would be based on the expert's familiarity with the history of pumps of similar design.

19.3 Introduction to Probability Distributions [2,3]

The probability distribution of a random variable concerns the distribution of probability over the range of the random variable. The distribution of probability, i.e., the values of random variables together with their associated probabilities, is specified by the *probability distribution function* (pdf). This section is devoted to providing general properties of the pdf for the case of discrete and continuous random variables as well as an introduction to the *cumulative distribution function* (cfd). Special pdfs finding extensive application in chemical engineering analysis are presented in the next section.

The pdf of a *discrete* random variable X is specified by f(x), where f(x) has the following essential properties

1. F(x) = P(X)=x) = probability assigned to the outcome corresponding to the number x in the range of X; i.e., X is a *specifically designated* value of x. (19.5)
2. F(x) ≥ 0 (19.6)
3. $\sum_{x} f(x) = 1$ (19.7)

Property 1 indicates that the pdf of a discrete random variable generates probability by substitution. Property 2 and Property 3 restrict the values of f(x) to nonnegative real numbers and numbers whose sum is 1, respectively.

The pdf of a *continuous* random variable X has the following properties:

1. $\int_{a}^{b} f(x)\,dx = P(a < X < b)$ (19.8)

2. $F(x) \geq 0$ (19.9)

3. $\int_{-\infty}^{\infty} f(x)\,dx = 1$ (19.10)

Equation (19.8) indicated that the pdf of a continuous random variable generates probability by integration of the pdf over the interval whose probability is required. When this interval contracts or reduces to a single value, the integral over the interval becomes *zero*. Therefore, the probability associated with any particular value of a continuous random variable is therefore zero. Consequently, X is continuous

$$\begin{aligned} P(a < X \leq b) &= P(a < X \leq b) \\ &= P(a < X < b) \\ &= P(a \leq X < b) \end{aligned} \tag{19.11}$$

Equation (19.9) restricts the values of f(x) to nonnegative numbers. Equation (19.10) follows from the fact that

$$P(-\infty < X < \infty) = 1 \tag{19.12}$$

The expression $P(a < X < b)$ can be interpreted geometrically as the area under the pdf curve over the interval (a, b). Integration of the pdf over the interval yields the probability assigned to the interval. For example, the probability that the time in hours between successive failures of an aircraft air conditioning system X is greater than 6 but less than 10 is $P(6 < X < 10)$.

Another function used to describe the probability distribution of a random variable X is the *cumulative distribution function* (cdf). If $f(x)$ specifies the pdf of a random variable X, then $F(x)$ is used to specify the cdf. The cdf of X is defined by for both discrete and continuous random variables

$$F(x) = P(X \geq x); -\infty < x < \infty \tag{19.13}$$

Note that the cdf is defined for all real numbers, not just the values assumed by the random variable. It is helpful to think of $F(x)$ as an accumulator of probability as x increases through all real numbers. In the case of a discrete random variable, the cdf is a step function increasing by finite jumps at the values of x in the range of X. In the case of a continuous random variable, the cdf is a continuous function.

The following properties of the cdf of a random variable X can be deduced directly from the definition of $F(x)$.

1. $F(b) - F(a) = P(a < X \leq b)$ (19.14)
2. $F(+\infty) = 1$
3. $F(-\infty) = 0$
4. Also note that $F(x)$ is a nondecreasing function of X

As noted above, these properties apply to the cases of both discrete and continuous random variable.

Shaefer and Theodore [3] provided numerous illustrative examples dealing with probability distribution.

19.4 Discrete and Continuous Probability Distributions [2,3]

19.4.1 Discrete Probability Distributions

There are numerous discrete probability distributions. However, this section simply lists the three distributions the chemical engineer is most likely to encounter in his/her career.

1. The Binomial distribution
2. The Hypergeometric distribution
3. The Poisson distribution

Each receives treatment in Shaefer and Theodore [3], including several other distributions.

There are numerous continuous probability distributions. The four distributions the chemical engineer is most likely to encounter in his/her career are listed below

1. The exponential distribution
2. The Weibull distribution
3. The normal distribution
4. The log-normal distribution

The literature [3] provides details on these as well as other continuous probability distribution topics [4].

19.5 Contemporary Statistics

This section examines topics that Shaefer and Theodore [3] have classified as contemporary statistics; their development includes numerous illustrative examples. The following subject areas are reviewed by the authors.

1. Confidence Intervals for Means
2. Confidence Intervals for Proportions
3. Hypothesis Testing
4. Hypothesis Test for Means and Proportions
5. Chi-Square Distribution
6. The F Distribution
7. Nonparametric Tests

19.6 Regression Analysis (3)

It is no secret that many statistical calculations are now performed with spreadsheets or packaged programs; this statement is particularly true with *regression analysis*, a topic of interest to all chemical engineers. The result of this shortsighted approach has been to reduce or eliminate one's fundamental understanding of this subject. This section attempts to correct this shortcoming.

Chemical engineers and scientists in numerous disciplines often encounter applications that require the need to develop a mathematical relationship between data for two or more variables. For example, if Y (a dependent variable) is a function of or is dependent of X (an independent variable), that is

$$Y = f(X) \tag{19.15}$$

one may be required to express this (X, Y) data in equation form. This process is referred to as *regression analysis*, and the regression method most often employed is the method of *least squares*.

An important step in this procedure – which is often omitted – is to prepare a plot of Y vs. X. The result, referred to as a *scatter* diagram, could take on any form. This topic received treatment in Chapter 2, but is provided again in Figure 19.1 (a)-(c). The first plot (A) suggests a linear relationship between X and Y, i.e.,

$$Y = a_0 + a_1X \tag{19.16}$$

The second graph (B) appears to be best represented by a second order (or parabolic) relationship, i.e.,

$$Y = a_0 + a_1X + a_2X^2 \tag{19.17}$$

The third plot (C) suggests a linear model that applies over two different ranges, i.e., it should represent the data as

$$Y = a_0 + a_1 X; \quad X_0 < X < X_M \tag{19.18}$$

and

$$Y = a'_0 + a'_1 X; \quad X_M < X < X_L \tag{19.19}$$

This multiequation model finds application in representing adsorption equilibria, multiparticle size distributions, and quantum energy relationship. In any event, a scatter diagram and individual judgment can suggest an appropriate model at an early stage in the analysis.

Some of the models often employed by chemical engineers are as follows

1. $a_0 + a_1 X$	Linear	(19.20)
2. $a_0 + a_1 X + a_2 X^2$	Parabolic	(19.21)
3. $a_0 + a_1 X + a_2 X^2 + a_3 X^3$	Cubic	(19.22)
4. $a_0 + a_1 X + a_2 X^2 + a_3 X^3 + a_4 X^4 + \ldots$	Higher Order	(19.23)

Procedures to evaluate the regression coefficients a_0, a_1, a_2, etc., are available in the literature. The least squares technique provides numerical values for the regression coefficients a_i such that the sum of the square of the difference (error) between the actual Y and Y_e predicted by the equation or model is minimized.

The correlation coefficient provides information on how well the model, or line of regression, fits the data; it is denoted by *r*. The procedure to calculate *r* and its properties are also available [2]. It should be noted that the correlation coefficient *only* provides information on how well the model fits the data. It is emphasized that *r* provides no information on how good the model is or, to reword this, whether this is the correct or best model to describe the functional relationship of the data. This topic is briefly discussed in the next section.

19.7 Analysis of Variance

Analysis of variance is a special statistical technique that chemical engineers may become faced with during their career. It features the splitting of the total variation of data into components measuring variation attributable to one or more factors or combinations of factors. The simplest application of analysis of variance involves data classified in categories (levels) of one factor. This topic receives extensive treatment in Shaefer and Theodore [3]. Interested readers should review the other literature resources of a statistics nature.

A detailed and expanded treated of probability and statistics is available in the following three references.

1. L. Theodore and F. Taylor, *Probability and Statistics*, A Theodore Tutorial, Theodore Tutorials, East Williston, NY, originally published by USEPA/APTI, RTP, NC, 1993 [2].
2. S. Schaefer and L. Theodore, *Probability and Statistics Applications in Environmental Science,* CRC Press/ Taylor & Francis Group, Boca Raton, FL, 2007 [3].
3. L. Theodore, =*Chemical Engineering: The Essential Reference*, McGraw-Hill, New York City, NY, 2014 [6].

19.8 Illustrative Open-Ended Problems

This and the last section provide open-ended problems. However, solutions *are* provided for the three problems in this section in order for the reader to hopefully obtain a better understanding of these problems which differ from the traditional problems/illustrative examples. The first problem is relatively straightforward while the third (and last problem) is somewhat more difficult and/or complex. Note that solutions are not provided for the 42 open-ended problems in the next section.

Problem 1: The problem of calculating probabilities of *objects* or *events* in a finite group – defined earlier as the *sample space* – in which equal probabilities are assigned to the elements in the sample space requires counting the elements which make up the events. The counting of such events if often greatly simplified by employing the rules for permutations and combinations. Define and describe these two terms.

Solution: Permutations and combinations deal with the grouping and arrangement of objects or events. By definition, each different ordering in a given manner or arrangement *with* regard to order of all or part of the objects is called a *permutation*. Alternately, each of the sets which can be made by using all or part of a given collection of objects *without* regard to order of the objexts in the set is called a *combination*. Although,

TABLE 19.1 Subsets of Permutations and Combinations

Permutations (With Regard to Order)	**Combinations (Without Regard to Order)**
Without replacement	Without replacement
With replacement	With replacement

permutations or combination can be obtained with *replacement* or *without replacement*, most analyses of permutations and combinations are based on a sample that is performed without replacement; i.e., each object or element can be used only once.

For each of the two *with/without* pairs (with/without regard to order and with/without replacement), four subsets of two may be drawn. These four are provided in Table 19.1. To personalize this, the reader could consider the options (games of change) one of the authors faces while on a one-day visit to a casino. The only three options normally considered are dice (often referred to as craps), blackjack (occasionally referred to as 21), and pari-mutual (horses, trotters, dogs, and jai alai) simulcasting betting. All three of these may be played during a visit, although playing two or only one is also an option. In addition the order may vary and the option may be repeated. Some possibilities include the following:

1. Dice, blackjack, and then simulcast wagering
2. Blackjack, wagering, and dice
3. Wagering, dice, and wagering
4. Wagering and dice (the author's usual sequence)
5. Blackjack, blackjack (following a break), and dice

Problem 2: Provide a general introduction to the subject of design of experiments (DOE).

Solution: One of the more difficult decisions facing the authors during the writing of this text was how to handle/treat the more advanced subject of "design of experiments" (DOE), not to be confused with Department of Energy. The material presented in the introduction discussed statistical techniques employed to analyze data obtained from "experiments", with little or no consideration given to the experiment itself. For most individuals desiring an understanding of statistics, the planning and design of experiments is not only outside their control, but also beyond the required background of that individual. It is only on rare occasions that the procedures and details of this topic are mandated. Ultimately, it was decided primarily to provide a very short introduction to DOE because of the enormous depth of the topic. That introduction follows.

The terminology employed in DOE is different to some degree from that employed earlier. Traditionally, experimental variables are usually referred to as *factors*. The factor may be continuous or discrete. Continuous factors include pollutant concentration, temperature, and drug dosage; discrete factors include year, type of process, time, machine operator, and drug classification. The particular value of the variable is defined as the *level* of the factor. For example, if one were interested in studying a property (physical or chemical) of a nanoparticle at two different sizes, there are two levels of the factor (in this case, the nanoparticle's size). If a combination of factors is employed, they are called *treatments* with the result defined as the *effect*. If the study is to include the effect of the shape of the nanoparticle, these different shapes are called *blocks*. Repeating an experiment at the same conditions is called a *replication*.

In a very general sense, the overall subject of DOE can be thought of as consisting of four steps:

1. Determining the importance of (sets of) observations that can be made to the variable or variables one is interested in
2. The procedure by which the observations will be made, i.e., how the data is gathered
3. The review of the observations, i.e., how the data are treated/examined,
4. The analysis of the final results

The reader should also note that obtaining data can be:

1. Difficult
2. Time consuming
3. Dangerous
4. Expensive
5. Affected by limited resources available for experimentation

For these reasons there are obviously significant advantages to expending time on the design of an experiment *before* the start of the experiment. The variables can have a significant effect on the final results; careful selection can prove invaluable. In summary, the purpose of experimental design is to obtain as much information concerning a "response" as possible with the minimum amount of experimental work possible.

Problem 3: A technical definition of both a probability distribution function (pdf) and a cumulative distribution function (cdf) was presented in the Overview. Provide some specific examples of both pdfs and cdfs.

Solution: Consider, for example, a box of 100 transistors containing 5 defectives. Suppose that a transistor selected at random is to be classified as defective or nondefective. Then X is a discrete random variable with pdf specified by

$$\begin{aligned} f(x) &= 0.05; \quad x = 1 \\ &= 0.95; \quad x = 0 \end{aligned}$$

For another example of the pdf or a discrete random variable, let X denote the number of the throw on which the first failure of an electrical switch occurs. Suppose the probability that a switch fails on any throw is 0.001 and that successive throws are independent with respect to failure. If the switch fails for the first time on throw x, it must have been successful on each of the preceding $x - 1$ trials. In effect, the switch survives up to $x - 1$ trials and fails at trial x. Therefore, the pdf of X is given by

$$f(x) = (0.999)^{x-1}(0.001); \quad x = 1, 2, 3, \ldots, n, \ldots$$

Note that the range of X consists of a countable infinitude of values.

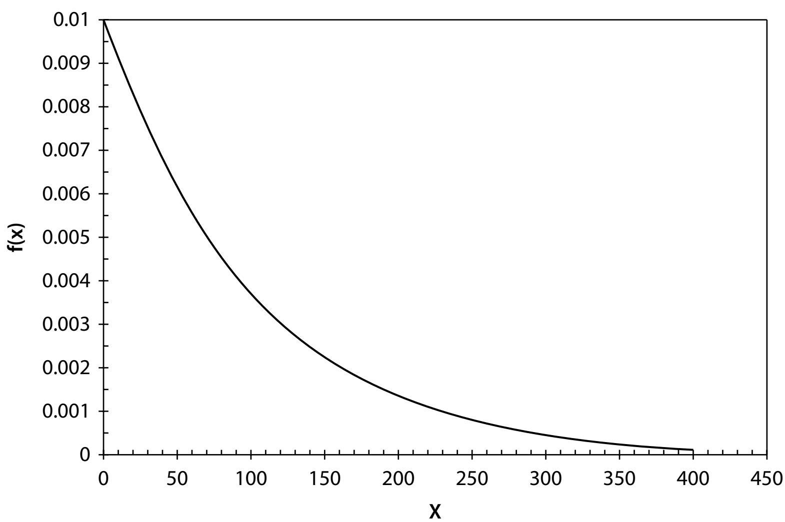

Figure 19.1 The pdf of time in hours between successive failures of an aircraft air conditioning system.

As an example of the pdf of a continuous random variable, consider the pdf of the time x in hours between successive failures of an aircraft air conditioning system. Suppose the pdf of x is specified by

$$\begin{aligned} f(x) &= 0.01e^{-0.01x}; \quad x > 0 \\ &= 0; \text{ elsewhere} \end{aligned}$$

A plot of $f(x)$ vs. x for positive values of x is provided in Figure 19.2. Inspection of the graph indicates that intervals in the lower part of the range of x are assigned greater probabilities than intervals of the same length in the upper part of the range of x because the areas over the former are greater than the areas over the latter.

To illustrate the derivation of the cdf from the pdf, consider the case of a random variable X whose pdf is specified by

$$\begin{aligned} f(x) &= 0.2; \quad x = 2 \\ &= 0.3; \quad x = 5 \\ &= 0.5; \quad x = 7 \end{aligned}$$

Applying the definition of the cdf provided earlier, one obtains for the cdf of X (see also Figure 19.3).

$$\begin{aligned} f(x) &= 0; \quad x < 2 \\ &= 0.2; \quad 2 \le x < 5 \\ &= 0.5; \quad 5 \le x < 7 \\ &= 7; \quad x \ge 7 \end{aligned}$$

It is helpful to think of $f(x)$ as an accumulator of probability as x increases through all real numbers. In the case of a discrete random variable, the cdf is a step function increasing by finite jumps at the values of x in the range of x. In the aforementioned example, these jumps occur at the values 2, 5, and 7. The magnitude of each jump is equal to the probability assigned to the value at which the jump occurs. This is depicted in Figure 19.2.

Finally, in the case of a continuous random variable, the cdf is a continuous function. Suppose, for example, that x is a continuous random variable with pdf specified by

$$\begin{aligned} f(x) &= 2x; \quad 0 \le x < 1 \\ &= 0; \quad \text{elsewhere} \end{aligned}$$

Applying the definition once again, one obtains

$$
\begin{aligned}
F(x) &= 0; && x < 0 \\
\int_0^x 2x\,dx = x^2&; && 0 \le x < 1 \\
&= 1; && x \ge 1
\end{aligned}
$$

Figure 19.3 displays the graph of this cdf, which is simply a plot of $f(x)$ vs. x. Differentiating cdf and setting pdf equal to zero where the derivative of cdf does not exist can provide the pdf of a continuous random variable. For example, differentiating the cdf of x^2 yields the specified pdf of $2x$. In this case, the derivative of cdf does not exist for $x = 1$.

19.9 Open-Ended Problems

This last section of the chapter contains open-ended problems as they relate to probability and statistics. No detailed and/or specific solution is provided; that task is left to the reader, noting that each problem has either a unique solution or a number of solutions or (in some cases) no solution at all. These are characteristics of open-ended problems described earlier.

There are comments associated with some, but not all, of the problems. The comments are included to assist the reader while attempting to solve the problems. However, it is recommended that the solution to each problem should initially be attempted *without* the assistance of the comments.

There are 42 open-ended problems in this section. As stated above, if difficulty is encountered in solving any particular problem, the reader

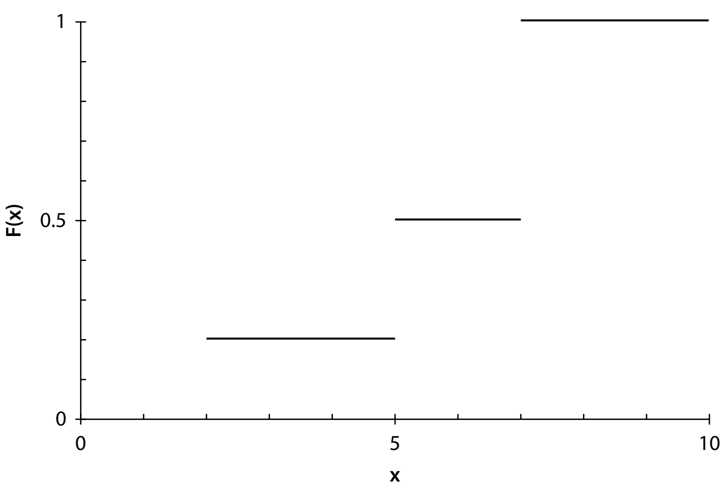

Figure 19.2 Graph of the cdf of a discrete random variable x.

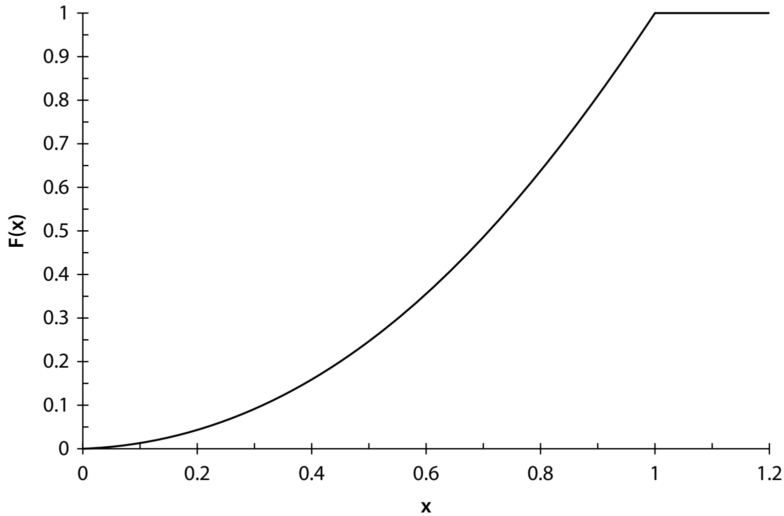

Figure 19.3 Graph of the cdf of a continuous random variable x.

should next refer to the comment, if any is provided with the problem. The reader should also note that the more difficult problems are generally located at or near the end of the section.

1. Describe the early history associated with probability.
2. Describe the early history associated with statistics.
3. Discuss the recent advances in statistics.
4. Select a refereed, published article on probability and statistics from the literature and provide a review.
5. Provide some normal everyday domestic applications involving the general topic of probability and statistics.
6. Develop an original problem on either probability or statistics that would be suitable as an illustrative example in a book.
7. Prepare a list of the various technical books which have been written on probability and statistics. Select the three best (hopefully including the text written by one of the authors) and justify your answer. Also select the three weakest books and, once again, justify your answer.
8. Why is the general subject of probability and statistics important to the chemical engineer?
9. Define set notation in layman terms.
10. Define conditional probability in layman terms.
11. What is a random variable?

12. What are random numbers?
13. Define Bayes' Theorem in layman terms.
14. Describe how the pdf and cdf for a random discrete variable differ from that of a random continuous variable.
15. Describe the hypergeometric distribution.
16. Discuss the relationships between the hypergeometric and binomial distributions.
    Comment: Refer to the literature [3] for additional details.
17. The average number of breakdowns of personal computers during 1000 hours of operating of a computer center is 2. What is the probability of no breakdowns during a work period in the 1 – 12 hour range? Comment on the results.
18. One of the authors recently bet on ten basketball games at the Mirage simulcasting center in Las Vegas. Assume the odds of winning the bet are 0.5. The probability of breaking even, i.e., winning 5 of the bets, can be calculated via application of the binomial theorem with $n = 10$, $p = 0.5$, and $x = 5$.

$$
\begin{aligned}
P(X = 5) &= \frac{10!}{5!5!}(0.5)^5(0.5)^5 \\
&= (252)(0.03125)(0.03125) \\
&= 0.246 = 24.6\%
\end{aligned}
$$

    Revise and rewrite the above Problem statement so that it relates to a technical application of your choice.
19. Revise and rewrite Problem 18 so that it applies to a domestic (everyday) application.
20. The probability that U.S. citizens of age 72 to 73 (once the age of one of the authors) will die within the year was 0.0417. With a group of 1,000,000 such individuals, what is the probability that exactly X will die within the year? Perform the calculations for values of X in the 1 x 10^4 – 20 x 10^4 range. Analyze the results.
    Comment: As one might surmise, one of the authors survived that year. The author is currently 80 and the probability of dying has unfortunately increased significantly.
21. Provide a layman's definition of the Weibull distribution.
22. One of the authors [4] is currently attempting to develop a failure model to replace the Weibull distribution. You have been assigned a similar task. Note that the Weibull distribution contains 6 coefficients—two for each of three

different ranges. The author(4) has developed a of 4-coefficient equation to replace the Weibull equation. You have been hired to develop "another" failure rate equation that consists of *less* than 4 coefficients.

23. Define the log-normal distribution and discuss its relationship to the normal distribution.
24. Attempt to develop another probability distribution.
25. Provide technical definitions for the following two terms.
    - Confidence interval
    - Confidence coefficient
26. Describe both the confidence limit and confidence coefficient in layman terms.
27. Provide definitions for the following 9 terms:
    - Statistical hypothesis
    - Null hypothesis
    - Alternative hypothesis
    - Test of a statistical hypothesis
    - Type I error
    - Type II error
    - Level of significance
    - Test statistic
    - Critical region
28. Describe both fault trees and event trees. What are the differences between the two trees?
29. A distillation column explosion can occur if the overhead cooler fails (*OC*) and condenser fails (*CO*) or there is a problem with the reboiler (*RB*). The overhead unit fails (*OUC*) only if both the coolers fail (*OC*) and the condenser fails (*CO*). Reboiler problems develop if there is a power failure (*PF*) or there is a failed tube (*FT*). A power failure occurs only if there are both operator error (*OE*) and instrument failure (*IF*). Add any other failure possibilities of your choice and then construct a fault tree.
    Comment: Refer to the literature [2] for possible assistance.
30. Describe hypothesis testing in layman terms.
31. One of the authors [5] once outlined a general procedure that may be employed for testing a hypothesis, as described below
    - Choose a probability distribution and a random variable associated with it. This choice may be based on previous experience, intuition, or the literature.

- Set H_0 and H_1. These must be carefully formulated to permit a meaningful conclusion.
- Specify the test statistic and choose a level of significance α for the test.
- Determine the distribution of the test statistic and the "critical region" for the test statistic.
- Calculate the value of the test statistic from sample data.

Accept of reject H_0 by comparing the calculated value of the test statistic with the critical region.

Apply this procedure to a technical application of your choice.

32. Provide a layman's definition of nonparametric tests.
33. Provide a layman's definition of analysis of variance.
34. Describe the role design of experiments can play on the career of a chemical engineer.
35. A redundant system consisting of three operating pumps can survive two pump failures. Assume that the pumps are independent with respect to failure and each has a probability of failure of x. Obtain the equations describing the reliability of the system in terms of x. Calculate the reliability of the system for various x values in the 0.05 – 0.25 range. Analyze the results.
36. Consider a standby redundancy system with one operating unit and one on standby, and a system that can survive one failure. If the failure rate is 2 units per year, what is the reliability of the system over a 2 – 24 month range? Comment on the results.
37. Consider the following numbers:

    8, 14, 23, 34, 42, 49, 57, 66

    Comment on whether these are random numbers, a result of a set progression, or the results of an analytical expression or whether there is some other explanation for this sequence. Comment: This is a tough one. Care should be exercised in interpreting numbers. The numbers given in the Problem statement do *not* represent random numbers. Refer to the literature [3] for the solution.
38. One of the major gambling options during the professional football championship game (Super Bowl) is to "buy a box" in a uniquely arranged square, usually referred to as the

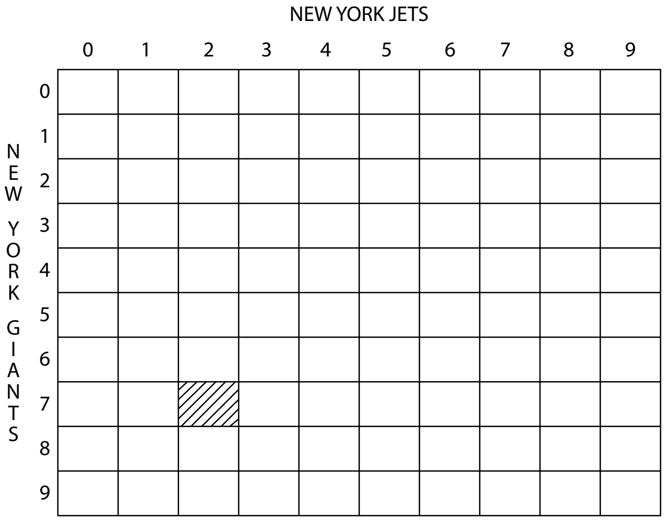

Figure 19.4 Football pool

pool. An example of a pool is shown in Figure 19.4. As can be seen, there are 100 boxes. If each box costs $1,000, the total cash pool is $100,000. The individual, who select the box with the last digit of the final score for each team takes home the bacon, i.e., wins the $100,000. If the final score of the Super Bowl is Jets 22/Giants 7, the owner of the shaded box is the winner. Scores such as Jets 12/Giants 27 or Jets 22/Giants 37 would also be winners. Note that this format does not provide each person buying a box an equal chance of winning (unless the numbers are selected at random).

Revise and rewrite the above Problem statement so that it relates to a technical application.

39. A series system consists of two electrical components, A and B. Component A has a time to failure, T_A, assumed to be normally distributed with mean 100 hours and standard deviation 20 hours. Component B has a time to failure, T_B, assumed to be normally distributed with mean 90 hours and standard deviation 10 hours. The system fails whenever either Component A or Component B fails. Therefore, T_S, the time to failure of the system is the minimum of the times to failure of components A and B. Estimate the average value of T_s on the basis of the simulated values of 10 simulated values of T_A and 10 simulated values of T_B using any convenient method of performing the calculation [2,4].

Comment: Can applying a Monte Carlo method of solution assist?

40. The life (time to failure) of a machine component has a Weibull distribution. Determine the probability that the component lasts at least 25,000 hours if
    1. the failure rate is constant and equal to 0.01 per thousand hours where t is measured in thousand of hours, and
    2. the failure rate is $t^{-1/2}$ where t is measured in thousand of hours. Also perform the calculation for different failure rate values. Comment on the results.
41. Table 19.2 below shows 8 pairs of observations on X and Y where Y is the observed percent yield of a chemical reaction at various centigrade temperatures, X. Select a model of your choice and obtain the least squares line of regression of Y and X and use it to estimate the average percent yield at 235° degrees C° [2]

    Comment: First plot the observed values of Y against the associated values of X and observer whether the scatter diagrams exhibit a linear pattern.
42. Refer to Problems 21 and 22. Attempt to develop a *two* coefficient failure rate model.

TABLE 19.2 Temperature – Yield for Problem 4

Temperature, °C **X**	**% Yield** **Y**
150	75.4
175	79.4
200	82.1
225	86.6
250	90.9
275	93.3
300	95.9
325	96.1

References

1. Webster's New World Dictionary, Second College Edition, Prentice-Hall, Upper Saddle River, NJ, 1971.
2. L. Theodore and F. Taylor, *Probability and Statistics*, A Theodore Tutorial, Theodore Tutorials, East Williston, NY, originally published by USEPA/APTI, RTP, NC, 1993.
3. Adapted from, S. Schaefer and L. Theodore, *Probability and Statistics Applications in Environmental Science,* CRC Press/ Taylor & Francis Group, Boca Raton, FL, 2007.
4. L. Theodore and R. Dupont, *Environmental Health and Hazard Risk Assessment: Principles and Calculations*, CRC Press/Taylor & Francis Group, Boca Raton, FL, 2012.
5. Personal notes: L. Theodore, East Williston, NY, 2013.
6. L. Theodore, *Chemical Engineering: The Essential Reference*, McGraw-Hill, New York City, NY, 2014.